Bio-mathematical aspects of the plasticity of proteins

Rodrigo Dorantes Gilardi

Communauté UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : Mathématiques appliquées

Arrêté ministériel : 25 mai 2016

Présentée par

Rodrigo DORANTES GILARDI

Thèse dirigée par **Laurent VUILLON**, USMB
et codirigée par **Claire LESIEUR**

préparée au sein du **Laboratoire de mathématiques** dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Aspects bio-Mathemiques de la plasticité structurale des protéines

Bio-mathematical a aspects of protein structure plasticity

Thèse soutenue publiquement le **24 avril 2018**,
devant le jury composé de :

Monsieur Laurent VUILLON
PR1, Laboratoire de Mathématiques Université de Savoie Mont-Blanc, Directeur de thèse
Monsieur Hilal LASHUEL
Professeur, Ecole Polytechnique Fédérale de Lausanne, Rapporteur
Monsieur Frédéric CAZALS
Professeur, Inria Sophia Antipolis Méditerranée, Rapporteur
Madame Claire LESIEUR
Chargé de Recherche, IXXI-ENS-Lyon, Co-directeur de thèse
Madame Luisa DI PAOLA
Professeur assistant, Università Campus Biomedico, Examinateur
Monsieur Kavé SALAMATIAN
Professeur, Université de Savoie Annecy, Président

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques Appliquées**

Arrêté ministériel : Arrêté du 7 août 2006 relatif à la formation doctorale

Présentée par

Rodrigo Dorantes-Gilardi

Thèse dirigée par **Laurent Vuillon**
et codirigée par **Claire Lesieur**

préparée au sein **LAMA, Université de Savoie et IXXI, ENS-Lyon**
et de **École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Bio-Mathematical aspects of the plasticity of proteins

Thèse soutenue publiquement le **24 Avril 2018**,
devant le jury composé de :

M, Frédéric Cazals
Directeur de Recherche, INRIA, Rapporteur
M, Hilal Lashuel
Professeur, EPFL, Rapporteur
Mme, Luisa di Paola
Maître de conférences, Campus Biomedico Roma, Examinatrice
M, Kavé Salamatian
Professeur, Université de Savoie Mont-Blanc, Président
M, Laurent Vuillon
Professeur, Université de Savoie Mont-Blanc, Directeur de thèse
Mme, Claire Lesieur
Chercheur CNRS, HDR, Co-Directeur de thèse

A Tina y Go

Acknowledgements

I would first like to thank my two supervisors Claire Lesieur and Laurent Vuillon. Claire's guidance has been a basis for my work, without her energy, advice and support I would simply not be writing these words. Laurent's encouragement throughout my Ph.D. has been a boost on the confidence to pursue my own ideas and his tolerance has always helped to correct the path when they led to a dead end. I was very lucky to have them both as supervisors, as their different backgrounds and expertise prompted me to inquire on a broad selection of topics that enriched my work.

I would like to thank Frédéric Cazals and Hilal Lashuel for accepting being examiners of my thesis, as well as Luisa di Paola and Kavé Salamatian for accepting being part of the jury. I also appreciate the helpful corrections Frédéric Cazals provided for a previous version of my manuscript.

I am grateful for my time at the *laboratoire de mathématiques*, where I got a very useful introduction on Delaunay triangulations by Jacques-Olivier Lachaud. I would also like to thank Michel Raibaut, Olivier Le Gal, and Xavier Provençal who were always available for a less formal chat. Fellow colleagues Mounia, Charlotte, Nadia, Bilal, Rodolphe, and Pierre were always ready to share cheerful and interesting moments. I would like to thank my two friends Paul and Lama for their incredible support, in and outside the laboratory.

My time at the *IXXI* was crucial for the development of my work, and where I spent most of my time. I would like to thank Márton Karsai and Nicolas Schabanel for their knowledge sharing. I am specially grateful with Jean-Pierre Chevrot for his advice and mentorship when it came to looking for postdocs. My colleagues Jacobo, Marija, Sam, Matteo, Si Cheng, and Misha made my time at the laboratory a most pleasant, for which I thank them.

I am thankful to Paul Sorba and Sylvie Ricard-Blum for critical reading of the manuscript leading to Chapter 2. Moreover, I am very grateful with Gelasio Salazar for his support that started long ago but whose influence

span includes my Ph.D. As well as Gallo Ugalde and Carlos Espinoza for their introduction to mathematics and coding.

My friends Memo, Wa, Victor and fellow *cuates* always showed their support throughout my work, as well as my family Ivan, Heidi, Sebastián, Aly, Zoe, and to all others which I cannot mention due to their abundance and the lack of space. I am deeply grateful for the constant love and care I received from Sarah, which helped me rise up to my feet more than once. On the same note, I would like to thank my sister Mariana who has been there since the beginning and my parents Go and Tina, which affection has no words to describe, and to whom this thesis is dedicated.

Last but not least I would like to thank Conacyt of Mexico for their financial support, without which this work wouldn't have been possible.

Abstract

Proteins are biological objects made to resist perturbations and, at the same time, to adapt to new environments. What are the structural properties of proteins allowing such plasticity? To tackle this question we first model protein structure as a network. Given the structural conformation of a mutation, a network approach allows the quantification of its structural change. Using large sets of mutations, we concluded that structural change is independent from the type of amino acid replaced, or replacing after mutation. Looking at the composition of amino acid neighborhoods, we noticed that the location of a type of amino acid in the 3D structure is arbitrary. Leading to the observation that the neighborhood of the amino acid in the 3D structure is the single property related to structural plasticity. Finally, we implemented three algorithms to measure the empty space around amino acids to look at the relation between void and structural plasticity. Results show a clear gap in small atomic distances, invariant across a large dataset, suggesting a cutoff to separate intra-atomic and inter-atomic void, based on the distances of covalent and non-covalent interactions.

Résumé

Les protéines sont des objets biologiques conçus pour résister des perturbations et, en même temps, pour s'adapter aux nouveaux environnements. Quelles sont les propriétés structurelles des protéines permettant une telle plasticité ? Pour aborder cette question, nous modélisons d'abord la structure des protéines en tant que réseau. Compte tenu de la conformation structurelle d'une protéine mutée, une approche en réseau permet la quantification de son changement structurel. En utilisant de grands ensembles de mutations, nous avons conclu que le changement structurel est indépendant du type d'acide aminé remplacé ou de celui de l'acide aminé remplacent après mutation. En regardant la composition des voisinages d'acides aminés, nous remarquons que l'emplacement d'un type d'acide aminé dans la structure 3D est arbitraire. Ceci menant à l'observation que le voisinage de l'acide aminé dans la structure 3D est la seule propriété liée à la plasticité structurelle. Enfin, nous avons implémenté trois algorithmes pour mesurer l'espace vide autour des acides aminés afin d'observer la relation entre le vide et la plasticité structurelle. Les résultats montrent un écart clair dans les petites distances atomiques, invariant à travers un grand ensemble de données, suggérant une coupure pour séparer le vide intra-atomique et inter-atomique, basé sur les distances des interactions covalentes et non-covalentes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Protein structure

In this section, we introduce the concept of the amino acid network and its relation to the structure of a protein. We'll see that there are different amino acid networks depending on the targeted structure of the protein. The regular blocks of the secondary structure can be coarse-grained into the tertiary and quaternary structures, in that order. The primary structure, however is not yet *folded* and cannot be coarse-grained into the secondary structure. Nonetheless, it is this the basic structure, that contains the necessary information for the conformation of the remaining structures [4].

Here, we explain how the building blocks of the protein structure, the so-called *amino acids* are at the center of each structural categorization of the protein. We start then with an overview on the amino acids followed by the explanation of the primary structure as a linear sequence of amino acids. Next, we present the secondary and tertiary structures as the atomic conformation of those amino acids once achieving their final position in the three-dimensional space. It is here that the concept of amino acid network is first introduced, and then further developed with the presentation of the quaternary structure of a protein as a set of tertiary structures.

1.1.1 Amino acids and primary structure

The first amino acid ever discovered was asparagine in 1806 by chemists Vauquelin and Robiquet [72]. Since then, other 20 standard amino acids have been discovered: They are the essential element or building block of the protein structure.

An amino acid is composed of an amine group (-NH_2), a carboxyl group (-COOH), and a side chain called the R group (Figure 1.1). The atoms of an amino acid consist mainly of carbon, hydrogen, oxygen, and nitrogen. As a matter of fact two amino acids only differ on their side chain composition and are thus categorized based on properties of their side chains into four groups: acidic, basic, un charged polar (hydrophilic) and non polar (hydrophobes).

Amino acids are commonly joined together by a peptide bond, in which the carboxyl group of one amino acid is linked to the amino group of another releasing a molecule of water (Figure 1.1).

A linear chain of more than two amino acids joined together in this fashion is named a *polypeptide*. Amino acids in a polypeptide are usually called *residues*, a reference to the fact that they release either a hydrogen (H) ion from the N-terminal or a hydroxyl (OH) ion from the C-terminal, or both. Proteins are polypeptides composed usually of more than 20-30 residues.

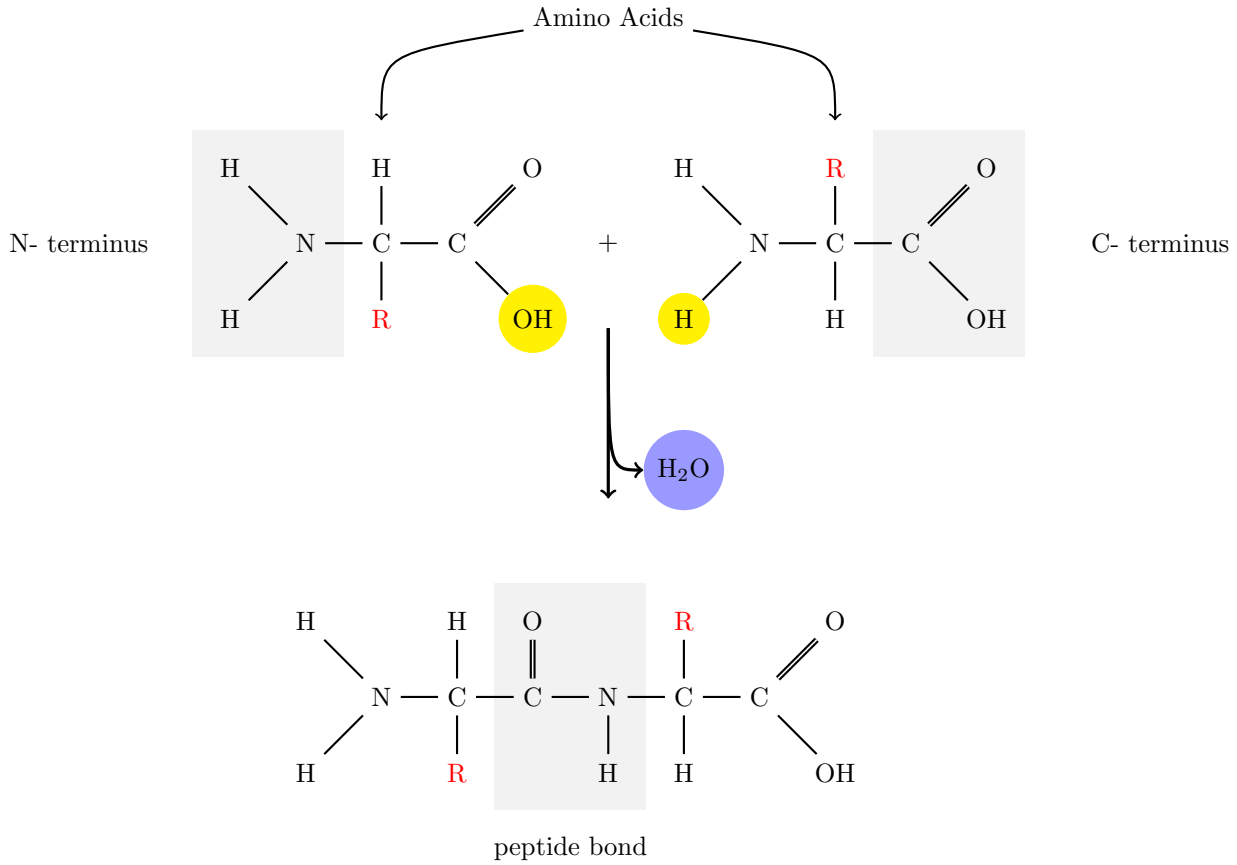


Figure 1.1: Representation of the formation of a peptide bond between two amino acids. The resulting *peptide* is read from the N-terminal to the O-terminal by convention.

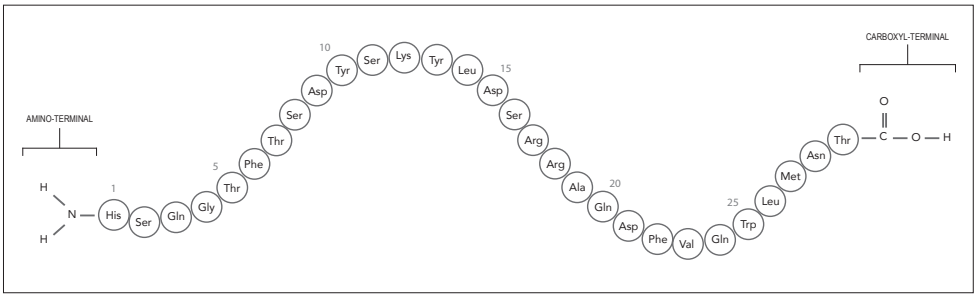


Figure 1.2: A polypeptide is a linear sequence of amino acids connected through a peptide bond. By convention it is read from the N-terminal to the C-terminal, assigning each amino acid a position in the sequence.

This linear sequence of residues is the primary structure of a protein. By convention, the sequence starts at the N-terminal and ends at the C-terminal. The order of a residue in the sequence is what it's called the residue's *position* (Figure 1.2).

The positions of amino acids in the sequence will be later used to label the nodes in the amino acid network. To define the amino-acid network, however, we first need to talk about the positions of amino acids in space.

1.1.2 Secondary and tertiary structures

The primary structure of a protein lacks any three-dimensional structure on its own. It is on a shapeless state called *random coil* where residues are only connected to their neighbors in the linear chain.

The random coils forms local conformations almost spontaneously. The ensemble of these local conformations is what is known as the secondary structure of the protein. They are composed by short segments of the linear sequence taking a three-dimensional shape. The two most frequent motifs in the secondary structure are the alpha helix and the beta sheet. The alpha helix is the most regular as well as the most prevalent in proteins, it was considered to be the secondary structure on its own when the concept of secondary structure was coined [37]. It has a helicoidal shape stabilized by a bond between an oxygen from the N-group of a residue and a hydrogen from the O-group of a residue four positions further in the linear sequence. The beta sheet consist of two or more adjacent beta strands stabilized just like the alpha helix with hydrogen bonds between the N-group hydrogen of

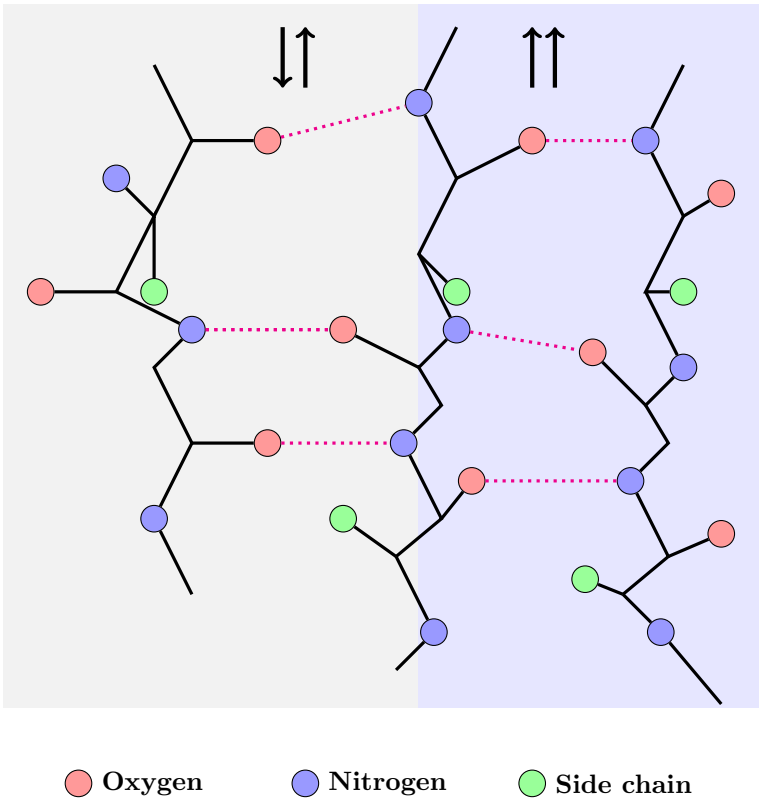


Figure 1.3: Three beta strands representing two beta sheets. The two beta strands on gray background form an anti-parallel beta sheet in which the two strands are in opposite directions. The two strands on blue background form a parallel beta sheet in which the two adjacent strand follow the same direction. The hydrogen bonds are represented by magenta colored dotted lines. The N-group hydrogen of a residue connects with the C-group oxygen of a residue in an adjacent strand.

Table 1.1: Non-covalent amino acid interactions in proteins, the distance at which they occur and their abundance [28].

Interaction	Distance (Ångströms)	Abundance
Van der Waals	4-8	Numerous
Hydrophobic	4-8	Numerous
Hydrogen	4	Moderate
Charged with uncharged groups	4-8	Moderate
Salt bridge	5	Few
Coordinate bond	2.5	Very Few
Disulfide bond	2.5	Very Few

a residue and the C-group oxygen of a residue in an adjacent strand. The sheet can be parallel when adjacent strands follow the same direction or anti-parallel otherwise (Figure 1.3). Together with the alpha helix, it was discovered in 1951 by Pauling, Corey and Branson [46].

Another form of secondary structure called loops, is a more irregular motif which can serve to redirect the direction of the polypeptide chain often to create beta-sheets, in which case it is called a beta turn [27]. Like beta sheets and alpha helices, loops are also stabilized by internal hydrogen bonding.

After the amino acid sequence folds locally almost spontaneously to give rise to the local structures, the sequence continues to fold to reach a stable three-dimensional shape. This shape is the tertiary structure of a protein and its defined by the atomic coordinates of the polypeptide chain in (Euclidian) space. The tertiary structure is a single folded-chain of amino acids containing one or more secondary structure motifs. The polypeptide chain without its side chain in the tertiary structure is called the *backbone* of the protein. Any two adjacent residues in the linear sequence are also adjacent in the backbone by a peptide bond. The compactness or closeness of adjacent amino acids in the tertiary structure is determined by non only covalent (i.e. peptide bond) interactions, but other types of non-covalent interactions. These non-covalent interactions in space can occur between amino acids being otherwise faraway in the linear sequence. Non covalent interactions take place at different distances, usually ranging from 2.5 to 8 Ångströms (1 Å = 10^{-10} meter). Non-covalent interaction distances are shown in Table 1.1.

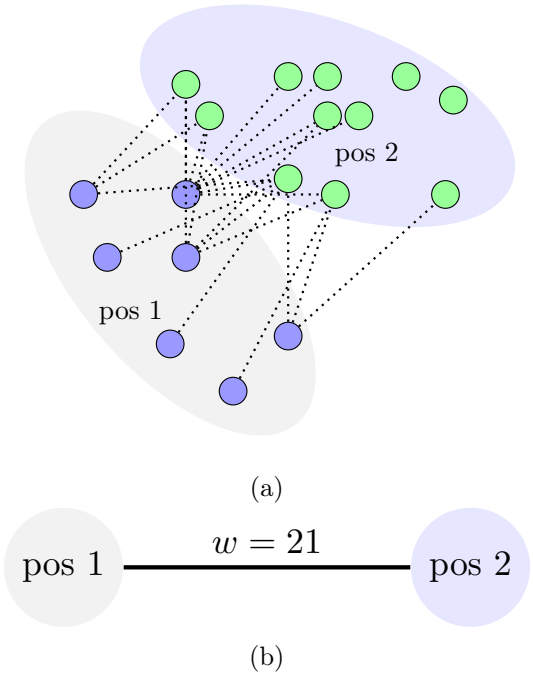


Figure 1.4: Atomic interactions happen at an atomic level. (a) A sketch of two amino acids in interaction labeled by their position number in the polypeptide chain. Two atoms are connected by a dotted line if their distance is less than a given threshold. (b) The representation of the same interaction in the amino acid network. Two nodes are labeled by the positions of the amino acids they represent and connected by a weighted link, where the weight (w) is equal to the number of atom pairs in interaction between the two amino acids.

As previously mentioned, the position of residues in the polypeptide chain will be used here to label the nodes of the amino acid network. It would now be natural to connect two such nodes by a link or an edge if they are interacting, i.e., if their distance is smaller than a threshold or *cutoff* (Figure 1.4).

Under this approach, consider two connected residues in the amino acid network; the closer they are from one another, it follows that the larger the probability that they are interacting. The concept of *weight* of a link can be used (and indeed it is!) to quantify the proximity between two already close residues: the weight of a link is equal to the number of shared atom pairs closer than a given cutoff (Figure 1.4).

To formally define the amino acid network of a protein, we first need to define the protein structure as a set of points in the Euclidian space $\mathbb{R}^3$ (Definition 1).

Definition 1. The structure of a protein, noted $S \subset \mathbb{R}^3$, is the set of atomic coordinates of the protein.

We can now think of an atom as a point in $\mathbb{R}^3$ and a residue as a set of atoms. Think of atoms and residues as such is convenient for the the parts of this work dealing with the spatial modeling of the protein. The notation $a \in S$ which pretends a to be a point in S should also be thought of as a is an atom of the protein. Similarly, $a \in r \subset S$ explains that the atom a is an atom of residue r, which in turn is a residue of the protein.

The term atom and point (resp. residue and set) will be use interchangeably throughout the text. Exceptions for this will be only found (we hope) under a sufficiently evident pure biological discussion.

A network or a graph, usually noted G, is an ordered pair (V, E) where V is the set of nodes or vertices and E is the set of links or edges. We use this terminology to formally define the amino acid network of a protein structure (Definition 2). The weight of an edge is defined as a function w where the weight of the edge is the number of atom pairs interacting between the two connected residues (Definition 3).

Definition 2. Given a protein structure $S \subset \mathbb{R}^3$ and a distance cutoff c, the amino acid network (AAN) of S (using c) is $G(S, c) = (V, E)$. Where V is equal to the set of residues in S, and a link uv connecting residues u and v exists if and only if there are two atoms $a_1 \in u$ and $a_2 \in v$, such that $d(a_1, a_2) < c$, where $d(a_1, a_2)$ is the distance from a_1 to a_2 (Figure 1.5).

Definition 3. Let $G(S, c) = (V, E)$ be the amino acid network of S and e an edge in E. The function $w : E \mapsto \mathbb{R}$, called weight, assigns to any edge $uv \in E$, the number of pairs $(a_1, a_2) \in S \times S$ where $a_1 \in u$, $a_2 \in v$ and $d(a_1, a_2) < c$ (Figure 1.5).

The interaction is calibrated by a distance cutoff c, depending on the target interaction. The concept of a distance cutoff, allows us to model different interactions varying on the distance at which they are defined. The amino acid network is mainly used to model interactions to quantify variations in the structure, as explained first in Chapter 2.

Many proteins are composed of more than one polypeptide chain or subunit. The quaternary structure is used to categorize these compounds, as seen in the next subsection.

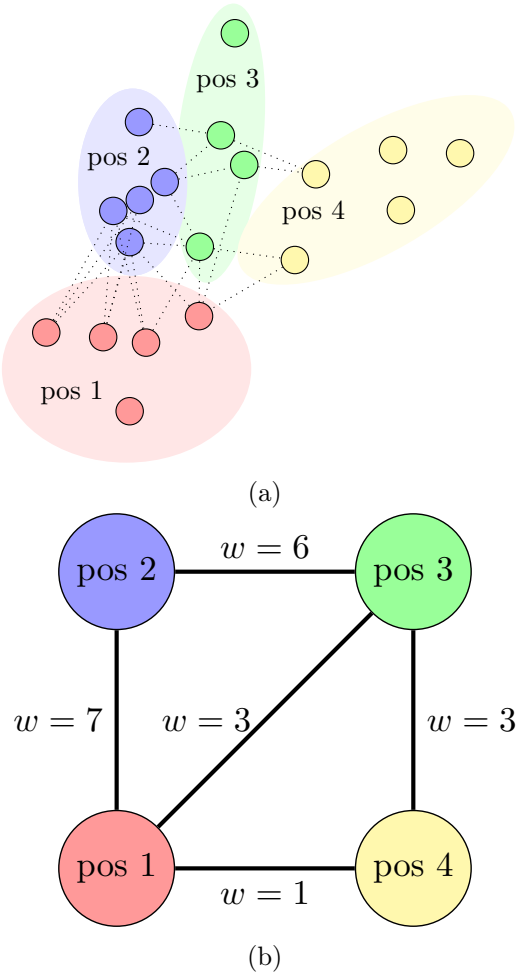


Figure 1.5: Example of a small amino acid network of a protein structure in $\mathbb{R}^2$. (a) The protein structure S. Four residues with their respective atoms are depicted in different colors, all atom pairs in interaction share a dotted line. The interaction between atoms is defined by a distance threshold or cutoff. (b) Draw of the amino acid network from the small structure shown above. Nodes are the residues in the structure and are labeled by position number in the linear sequence. By definition, each edge in the network has a positive weight.

1.1.3 Quaternary structure

Several polypeptide chains can fold into a unique protein. Because one gene codes for one chain, proteins with several chains can contain more than one *subunit*. A monomer, is a molecule unit that through polymerization, can group with other molecules to form a larger compound. While being part of the compound, the monomer is called a subunit. The quaternary structure of a protein is composed by the total number of subunits in the compound. Monomers, dimers, and oligomers are proteins composed of one, two and several subunits, respectively.

Two amino acids in an oligomer, can therefore be part of an intermolecular interaction, i.e., an interaction on two different polypeptide chains. Otherwise, the interaction is called intramolecular. Zones where there are intermolecular interactions are called interfaces; in other words, interactions happening only at the level of the quaternary structure. Residues laying on protein interfaces are referred to as "hot spots".

The stability of interfaces is of crucial importance to the overall stability of the protein structure. There is therefore an interest to study the structure of protein interfaces on their own. In chapter one, we propose a model of amino acid network, called the hotspot network used to study structural variations on the protein interfaces belonging to the cholera toxin. The hotspot network of an oligomer is a subnetwork of its amino acid network (Definition 4). A network or a graph $G = (V, E)$ contains the network $G' = (V', E')$ if and only if $V' \subset V$ and $E' \subset E$. In which case G' is called a subnetwork or subgraph of G.

Definition 4. The hotspot network (HSN) of a structure S given a cutoff c, noted $H(S, c) = (V^H, E^H)$, is a subgraph of the amino acid network $G(S, c) = (V, E)$, where the set $V^H \subset V$, is equal to the set of hotspots in S, and an two hotspots u and v share an edge uv, only if u and v lie on different chains and $uv \in E$ (Figure 1.6).

The majority of the protein structure networks used in this work are subnetworks of the amino acid network as well and will be defined in the following chapters.

Several structural properties of an amino acid are found the amino acid network. The degree of an amino acid in the network, that is, the number of links that are connected to it, depends on the distribution of other amino acids around it in the protein. Similarly, the weight of an amino acid is the number of the atoms belonging to other amino acids that are close. Will se this more in detail in Chapters 2, 3 and 4.

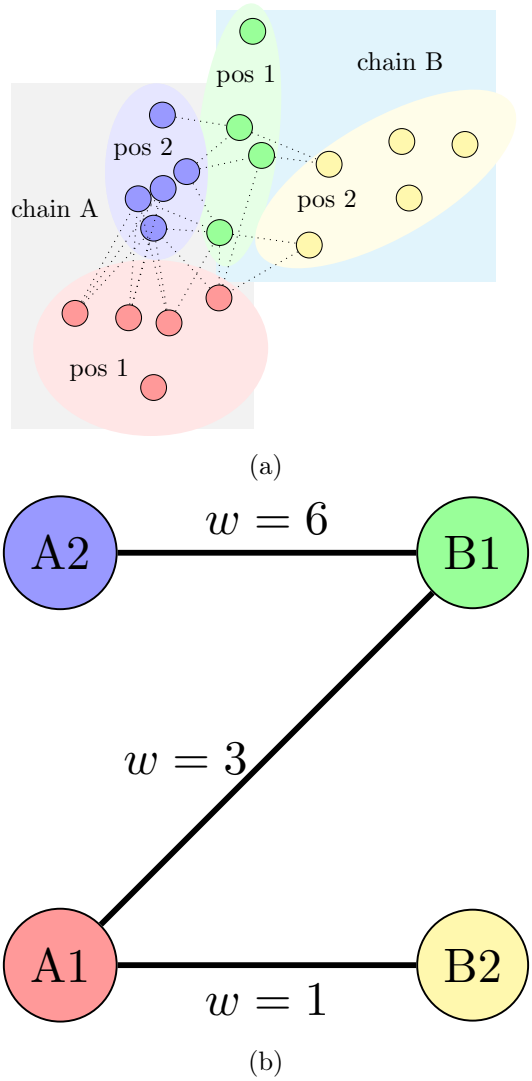


Figure 1.6: An example of a hotspot network of a structure in $\mathbb{R}^2$. (a) The protein structure is composed of two chains each having two residues. The atoms of each residue are shown in a different color. Atom pairs at interaction distance are represented by dotted lines. (b) The hotspot network. Only residues in opposed chains can be interacting. Note that the labeling of nodes is prefixed by the name of their chain.

Other parameters can be calculated like the betweenness centrality of a node, which is the number of paths in the network passing to that node. The closeness centrality is the length of the paths between the node an all other nodes in the network. These parameters can be used to study the relevance of the node in the network in terms of interaction paths.

In the protein structure, same type (and size) amino acids can be different in volume as we'll see in Chapter 5. We'll see that also the empty space around amino acids or void in the structure varies considerably between close residues. This is a consequence of the combinatory power of amino acid neighborhoods in proteins as shown in Chapter 5.

There is one subject that is omnipresent in different forms during the entire spread of this work: mutations. A mutation is an evolutionary phenomenon happening continuously in proteins.

During the first part of this thesis, we study the effects of mutations on the protein amino acid mutations. We consider mutations to be (only) variations of the atomic coordinates of proteins in the three-dimensional space. However, the variation in the structure is only a consequence of mutations. In reality, a mutation is a change in the genetic sequence of the protein happening during DNA replication. Subsequently, the protein is constructed from a segment of DNA or a gene into an amino acid sequence by the process called *protein synthesis*.

1.2 Protein synthesis

The process of protein gene expression starts with a segment of DNA (deoxyribonucleic acid) called gene and ends with the synthesis of the protein amino acid sequence. In this section we briefly explain how this process is carried first in the (Eukaryotic) cell nucleus and then in the cytoplasm. Protein gene expression can be divided into two subprocesses: *transcription* and *translation*.

1.2.1 Transcription and translation

Transcription of a gene refers to the process of *copying* the information in the gene stored in the nucleus of the cell into another molecule, called RNA (ribonucleic acid). Transcription can be divided into three steps: initiation, enlarging, and termination. Initiation starts when the molecule RNA polymerase, while bound to the DNA, encounters a *promotor* site signaling the start of a gene. The RNA polymerase then unwinds the DNA double helix and in a complementary fashion starts copying the nucleotides from one of

the DNA strands. The strand used is the template strand, to which RNA is going to be complementary. The RNA polymerase copies one nucleotide at a time by covalently linking the new complementary nucleotide of the RNA to the one previously added forming the RNA backbone. This is the phase of elongation. The RNA polymerase eventually finds a *terminator* or stop site and halts the synthesis of the RNA and releases from the template DNA strand. The double helix is closed again and the RNA molecule is ready to be used by the cell. The resulting RNA can be used for other jobs in the cell besides the creation of a protein molecule, here we'll only focus on the messenger RNA (mRNA), the molecule in charge of passing the genetic message (Figure 1.7). The mRNA exits the cell's nucleus by its pores and enters the cytoplasm, where the subprocess called *translation* takes palace.

The translation of the mRNA into a new sequence of amino acids is done by a protein called ribosome. The ribosome translates the information encoded in the mRNA into amino acids. The sequence of the mRNA is divided in sets of three consecutive nucleotides called codons. Each codon translates to one of twenty amino acids, with the exception of the stop codons, which terminate the translation. All 64 codons ($4 \times 4 \times 4$) with their translation compose the genetic code. In this code, every amino acid is coded by more than one codon except for Methionine, and three codons are used to signal the termination of the translation.

Translation can be divided into four steps: activation, initiation, elongation, and termination. The activation phase consists of the amino acids binding to the transfer RNA (tRNA) molecules, which will transport the amino acids to the ribosome. The activation phase is when a small subunit of the ribosome binds the end of the mRNA (first codon). The elongation phase consists of the charged tRNA (tRNA with its corresponding amino acid) matching the codon, and binding to the ribosome. Finally, the termination phase occurs when the ribosome encounters either a nonsensical codon or a stop codon, and finally detaching from the amino acid sequence (Figure 1.7). As we mentioned before, a mutation which is explained in the next subsection, can alter the structure of a protein. This alteration will be measured with the use of amino acid networks by comparing the networks of the mutated structure and the wild type.

1.2.2 Mutations

A point mutation is a variation happening in the nucleotide sequence of the DNA (or RNA). They occur in nature mainly during DNA replication, but can happen also during the transcription and translation. Most point

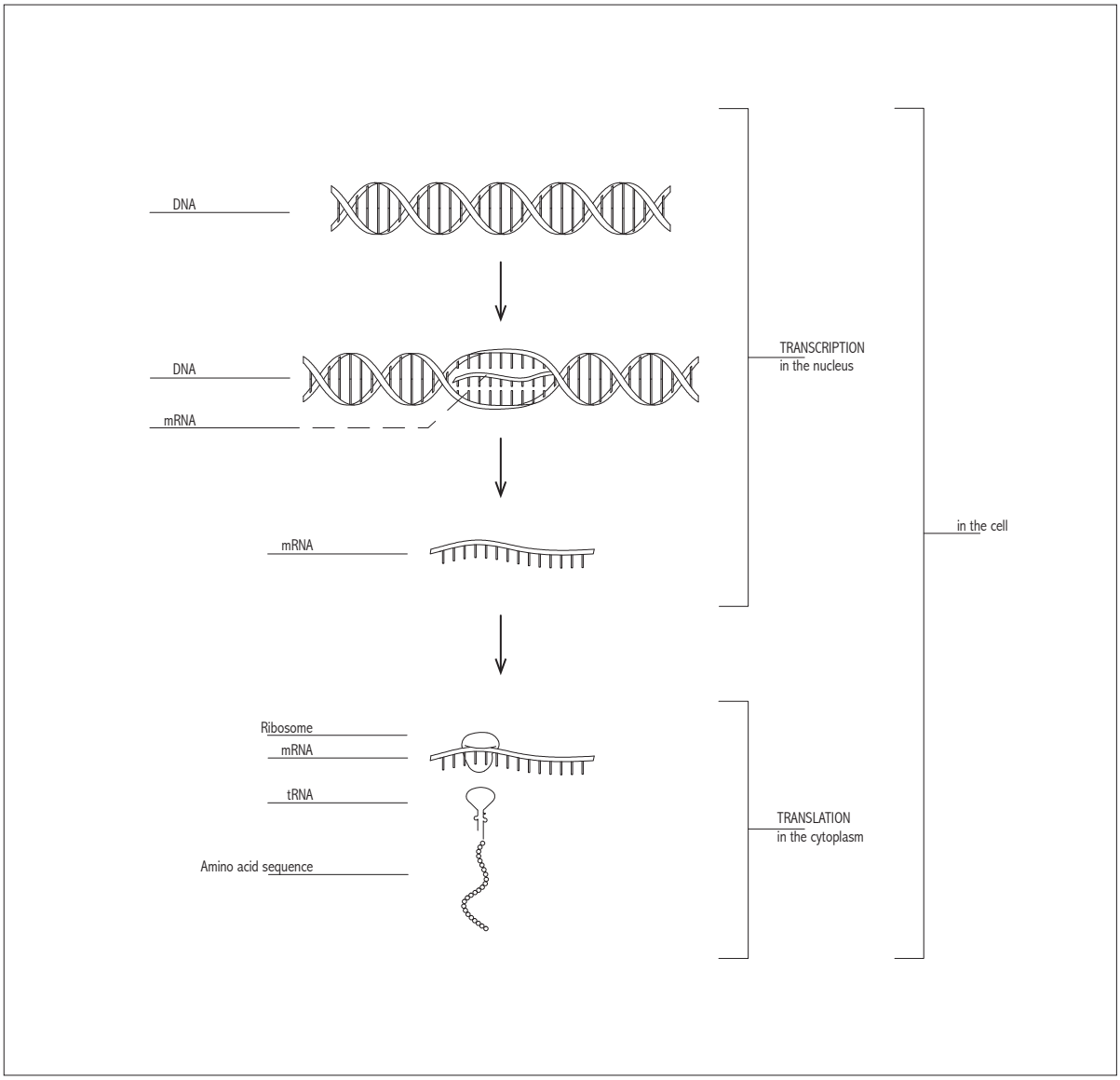


Figure 1.7: Representation of protein gene expression, starting in the nucleus of the cell where genes are transcribed into an RNA molecule, and ending with the translation of the RNA into an amino acid chain in the cytoplasm.

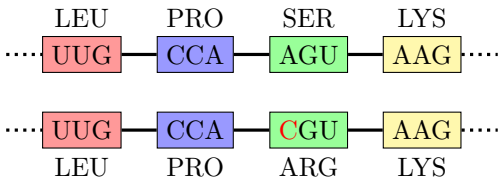


Figure 1.8: A point mutation.

mutations do not impact the amino acid sequence of a protein, they mostly happen in non-coding regions of the DNA. The robustness of the protein also relies on the redundancy of the genetic code, where a point mutation happening in a codon do not change the amino acid coded (e.g. CUU and CUC both code for amino acid Leucine). Even if the point mutation yields a change in the amino acid sequence (Figure 1.8), the alternative amino acid is likely to adapt to the previous structure and function.

However, a mutation can produce a change to the structure of a protein and therefore to its function. In order to understand the role of mutations in the protein function and structure, we first need to mention the relation between the amino acid sequence and the structure and function of a protein.

The function of a protein depends on the underlying protein structure. Allosteric shifts or intrinsically disordered regions in the protein structure can yield several functions in a same protein [79]. The main purpose of the protein structure is indeed to accomplish one, or several functions. Therefore the protein must fold in the conformation allowing the protein to well-function. This interdependency between function and structure is one of the motivations to study the structure of proteins. This is supported by the fact that there are extensive databases of crystalized protein structures available online.

1.3 Amino acid networks

1.3.1 Protein structure and function

An additional variation in the definition of amino acid networks lies in the distance used to consider two atoms to be interacting. When, for example, only alpha carbons are considered, the distance cutoff is considerably larger than when all atoms are taken into account. Finally, some authors consider the side chain atoms only, neglecting the backbone.

Networks have been applied to biological hot topics related to the func-

Table 1.2: Distance cutoffs used in literature.

Nodes	Cutoff(s) used in literature (Å)	Network type
C_α	7, 8, 8.5	Unweighted
C_β (C_α for Gly)	7, 8.5	Unweighted
Centroids of side chain	8.5	Unweighted
Amino Acid	4.5, 5	Weighted

tion and the structure of the protein. Here we give a general presentation of some of those applications disregarding the methods of construction of the amino acid networks.

1.3.2 Mutations

The use of amino acid networks in the study of important sites related to functional change has also been applied. In one study, the authors compared the centrality values in the amino acid network to relate residues with high values to destabilizing sites. Where mutations are usually detrimental to the protein. In this work, we propose several other network parameters, not to identify functional sites, but to measure the impact of mutations of those sides in the underlying amino acid network.

1.4 Tools

1.4.1 Computational

The entirety of the computational tools used in this work, relative to the computations on networks and structure but not including the visualization, are done in the programming language Python (version 3.6). We used several python "modules", or sets of functions previously constructed, to elaborate our own algorithms. Here we present the main modules used for our own computational tools and, next, our own computational tools also written in Python. The previously constructed modules by other authors are not thoroughly surveyed, as the documentation is easily found online (and it drifts away from the purpose of this methodology). Instead, we show some examples on how to use those tools in the work relative to network theory and bioinformatics (Figures 1.9 and 1.10).

The tools we developed for this work are in the form of Python functions gathered in the package called "biographs". The package contains five

```
In [1]: %matplotlib inline
        import networkx as nx
        G = nx.Graph() # A graph in networkx
        V = ['a', 'b', 'c', 'd'] # A set of 5 nodes
        E = [('a', 'b'), ('a', 'd'), ('b', 'c'), ('b', 'd'), ('c', 'a')] # A set of 5 links
        G.add_edges_from(E)
        nx.draw_networkx(G, node_color='w') # This gives the following figure
```

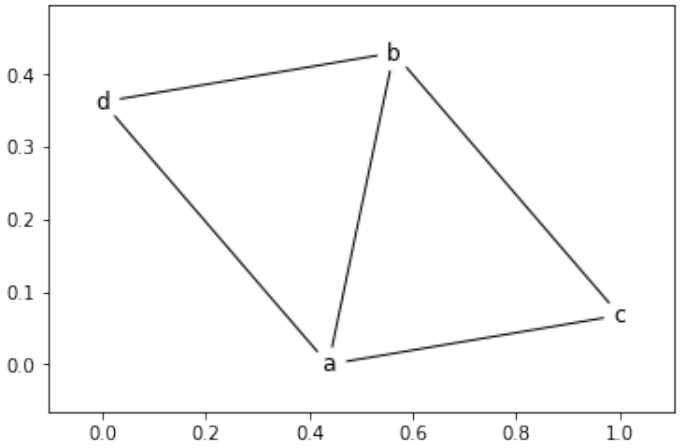


Figure 1.9: **networkx** is a library for python comprising multiple functions dealing with networks. The module includes functions to display the networks (as in this example), compute the degree of nodes, the assortativity coefficient, the clustering coefficient, etc [23].

```
In [3]: import Bio.PDB
        pdb_file ='/Users/rdora/Desktop/1be9.pdb' # The pdb file containing the protein
        protein_parser = Bio.PDB.PDBParser(PERMISSIVE=1)
        # The variable `structure' contains the structure of the protein
        structure = protein_parser.get_structure('PDZ protein', pdb_file)
        A = Bio.PDB.Selection.unfold_entities(structure[0], 'A') # Atoms
        R = Bio.PDB.Selection.unfold_entities(structure[0], 'R') # Residues
        dis = A[0] - A[1]
        print 'Number of atoms: ' + str(len(A))
        print 'Number of residues: ' + str(len(A))
        print 'Residue %s%i is a %s' %(R[0].parent.id,R[0].id[1],R[0].resname)
        print 'Atoms %s, %s of %s%i are at dis: %f'
                %(A[0].id,A[1].id,R[0].parent.id,R[0].id[1],dis)

Number of atoms: 1045
Number of residues: 1045
Residue A301 is a PHE
Atoms N, CA of A301 are at dis: 1.494597
```

Figure 1.10: The module biopython parses and information in a PDB file and has algorithms for structural biology. Here we show an example of the computation of atoms and residues of the protein with PDB identifier "1BE9". Moreover, we can easily check the types of residues in the protein and compute the distance between any pair of atoms.

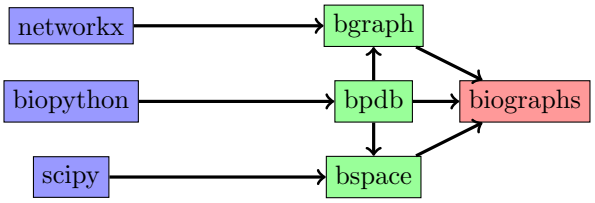


Figure 1.11: The architecture of the package *biographs*. The network represents the architecture of the package, a directed link indicates source node (module) is used in the target node (module or package).

modules, including one producing the amino acid networks present above and another module for the creation of Python objects from PDB files [10]. Two other modules are mainly used for the purpose of the two previously mentioned. A last module contains functions dealing with the spatial properties of residues and their atoms, to compute their void, volume, and spatial centrality (Figure 1.11). The package will be available at `http://rodogi.github.io` together with a user interface (future work).

In this section, we present the process to produce the three-dimensional structure of a mutation *In Silico*. The process is divided in two main steps: The retrieval of the atomic coordinates of the native protein and the computation of the new atomic coordinates given a mutation (or a set of mutations) by the software FoldX.

The computed structures of the proteins are deposited in the protein data bank (PDB). For this work, we retrieved all the structures used from the Research Collaboratory for Structural Bioinformatics (RCSB) protein data bank, a member of the worldwide PDB. The atomic coordinates of the structure, together with additional information relative to the process of obtainment of them, including the authors, the process and the journal where the work was published, are saved in a pdb file (Figure 1.12).

The file of the native protein to be mutated is used as part of the input in the FoldX software to the computation of the three-dimensional conformation of the mutation. In combination with the PDB file, FoldX takes as an input the name of the mutation(s). A separate file, called "Individual_list.txt" must be provided with the mutations to make (Figure 1.13). Each line of this file is a mutation or group of mutations to be made simultaneously. The mutation(s) is produced using the function BuildModel, included in the software. The function first mutates the selected position to Alanine and annotates the side chains of the neighboring positions. Those who exhibit energy differences are then mutated by themselves to minimize

```
HEADER    EXTRACELLULAR MATRIX                    22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

Figure 1.12: An extract of a PDB file, taken from the PDB file of the seal myoglobin (PDB identifier: 1MBS). The header includes the name of the molecule, the authors, the method used (in this case X-ray diffraction) and remarks. The atomic coordinates of the molecule are then shown for each atom of the structure together with the amino acid to which it belongs, the chain and the type of atom.

```
FA39L,FB39L;
LA41F,LB41F;
KA42I,KB42I;
GA46S,GB46S;
AA47V,AB47V;
LA48S,LB48S;
LA52S,LB52S;
```

Figure 1.13: An example of the file "individual_list.txt" used as input for the obtaining of the structure of a mutation using the software FoldX. Each line represents a mutation or a number of mutations (as in this example). The format used to indicate the mutation is the type of amino acid in the selected position, the chain, the position, and the mutant amino acid. Mutations in a same line are produced simultaneously and lines are separated by a semicolon.

the energy change [71]. Finally, the selected position is mutated to itself and two pdf files are created, a "individual wild type" file and the file containing the structure of the mutation.

Chapter 2

Protein structural robustness to mutations: an *in silico* investigation

Abstract

Protein functional performances depend on the protein capacity to handle the structural changes responding to functional purposes and perturbations (e.g. mutations). The present study identifies in-built parameters responsible for such structural supervision from the survey of 736,149 amino acids and of their spatial neighborhoods. It appears that regardless of amino acid type or position in a structure, amino acids interact with their neighbors *via* a moderate average number of atomic links, achievable by all, and following a Goldilocks principle: not too many links, not too few. The structural responses to mutation depend on the reproducibility of the average number of atomic links at the mutated position, condition accomplished by customizing neighbors, *via* amino acid alternative solutions or compensatory mutations. On the other hand, the modulation of pairwise atomic interactions governs structural transitions such as protein folding.

2.1 Introduction

How proteins sustain and adapt their biological functions, or fail to do so, is a complex phenomenon. The structure and function of a protein are defined by amino acid sequences that naturally vary upon genetic mutations. The robustness of proteins against mutations depends on the impact on the protein function of the structural changes arising from the mutations, changes which are not much investigated [28]. Proteins are strongly resistant to single amino acid mutations: most amino acids can be mutated without loss of function [68], i.e., such mutations are functionally neutral. Less frequently, with a frequency about 10^{-9} per site, mutations lead to the emergence of new functions (innovation) [3]. Alternatively, there are pathological mutations which lead to a loss of function. The present view of neutral mutations is that some are adaptive because their combination with other mutations drives functional evolution through non-additive effects (e.g. functional promiscuity or epistasis) [3]. Non-additive effects are also involved in rescue mechanisms, wherein the negative effect of a pathological mutation is neutralized by a mutation at a second site [13, 43, 57, 68]. In general, protein robustness, protein innovation and protein adaptation, refer to the impact of mutations on the biological function of proteins.

On the other hand, the structural changes which are tolerated by a protein without jeopardizing the protein's functionality (functional robustness or emergence of a new function) or those that, on the contrary, lead to loss of function, are rarely looked into. Therefore, even little understanding of the underlying structural changes would be instructive for addressing pathological mutations or help designing new enzymes. The gap between the studies on functional and structural robustness is due to several issues. To investigate functional robustness, a protein prototype is chosen, every individual amino acid is mutated and the function of each mutant is tested experimentally [42]. Likewise, studying structural robustness, namely, maintenance of the structural integrity necessary for a biological function, implies choosing a protein prototype, mutate every individual amino acid, crystallizing each mutant, solving each structure and comparing the ones which share the same function. First, this is technically and financially challenging as well as time consuming. Second, the goal is to understand if a protein structure is built to bear mutational changes and if so, to investigate by what mechanisms. Furthermore, an experimental approach is not appropriate, because some mutations would fail to produce a structure for reasons not necessarily related to structural robustness. A mutation might prevent folding and acquisition of a stable structure, yet have no impact on the structural ro-

bustness. For instance, the B subunits of the pentamers of the cholera toxin and the heat labile enterotoxin, maintain a pentamer at ph 5.0 but do not reassemble at this ph [11, 55, 56, 80]. Moreover, a mutation leading to a new structure and a new function might not easily be identified as such, experimentally. On the other hand, *in silico* mutations produce structural changes in order to generate a stable structure. *in silico* methods cannot create a new structure or destroy a structure from a mutation, and they produce a set of conformations close to the wild-type structure. This is a relevant framework to investigate the structural changes that underlie structural robustness as a general issue rather than having to restrict the study to specific mutations. The third issue is the lack of tools available to measure and compare the effects of mutations on a structure, comparison needed to understand the mechanisms by which the protein structure bears the changes. There exist programs to compare global structure features (e.g. rmsd) and visualize structural differences [5, 60, 62, 78]. However, the present study is about following changes from a local perturbation, the site of the mutation, to the entire protein structure.

To circumvent these difficulties, we have adopted the following strategy. We have worked on the atomic structure of the pentamer of the cholera toxin B subunit ($CtxB_5$) because it is a stable protein with an ob-fold, structure common to many other proteins with different sequences. We can therefore assume that the structure is naturally robust to mutational changes. We generated a set of *in silico* mutations using foldx, which produces structural changes maintaining a reliable structure [61]. Let us recall that the goal of the study is not to predict the effects of experimental mutations on a structure, but to have a set of mutations appropriate to explore structural robustness. The dataset is the individual mutation of all the amino acids which compose the toxin interface. To analyze the structural changes due to mutations, we modeled the toxin interfaces as networks of amino acids in interaction, such that the structural properties are compared through network comparison. The analysis of the networks helped us build an *ad hoc* algorithm, called amino acid rank (aar) which takes into account all structural changes observed in the dataset, quantifies them, and ranks the mutations accordingly.

Finally, we analyzed the results of aar in terms of structural robustness. The results indicate that mutations generate structural changes at different scales (local or long range) in a cascade mechanism and independently of the local changes on the mutation site and of the nature of the mutation. Structural robustness relies not only on mutations producing zero or a few little changes, but also on mutations producing significant structural changes

while generating redundant conformations, in good agreement with the recent definition of protein as an ensemble of conformations fulfilling one function. Redundancy produces the alternative structures necessary for having conformations functionally distinct upon secondary mutations, consistently with "adaptive neutral mutations". An example of non-additive mutations is provided not in the context of emerging functions, but as a correction mechanism of a cancer-related mutation reported in the tetrameric domain of the tumor suppressor p53. This error-correction mechanism is not conceivable if structural robustness is based only on a lack of structural changes upon mutation. The identification of a second site mutation capable of correcting the fault is possible because of the new algorithm aar.

2.2 Methods

2.2.1 Aminoacidrank (aar) algorithm

function spectralpro

The goal is to model a protein interface by a hotspot network. A protein interface is made when the amino acids of one chain with the amino acids of adjacent chains. These amino acids are referred to as "hotspots", in this work. To construct a hotspot network, we first define its atomic network. Using the atomic coordinates from a pdb file, all distances between atoms of one chain and atoms of adjacent chains are computed. Two atoms share a link if they are within a 5å distance from each other. Two hotspots share a link if they have at least one of their respective atoms within 5å distance from one another. It is convenient to represent the hotspot network as its adjacency matrix a. If n is the number of hotspots in the protein, then a is the $n \times n$ matrix with value $a_{i,j}$ in row i and column j, if i and j are connected by a link or 0, otherwise. The weighted adjacency matrix w is defined by $w_{i,j}$, the weight of the link connecting i and j, that is, the number of atomic links between amino acid i and amino acid j. In the adjacency matrix a, $a_{i,j} = 1$ if $w_{ij} > 0$, and $a_{ij} = 0$, otherwise.

Function Arank

A mutated pdb file is generated with foldx introducing a single hotspot mutation of a residue at position r. The function spectralpro is then applied on the mutated pdb file. To compute the quantity of structural changes produced by the mutation, a $n \times n$ difference matrix d is defined as follows: $d_{i,j} = w_{i,j}^{mut} - w_{i,j}^{wt}$, where $d_{i,j}$ is the entry value of d at row i and column

j, and $w_{i,j}^{mut}$ and $w_{i,j}^{wt}$ are the weights at row i and column j of the mutated network and the wild type network, respectively.

The structural changes produced by the mutation on the entire structure (global changes, arank_r) are computed as the sum of the absolute value of all the entries of d (that is, $\sum_{i,j} |d_{i,j}|$). The structural changes at the position of the mutation r (local changes, local_r), are computed as the sum of the absolute values of all entries of d at row r (that is, $\sum_j |d_{r,j}|$). The arank_r values are used to rank mutations according to the amount of structural changes they produce.

Function backup

This function computes the redundancy of every link of the wild type (wt) hotspot network. The backup links are sought within the local secondary structure around every hotspot link, based on the known hydrogen bonding of secondary structure. The set of backup links of link (i, j), includes any link incident to a residue located within four residues along the sequence of residues i or j. The aar pseudocode is provided in section **??** (subsection 2.7.2).

2.2.2 Foldx

Mutations were computed using the protein design tool foldx (version 3 beta) [22, 61]. Only the protein design function was used for mutagenesis using the pdb file with code 1eei as the wild type structure (details and run parameters to be found in subsection 2.7.1). Essentially, the run parameters are chosen to minimize their impact on the network construction, to be applicable broadly on x-ray structures, and not to depend too strongly on a high quality structure. Herein the quality of the structures need to be at $\sim$ 2.5å resolution or above.

2.3 Results and discussion

The aim is to investigate the structural changes that a protein may go through from individual mutations of its amino acids, while maintaining a stable structure. As a model of study, we use CtxB_5, focusing on the amino acids that compose the toxin interface, the so-called hotspots. A protein structure is built on atomic interactions between its amino acids, likewise for a protein interface. Thus, to analyze the structural changes that take

place in the toxin interface upon mutation, first intermolecular atomic interactions need to be established. The exact atomic interactions are intractable due to the large size of the system. Atomic interactions rely on chemical nature of atoms, distances between atoms and the atom environment (atomic neighbors). In order to take these parameters into account, the following procedure is undertaken (section 3.4.2). The distances between all atoms of one chain and all atoms of an adjacent chain, referred to as interatomic distances are calculated from the x-ray coordinates of $CtxB_5$ provided by the rcsb protein data bank (pdb code 1eei). All interatomic distances within 5å are considered as chemical interactions, without distinguishing the nature of the atoms (section 3.4.2). This approximation is reasonable because every type of chemical interactions (van der waals, electrostatic, hydrogen bonds, etc.) between the atoms of amino acids, i.e., carbon, oxygen, nitrogen, sulfur and hydrogen fall within a distance of less than 5å [21]. The chemical nature of atoms is not considered also because it is assumed that two atoms in the x-ray structure would not be close if they ought to chemically clash. They are either necessarily chemically compatible or their neighbors' shielding prevent them from clashing.

To each hotspot is associated a weight w_i equal to $\sum_j w_{i,j}$, which is the total number of its links (intermolecular atomic distances within 5å, section 3.4.2). The pairs of atoms that are within a 5å distance are coarse-grained to their respective amino acids in order to associate to each hotspot a number of amino acids in physical contacts (degree), noted a_i and equal to $\sum_j a_{i,j}$. Because all the distances within 5å of every atom are considered, the algorithm intrinsically accounts for the neighboring atoms. The weigh and the degree can be considered as a proxy of the probability of interactions of the amino acid, the higher the degree the more likely the amino acid is to have an interaction.

2.4 Survey of the structural changes

Our algorithm amino acid rank (aar) after establishing the amino acids and the interactions that composed the toxin interface with the above procedure, models the interface as a network of amino acids in intermolecular interactions (section 3.4.2, function spectralpro). The amino acids that have at least one intermolecular atomic distance within 5 å are linked and referred to as hotspots. The $CtxB_5$ interface has 58 hotspots forming the nodes of the network, these are also recognized as hotspots by other programs available [2]. There are no histidine nor cysteine hotspots.

We systematically mutate every hotspot one by one. In the current study we restrict ourselves to mutations to asparagine residue for simplicity, as it has average chemical and geometrical properties. For example it is a residue that is polar rather than hydrophobic or charged, and has an average number of atoms compared to other amino acids. Mutations to other amino acids will be considered in future work.

In silico mutations are performed using foldx (section 3.4.2) [61], to generate a mutated structure, from which a mutated toxin interface and a mutated network are produced by the aar algorithm. To capture the structural changes associated with a mutation, aar compares the networks after and before mutation and extracts all modified amino acid links (section 3.4.2, function arank). Mutations change the positions of atoms which modify the intermolecular atomic distances and thus the degrees and weights of nodes of the network. To quantify the structural changes produced by a mutation at position r within the entire structure (arank_r), aar sums the absolute values of the differences between the weights after and before mutation of all the nodes of the networks; the higher the arank_r the larger the structural changes (table 2.1). A change in weight means some atoms have become closer or further away, implying atomic interaction rearrangements. Depletion of an amino acid link means that the two hotspots have no more atoms within a 5 å distance. Addition of a new link means that the two hotspots have moved closer so that they have atoms within 5 å distance. These are amino acid link rearrangements. To qualitatively describe the mutations, a "sphere of influence" is defined as the number of modified amino acids by the mutation and the distance between the site of the mutation and the modified residue the furthest from it (table 2.1). Two distances are measured, geodesic and euclidian. The geodesic distance is measured by the number of chemical links to be crossed to go from the site of mutation to the modified residue the furthest from it, by the shortest path, and the euclidian distance is measured between the two residues in ångströms (figure 2.1). The spheres of influence of the fifty-eight mutations are shown on their respective x-ray structures in figure 2.7 (section **??**), highlighting the broad diversity of structural changes in quantity and quality.

The arank_r values vary from 182 to 2, ten mutations have an arank_r below the first quartile while fifteen have an arank_r above the third quartile, and thus most mutations generate significant changes (Table 2.1). The changes involve side chain atoms only since the RMSD is zero for all mutations. No more than 10% of the native interfacial contacts are lost upon mutations. On average the mutations modified eight hotspots; a quarter modified only up to five hotspots and a quarter modified more than eleven. Thirteen muta-

Table 2.1: mutational features

Hotspot i has degree a_i and weight w_i

	Global Changes					Local Changes			
Mutations	arank$_r$	# modified hotspots	Geodesic	Euclidian (Å)	Intra	a_i^{wt}	w_i^{wt}	$\Delta a_{i(mut-wt)}$	$\lvert\Delta w_{i(mut-wt)}\rvert$
K69N	182	22	3	15	0	2	30	-1	23
R67N	178	16	4	9	0	9	101	-5	24
Y76N	143	8	2	6	0	4	40	-3	37
Q3N	120	4	1	5	0	4	43	-1	26
A64N	112	18	3	10	0	4	20	5	41
Y12N	102	10	4	9	0	4	41	-4	44
T78N	97	5	3	8	0	1	2	0	1
A32N	94	11	2	5	0	5	35	2	44
E29N	90	11	3	10	0	6	77	0	30
R73N	88	18	3	15	0	4	42	-1	32
Y27N	86	16	3	13	0	5	41	-2	8
E66N	82	15	3	13	1	2	33	0	8
A98N	80	8	2	11	0	3	23	1	38
101N	76	11	2	13	0	6	60	0	5
F25N	72	6	2	5	0	3	39	0	28
103K	70	7	2	5	0	4	44	-2	33
A80N	66	9	3	9	0	1	1	1	24
K23N	66	6	3	11	0	1	7	-1	7
G33N	60	7	2	6	0	3	25	1	29
T71N	56	12	10	10	0	3	31	0	4
K81N	54	5	3	9	0	1	1	0	0
D70N	53	16	5	16	1	2	28	-1	10
L77N	51	14	3	10	0	4	8	1	3
S26N	48	5	1	5	0	2	15	2	25
P2N	48	7	2	7	0	4	19	0	14
V50N	46	14	4	17	1	1	1	0	1
R35N	46	9	1	6	1	5	47	-1	9
E36N	42	15	2	14	1	5	36	-2	1
Q61N	38	10	3	8	0	4	39	0	6
A97N	38	8	2	5	1	3	34	0	15
T28N	36	8	1	5	0	4	35	3	16
E11N	36	4	2	5	0	1	15	0	12
100N	34	5	1	5	0	2	22	1	17
T1N	33	7	1	5	0	5	32	0	1
I99N	32	8	2	10	0	3	36	1	15
P93N	32	6	2	5	0	3	31	0	3
S30N	30	7	2	5	0	5	31	2	14
I58N	26	6	2	5	0	3	10	-3	10
I74N	20	10	4	9	0	3	7	-2	5
K34N	20	4	1	5	1	3	11	0	4
L31N	18	8	1	5	0	9	74	-1	1
S60N	16	8	3	7	0	2	18	0	3
L8N	16	11	2	5	1	5	19	-2	3
K63N	16	9	3	12	0	4	19	-2	6
W88N	16	7	2	11	1	3	11	-2	7
I65N	16	8	2	11	0	1	7	0	3
M68N	16	7	2	5	0	3	31	-1	8
Q49N	16	4	1	7	1	1	7	-1	7
N4K	15	5	2	5	0	1	4	3	11
I39N	14	7	1	5	0	4	16	-1	5
P53N	12	5	2	5	0	1	3	2	3
M37N	12	4	1	5	0	3	8	-2	4
I24N	12	4	2	8	1	1	1	0	0
102N	12	3	2	5	0	3	26	0	6
T92N	8	2	1	5	0	2	16	0	4
I96N	2	3	2	5	0	1	6	0	0
I5N	2	3	2	6	1	1	7	0	0
T47N	2	2	1	5	0	1	10	0	1

tions out of fifty-eight produced only local perturbations, namely, structural changes of residues in physical contact with the site of the mutation and so located within the chemical reach of the mutated residue (Euclidian and geodesic distances within 5 Å and 1, respectively). Forty-five mutations produced global changes, namely, changes beyond physical contact and chemical reach of the mutated residue. Eighteen modified residues located at distances above 10 Å. The maximum long range modification is 17 Å. The mechanism of the long range modifications is chemically sound since the changes are going from hotspots chemically linked to hotspots chemically linked in a step-by-step manner as determined from the geodesic distances (Figure 2.1). This cascade mechanism seems related to the secondary structure of the mutated residue since out of eighteen residues belonging to α-helices, seventeen produce a cascade (long range changes) upon mutation (95%). Out of twenty-six which belong to a β-structure, thirteen produce a cascade (50%) while out of fourteen which belong to a loop, twelve produce a cascade (86%). This relation would need to be verified and further explored on a dataset. There are twelve mutations for which the changes did not go from hotspots to hotspots but went from the mutated residue to its intramolecular contacts, which subsequently modified their hotspots (Table 2.1, column *Intra*). It was still a step-by-step mechanism, but through intramolecular and intermolecular links. Thus the results highlight paths of changes between amino acids of the interface and amino acids outside it. Likewise, mutations of amino acids outside the interface are capable of modifying the degrees of hotspots (work in progress). This is consistent with the mechanisms of protein assembly combining folding and association steps in a coordinated manner (for review [34]). A step-by-step mechanism is described in other real networks as Peer-to-Peer mechanisms (P2P) [44].

2.4.1 Specific examples

As selected examples, the mutations K69N, A64N, L31N and I39N are considered in detail because they allow covering the chemical and geometrical properties of amino acids (small, medium and large side chain, hydrophobic, charged and polar chemical nature). Their spheres of influence are shown in the X-ray structures of the respective mutants (Figure 2.2). Large modifications are seen for the K69N and A64N mutants while fewer modifications take place for the mutants I39N and L31N. The mutations K69N and A64N are among the top disruptive ones with arank_r values equal to 182 (first rank) and 112 (fifth rank), respectively (Table 2.1). This highlights that the extent of the structural changes cannot be inferred by the difference between

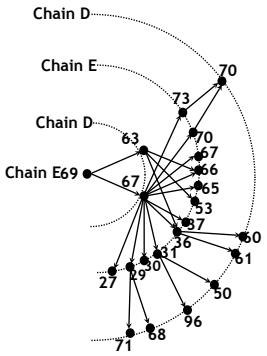


Figure 2.1: Schematic of the cascade mechanism underlying the structural changes associated with mutations. As the most disruptive mutation, K69N is chosen to illustrate the paths of the structural changes going from the site of mutation to elsewhere in the interface. The K69N mutation modified the atomic interactions of twenty-two hotspots of the interface, covering a distance of fifteen Ångströms. The paths of changes are schematically described by arrows going from hotspots (nodes, black circles) chemically linked to hotspots chemically linked. The chemical distances (5 Å) are illustrated by dotted semi-circles. However, because the structure is a three dimensional object, the Euclidian distance between the site of mutation and the residue modified the furthest from it cannot be calculated from the schematic. The geodesic distances are the number of chemical links crossed to go from one hotspot to another. The structural changes of K69N cover three chemical links.

the nature of the original and mutated residue, since lysine is bigger and has more atoms than asparagine, while alanine is smaller and has fewer atoms. This is further supported by the fact that the mutations of other lysine or alanine such as K34N and A102N have different AAR values (Table 2.1). To consolidate this point, the spheres of influence shown in Figure 2.7 are sorted by amino acid type, and subsequently sorted by decreasing values of arank_r.

Now if the mutation K69N is compared to the mutation L31N, the latter has an arank_r value ten times lower than 182. Yet, the residue L31 has degree 9 and weight 74, significantly higher than the degree and weight of the residue K69: 2 and 29, respectively. Like the nature of the residue, the degree or the weight does not condition the extent of the structural changes. This is further evidenced by plotting the arank_r values against the weight of the original

residue before mutation for the fifty-eight mutations (Figure 2.2). The linear correlation is weak (Figure 2.2, $R^2 = 0.27$), indicating that mutation of an amino acid with a high weight does not systematically lead to large structural changes, and likewise mutation of an amino acid with a low weight does not necessarily lead to few structural changes.

The arank_r values are then plotted against the local weight changes (local_r, weight differences on the mutated residue after and before mutation, Section 3.4.2), and again a rather weak linear correlation is observed (Figure 2.2, $R^2 = 0.44$). This indicates that global changes are not proportional to local changes. Moreover, only some mutations have arank_r values which fall on the straight-line of slope two implying local changes (Figure 2.2, red line). Most mutations have arank_r values outside this line and so they produce global changes and involve cascades. If there are only local changes, that is, weight changes on the mutated residue and nowhere else, then the global changes are twice the local changes because the global changes count the weight changes on the mutated node and on its endpoint nodes. This confirms that mutations produce changes at different scales as shown by the spheres of influence (Figure 2.7). The absence of correlation between arank_r values and local weight before mutation, or the local weight changes, remains true even if the networks are built with cutoffs 4 and 6 Å instead of 5 Å. Thus, these properties are invariant within the experimental error of X-ray structures ($\sim$ 1Å). It is interesting to discuss the two AAR outliers, the mutation R67N and the mutation K69N, because they have similar local and global changes (Table 2.1). What is different, however, is their fraction of local changes: R67N lost 24% of its interactions (24/101, ratio local weight difference to weight before mutation), whereas K69N lost 77% (23/30). The fraction of local changes does not correlate either with the global changes measured by AAR (not shown).

2.5 Structural Robustness, Fragility and Adaptation

To assess whether the structure of a protein is built to bear mutational effects, we propose to consider the structural changes produced in the CtxB_5 interface by the mutations and see if they are consistent with all known mutational effects: robustness, innovation, adaptation/rescue and pathology.

The first key point is that the mutations yield structural impact at different scales (Table 2.1, Figure 2.2, Figure 2.7). This means there is no *a priori* specific scale (e.g. 5 Å) at which structural changes can be detectable,

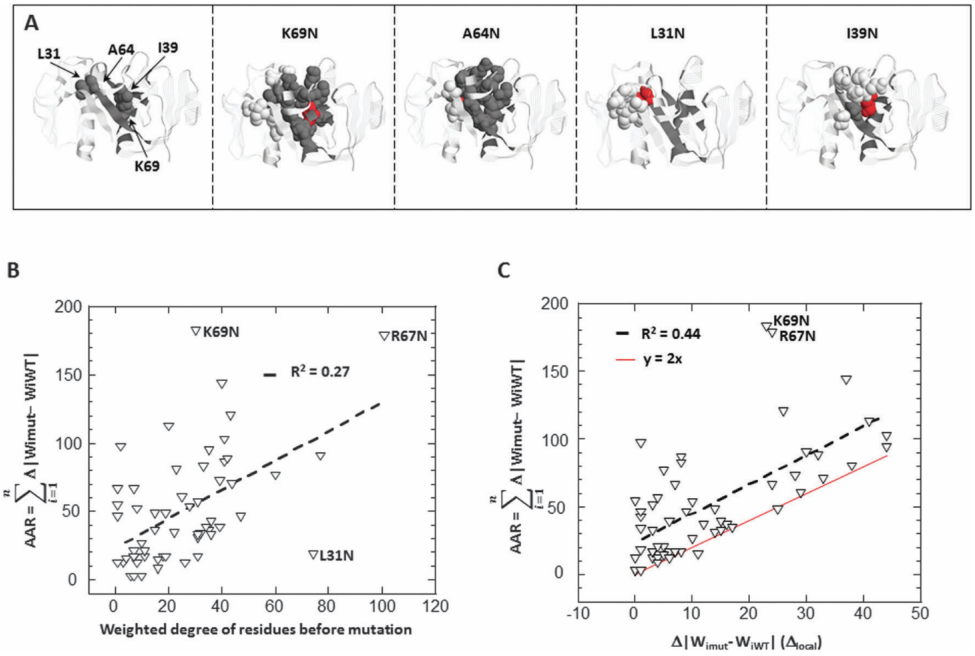


Figure 2.2: Local degrees and global changes. (**A**) Spheres of influence. Only two adjacent chains D and E of $CtxB_5$ are represented in pale and dark grey strands, respectively (PDB code 1EEI). The toxin interface is in ribbon. The residues modified by mutations are space-filled and the mutated residues are red. The left panel shows the location of the four mutated hotspots K69, A64, L31 and I39 on the WT structure. The other panels on the right are their respective spheres of influence, as shown on their respective X-ray structures. (**B**) Weak correlation between the original weighted degree of the mutated residue and the amount of structural changes after mutation measured by AAR. Values of arank_r are plotted against the weights of each hotspot i before mutation, w_i^{wt}. The dotted line is the linear correlation. (**C**) Global vs. local changes. Arank_r values are plotted against local_r values (Section 3.4.2, local weighted degree differences ($|w_i^{mut} - w_i^{wt}|$). The dotted line is the linear correlation and the red line is for $y = 2x$.

and it is necessary to measure them locally as well as globally. This is in good agreement with other studies showing both direct and indirect physical interactions in co-evolving residues [28]. Local structural changes, namely, modification within the chemical reach of the site of the mutation are consistent with enzymatic innovation or adaptation that does not lead to a full reorganization of the global structure. Global structural changes are consistent with pathologies where a single mutation is enough to jeopardize a structure and consequently a function. Of course, this does not imply that enzymatic innovation and pathology occurs only via local and global changes, respectively. This all depends on the scale at which the function is regulated by the structure.

The scaling does not explain adaptation through epistasis, a rescue mechanism, or compensatory mutations (non-additive effects). Let us consider the pre-requisite for such effects: a mutation at a site 1 with an effect 1 (Mutant 1) and a mutation at a site 2 with an effect 2 (Mutant 2). Non-additive effects mean the consequences of the combination of mutations 1 and 2 are different from the consequences of mutation 2 (or of mutation 1) individually. This implies that the structures of the mutant 1 (or of mutant 2) and of the wild-type are different, otherwise they would react similarly upon the secondary mutation (Figure 2.3). In other words, a robust mutation that leads to a rescue mechanism or a compensatory effect upon a second site mutation necessarily has a structure distinct from WT. This suggests that functional robustness is built on mutations with no structural impact (neutral mutation) as well as on mutations producing distinct structural solutions functionally equivalent to WT (adaptive mutations). If true, this means that among networks different from WT (i.e. $\text{arank}_r \neq 0$), some should be WT-alternative and other should be dissimilar. To investigate this possibility, the four mutations K69N, A64N, L31N and I39N are considered again. The structural changes due to these mutations are schematized by networks before and after mutation on Figure 2.4. Let us first consider the mutations K69N and A64N which both have significant structural changes, namely, high arank_r (Figure 2.4). The K69N mutation modifies the layout of the WT network substantially, since it reduces the atomic interactions between the region of interface composed of residues 63 to 67 of one chain, and residues 73 and 65 of the adjacent chain, and simultaneously increases the atomic interactions between the residue 67 of one chain and the residues 27 to 37 on the adjacent chain. This is well-illustrated in the X-ray structures (Figure 2.4). Moreover, the mutation also depletes the only two weak ties of the WT network, namely the links (31, 50) and (63, 53) which connect two regions of interface otherwise disconnected.

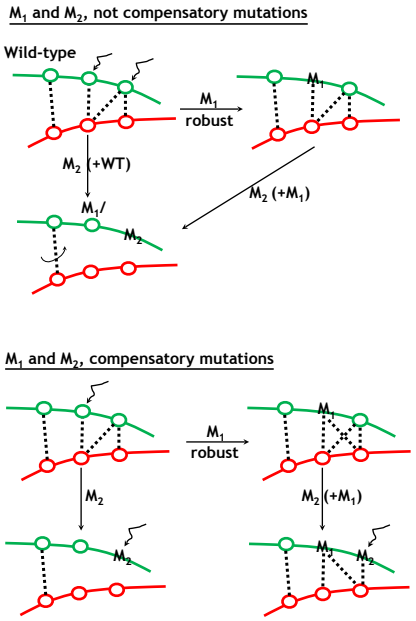


Figure 2.3: Schematics of additive and non-additive mutational effects. A WT network maintaining two segments together through four links of amino acids is drawn. Two sites of mutations M_1 and M_2 are considered. Non compensatory mutations (Upper schematic). If M_1 implies no structural and network reorganization, then M_2 has the same effect on the WT and M_1 mutated network. Compensatory mutations (Lower schematic). Mutation M_2 produces structural defaults disconnecting four nodes (Left). Mutation M_1 reinforces the connectivity of WT (right). A compensatory mechanism (non-additive effect) is set when mutation M_2 happens after mutation M_1.

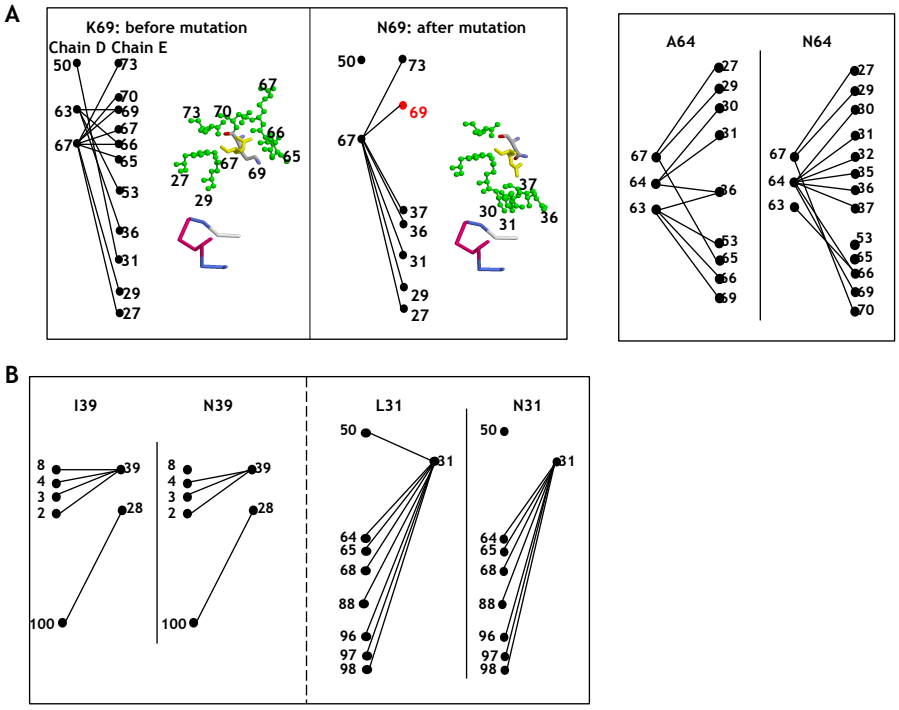


Figure 2.4: Structural robustness. (**A**) Networks of K69 and A64 residues, before and after mutation. Networks of the sphere of influence with hotspots nodes and links of hotspots as links. Zoom on a subset of interfacial residues in the X-ray structures of K69 and N69 (ball-and-stick representation). The numbers are the sequence positions of the residues. The residue 69 of chain E and the residue 67 of chain D are shown in CPK and yellow, respectively. The residues of the chain E are otherwise colored in green. The backbone shows that both structures are in the same position. (**B**) Networks of the spheres of influence of the residue I39 and L31, before and after mutation. Legend as in (**A**).

In contrast, the networks A64 and N64 have a similar layout (Figure 2.4). In fact, the N64 network appears like a WT alternative network with more amino acid links, but the same regions are connected. The K69N and A64N mutations well-illustrate the distinction between structural changes and alternative structural solutions. The mutations I39N and L31N have low arank$_r$ (14 and 18, respectively), but a similar result can be observed (Figure 2.4). Only the link (39, 8) is depleted in the I39N mutation, not modifying the network significantly since there are other linked residues in the vicinity of the link (39, 8) (Figure 2.4). In contrast, the L31N, even though it also yields a single link depletion (31, 50), the mutated network is not equivalent to WT because it lacks the only link that was connecting the regions 50, 64–68, 88 and 96–98 through the intermolecular link (31, 50) (Figure 2.4). It is therefore important to acknowledge that structural changes, large or small, yield alternative networks or not. Therefore, the quality of structural changes must also somehow be incorporated in order to anticipate the impact of a mutation. Because of the scaling issue and the cascade mechanism, establishing the appropriate measure for alternative networks to sort out robust (neutral and adaptive) and fragile mutations is complex and beyond the scope of the present work.

The obvious difference between the A64N and I39N alternative networks and the altered K69N and L31N networks is the redundancy of amino acid and atomic links in the former. This is reminiscent of peer-to-peer networks, which are robust to perturbation because they have more links than necessary—back up links—such that depletion or addition of links is tolerated by generating several alternative networks [8]. To see if alternative structures and networks exist in proteins, we measured backup amino acid links in the interface of CtxB$_5$. Two amino acid links (i, j) and (i', j') belonging to the same secondary structural element, implies that residues i and i' are four amino acids apart along the sequence; as well as j and j' residues, and are considered to backup each other. This is because the integrity of the secondary structure relies on at least the amino acid links which participate to the hydrogen bonding. The maximum distance of four amino acids apart along the sequence corresponds to a helix turn, so backup links are counted within this range of distance along the backbone. Based on this definition of backup, AAR calculates the number of backup links for each link of the WT network (Section 3.4.2, function Backup). Out of 92 links of amino acids, only the two weak ties have no backup. Eleven links have 1 to 3 backups, fifty-two have 4 to 13 backups and twenty-seven have more than 14 backups. A backup network of the WT toxin interface is shown in Figure 2.5, with the number of backups of each link described by a color code. The net-

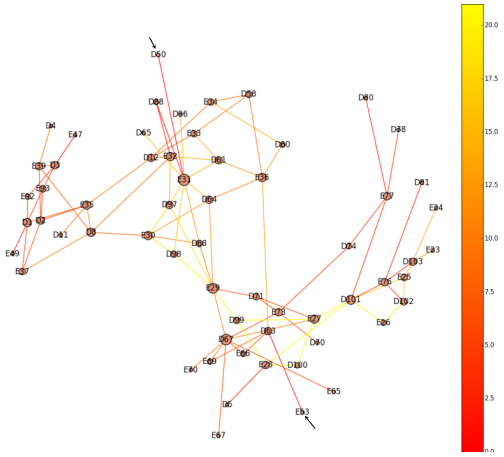


Figure 2.5: Backup network of the WT interface. Structural robustness is based on the presence of backup links that allow bearing of addition and depletion of links without structural impact. The nodes of the backup network represent the hotspots, and the size of nodes represents their degree. The links represent pairs of hotspots and the colors of the links represent the number of backup for each link within a range indicated by the color scale on the right. The redder the link, the fewer backup interactions the pair of amino acid has. The arrows indicate the positions of the two nodes with weak ties (50, 31) and (53, 63). The letters on the network are the chains on which the hotspots are located.

work shows a non-uniform distribution of the number of backup per hotspots within the structure which may indicate fragile areas. This result supports the possibility of having neutral structural changes through addition and/or depletion of links producing alternative networks and structural robustness (Figure 2.5). The backup of the residues K69, A64, L31 and I39 are 16, 41, 80 and 26, respectively. The mutations A64N and I39N which have a redundant network, also have a higher backup than K69N. The mutation L31N has the highest backup but the amino acid link (31, 50) has none. This illustrates the complexity in assessing robustness due to the scaling problem (robustness of a node, of a link or of a region/community). Nevertheless, the results are encouraging to further explore the concept of backup as a measure of robustness and fragility.

WT alternative networks lay the ground for non-additive mutational effects because different atomic interactions cope differently with secondary

mutations. A mutation not tolerated in a WT network/structure might be tolerated in a mutated WT alternative network. We tested this possibility to further support a mechanism of robustness *via* alternative WT networks. The cancer-related mutation G334V reported for the tetrameric domain of the tumor suppressor p53, is used as a default mutation case [25]. The goal is to find a second site mutation which alone produces neutral structural changes and a WT alternative network when coupled with the G334V mutation prevents the structural damages associated with G334V, corroborating non-additive effects through alternative networks. The impact of the G334V mutation on the protein conformation is such that X-ray crystallography is inapplicable, and there is no fiber structure available yet. The mutation G334V is generated *in silico* from the WT atomic structure (PDB code 1SAK) using FoldX instead. The interface between chains D and B was analyzed. The G334V mutation leads to a large amount of structural changes, with an AAR value of 286. Moreover, there are side chain and backbone atom rearrangements since the RMSD is 0.03 Å. The sphere of influence reveals long range changes up to residues at geodesic distances 5 and Euclidian distance 15 Å from the residue 334 (Figure 2.6). The structural changes go from the residue 334 up to the residue 324 on the N-terminal end and up to the residue 352 on the C-terminal end (Figure 2.6). The mutation does not change the degree of the residue 334, but it changes the degree of its intramolecular amino acid neighbors, residues 333 and 337, in a cascade mechanism (Figure 2.6). As a result, the residue 337 loses its pairing with the residues 345, 349 and 352, and maintains its pairing only with the residue 348, reducing the connectivity within the interface region composed of the residues 345 to 352 and 337 to 341 (Figure 2.6). Moreover, the residue 333 also loses pairing with the residue 345, removing a link between the interface region composed of residues 330–334 and 325–328, and the interface region composed of the residues 337–341 and 345–352 (Figure 2.6). It is possible that the rigidity between these two regions loosen up after depletion of the link 345–333. The residue N345 is at the crossroad of the structural changes produced by the mutation G334V. We tested whether a mutation at this position could reinforce the atomic interactions of the network such that it becomes robust to the G334V mutation. Again *in silico* mutations are performed using FoldX. The network of the single mutant N345D is similar to the WT network, except for an increase of the weights (number of atomic interactions) of the links (345, 333), (345, 341), (337, 348), and (337, 349), and a decrease of the weight of the link (337, 345) (Figure 2.6). The double mutant N345D+G334V has structural changes on half as many residues as the mutant G334V, it maintains both links (345,

333) and (345, 337), and its network looks like the WT network, apart from an additional link between the residue 333 and 352 found as well in the single mutant G334V (Figure 2.6). The small changes in the atomic interactions produced by the N345D prevent the residue 337 from moving away after the mutation of the residue 334 and prevent the loss of the link (333, 345). This is a non-additive mutational effect, since the effects of the individual mutations differ from the effects of combined mutations; the effects of the G334V are lost when combined with the N345D mutation. This suggests that a second site mutation producing a compensatory effect is to be found among the residues modified by the first site mutation, namely, it is on the sphere of influence of the first site mutation. This hypothesis is supported by the observation that on average, in the interface of $CtxB_5$, eight amino acids are modified by mutation and on average, deleterious mutations can be compensated by nine mutations [28, 48].

2.6 Conclusions

The study investigates the mechanisms proteins use to resist structural changes upon mutations, as a groundwork to understand functional robustness. Assuming that all proteins bear mutations by similar mechanisms, a case of study is a good model of investigation. The first challenge is to elaborate a set of mutations producing structural perturbations still maintaining a viable structure to look at. The solution proposed is to mutate *in silico* every amino acid of the interface of the B subunit pentamer of the cholera toxin and to monitor structural changes *via* a network model of the interface. A network representation is interesting because it allows measuring local to global changes and to investigate the capacity of proteins to cope with perturbation [74]. The relevance of network models in the study of structures for protein dynamics is now well established [7, 9, 14, 18, 19, 32, 33]. The second achievement is the AAR algorithm which quantifies all structural changes between wild-type and mutant structures by simply counting the changes in their number of atomic interactions. AAR is fast (less than one second for a protein of 103 amino acids), thorough and applicable on the Cartesian coordinates of any atomic structure.

One novel finding is that structural changes follow a cascade mechanism where the local reorganization of the atoms at the site of the mutation disturbs the chemical neighbors of the mutated residue which in turn disturb their chemical neighbors, as in a domino effect. What triggers the cascade is not yet identified but it is neither the degree nor the weight of the original

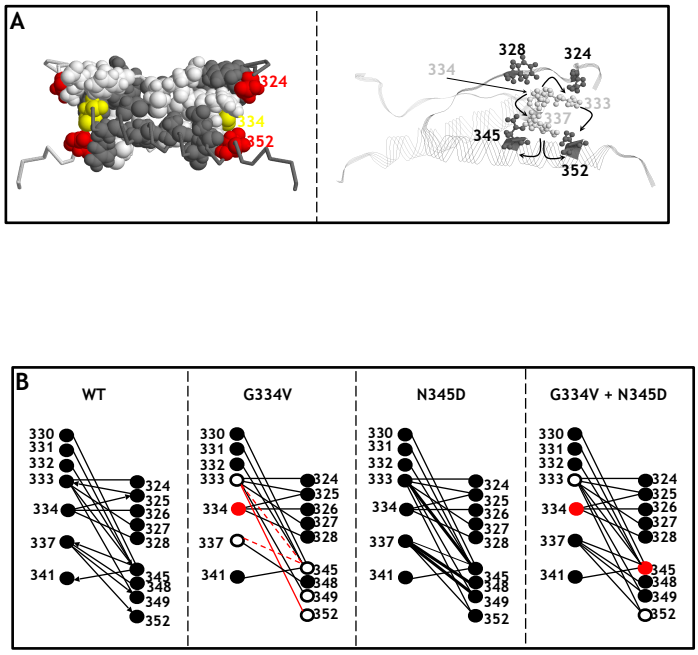


Figure 2.6: Non-additive *in silico* mutations G334V and N345D in the p53 tetrameric domain. (**A**) p53 WT. Left panel: The chains B (light grey) and D (dark grey) of the WT p53 are shown in backbone representation (PDB code 1SAK) except for the residues of the sphere of influence of the mutation G334V, space-filled. Right panel: As on left panel, but with a strand representation except for the residues indicated in ball-and-sticks. The cascade of changes is illustrated by arrows. (**B**) Spheres of influence of networks belonging to WT, G334V, N345D, and G334+N345D. Legend as in Figure 2.4. The mutated residues are in red. The open circles are the residues whose degrees are modified by the mutation. Arrows illustrate the path of structural changes going from the residue 334 to the residue 352. The red lines are for added (continuous) and depleted (dotted) links of amino acids. Black thick and thin lines are for increased and decreased weights, respectively.

residues, nor the fraction of local changes. This differs from networks where perturbations propagate through hubs (highly connected nodes) [6]. Instead, the changes propagate stepwise from hotspot to hotspot, from the site of the mutation to its neighbors (local change) to the rest of the protein (global change). This cascade mechanism results in major changes in interactions stretching out to large distances, or to more subtle changes. As mentioned already, the former are consistent with pathological mutations, while the latter accommodate adaptability and emergence of new functions through structural rearrangements which do not completely modify the protein conformation [42]. A cascade mechanism is also consistent with allostery, although multiple perturbations—as found in binding—are not tested here [67]. The cascade mechanism is more reliable than propagation of changes through hubs in a network with a power law distribution (few hubs, many low degrees) because it tallies with experimental evidences on the functional impact of mutations. In a hub-regulated network, the mutation of hubs would lead to massive changes, and pathologies; the mutation of residues with low degree would lead to local changes and explain robustness [29,39]. Yet, it would be difficult to account for the emergence of new function through few subtle changes as well as for adaptive mutations (non-additive mutation effects), since there would be little or large changes. Moreover, proteins do not have hubs in terms of having nodes with a significantly higher degree than other nodes; they have nodes with average degree [74].

The second novelty is the mechanism of robustness through alternative structures, rather than just unchanged structures. This fits the updated definition of protein function: an ensemble of conformations [45]. This also lays the ground for adaptability because it allows for non-additive effects, error corrections or epistasis [12, 13]. The presence of backup links in the WT network, which allows addition and depletion of links without altering substantially the layout of the network, might be a clue for identifying what triggers the cascade. Backup and alternative solutions are a current mechanism of robustness, reported for other real networks such as peer-to-peer networks or other biological networks [47, 77].

In summary, the extent of structural changes produced by mutations does not depend on the degree of the mutated residue, and it does not condition the impact of a mutation on the structure. The impact of mutation involves more complex mechanisms which remain to be deciphered [38]. Altogether, the mechanisms of structural changes observed through an *in silico* approach are consistent with all known functional effects of mutations (robustness, innovation, adaptation and pathology) supporting the approach as well as the hypothesis that structural robustness is embedded in the structure of the

protein.

2.7 Supplementary Information

The supplementary information contains: supplementary methods, supplementary Figure 2.7 and the Amino Acid Rank pseudocode.

2.7.1 Supplementary methods

FoldX

The run parameters are as follows: The PDB file of $CtxB_5$ (code 1EEI) and only two chains (Chains D and E) are considered for generating the mutation. Backbone and side chain atoms were allowed to move as well as neighboring atoms upon mutation (parameter "moveNeighbours" is set to "True"). The option "pdbHydrogens" was set to "False" so that hydrogen atoms were not included in the PDB output, because the positions of hydrogen atoms are not always calculated in a X-ray structure. Temperature was set to 298 K, the pH to 7.0, the ionic strength to 0.05 M. The "crystalwaters" parameter was set to "True", so crystallographic water bridges were considered in the PDB output if available in the crystal. Paramter "OutPBD" was set also to "True" in order to generate a mutated PDB file. Finally, "Complex with DNA" was set to "False".

2.7.2 Amino Acid Rank (Pseudocode)

The function SpectralPro is used to obtained the hotspot network with node set V and link set E. "SpectralPro" is used also to obtained the weighted adjacency matrix W of the hotspot network (Listing 2.1). The function "arank$_r$" computes the structural changes induced by the single mutation at sequence position r. The terms wt and mut are used to refer to the wild type and mutated weighted adjacency matrices, respectively. For a given sequence position r, it returns the values arank$_r$ and local$_r$ (Listing 2.2). Finally, the function Backup calculates for each link of the hotspot network its number of backup links. Finally, the function "Backup" obtains the number of backup links given a hotspot network and one link (Listing 2.3).

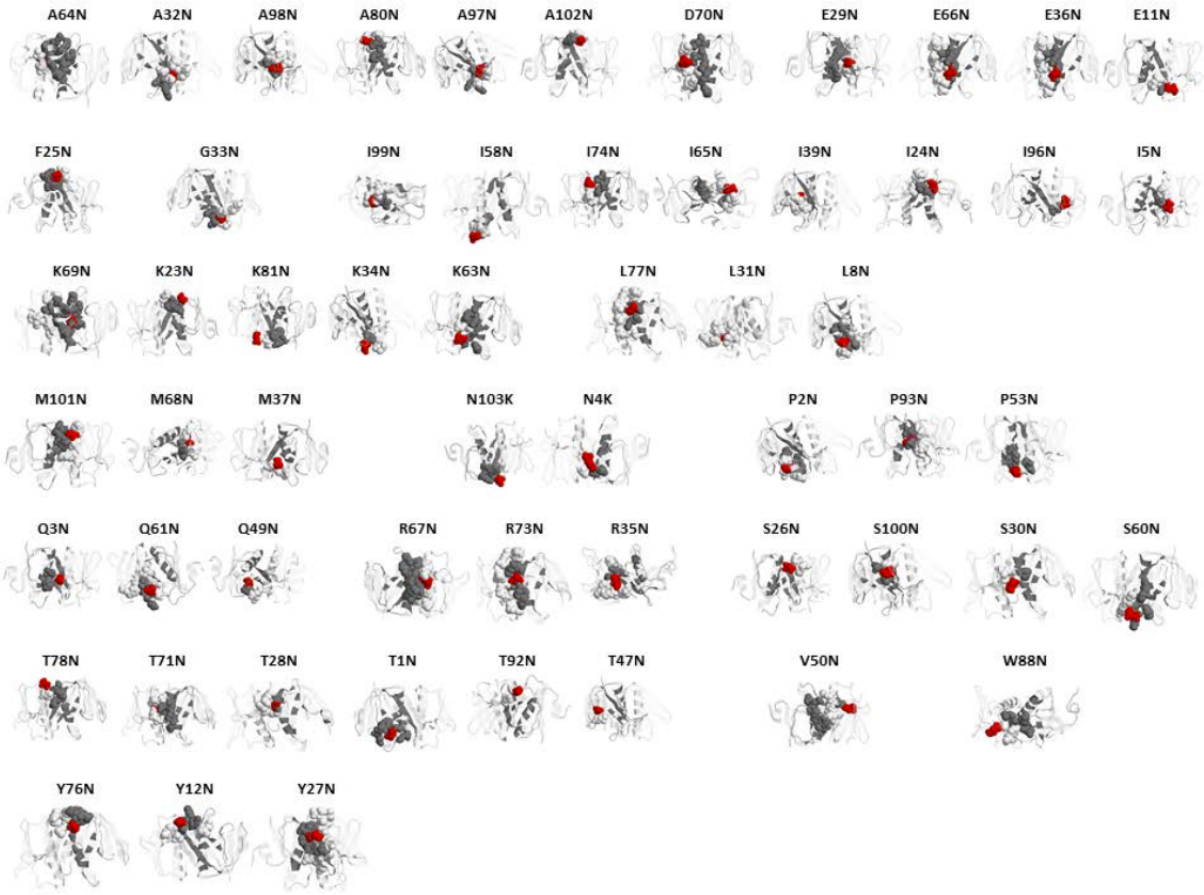


Figure 2.7: Spheres of influence as seen on the X-ray structures of the 58 mutants. Two adjacent chains of the $CtxB_5$ are shown in strands, the interface is shown in dark and light grey ribbons to distinguish both chains (PDB code 1EEI). The hotspots modified by the mutation are shown in space-fill, the mutated residue in red. The mutations are ordered per type of amino acids and decreasing $arank_r$ values within each type.

```
WRITE 'Input protein'
C = 5
E = list()
V = list()
W = dict()
FOR chain in protein DO
  FOR i in chain Do
    For j in (chain\i) DO
      a = set((a1, a2) for a1 in i and a2 in j)
      b = set(x for x in a if dis(x[0],x[1])<C)
      IF len(b) > 0 DO
        Append (i,j) to E
        Extend (i,j) to V
        W[(i,j)] = len(b)
      ELSE
        W[(i,j)] = 0
      END IF-ELSE
    END FOR
  END FOR
END FOR
RETURN V,E,W
```

Listing 2.1: Function SpectralPro.

```
WRITE 'Input MUT AND WT'
WRITE 'Input position r'
N = size(MUT) = size(WT)
A = array((N, N))
FOR (i,j) in N x N DO
  FOR i in chain Do
    A[i][j] = MUT[i][j] - WT[i][j]
  END FOR
arank_r = sum(abs(A))
local_r= sum(abs(A[r]))
RETURN arank_r, local_r
```

Listing 2.2: Function arank_r.

```
1  WRITE 'Input hotspot network H'
2  WRITE 'Input link (i,j)'
3  B = 0
4  E = edges(H)
5  V = nodes(H)
6  V_i = list(v for v in V if (chain(v)==chain(i)))
7  V_j = list(v for v in V if (chain(v)==chain(j)))
8  C_i = list(r for r in V_i if abs(r-i)<=4)
9  C_j = list(r for r in V_j if abs(r-j)<=4)
10 FOR (u, v) in E DO
11   IF u in C_i and v in C_j DO
12     B = B + 1
13   ELIF v in C_i and u in C_j DO
14     B = B + 1
15   END IF-ELSE
16 END FOR
17 RETURN B
```

Listing 2.3: Function Backup.

Chapter 3

Protein structure plasticity: the Neighborhood Watch

Abstract

Proteins possess qualities of robustness and adaptability to perturbations such as mutations, but occasionally fail to withstand them, resulting in loss of function. Herein the structural impact of mutations is investigated independently of the functional impact. Primarily, we aim at understanding the mechanisms of structural robustness, pre-requisite for functional integrity. The structural changes due to mutations propagate from the site of mutation to residues much more distant than typical scales of chemical interactions, following a cascade mechanism. This can trigger dramatic changes or subtle ones, consistent with a loss of function and disease, or the emergence of new functions. Robustness is enhanced by changes producing alternative structures, in good agreement with the view that proteins are dynamic objects fulfilling their functions from a set of conformations. This result, robust alternative structures, is also coherent with epistasis or rescue mutations, or more generally with non-additive mutational effects and compensatory mutations. To achieve this study, we have developed the first algorithm, referred to as amino acid rank (aar), which follows the structural changes associated with mutations from the site of the mutation to the entire protein structure and quantifies the changes so that mutations can be ranked accordingly. assessing the paths of changes opens the possibility to assuming secondary mutations for compensatory mechanisms.

3.1 Introduction

Proteins appeared 3.8 billion years ago, illustrating the resiliency of amino acid interactions through time and conditions [49]. This is also revealed by protein half-lives which cover orders of magnitude, ranging from minutes to days, even years (e.g. collagen), depending on function [69]. The outstanding functional resiliency of proteins makes the understanding of their design particularly worth investigating.

Such a resiliency relies on the output of perturbations such as mutations or changes in the environment, on protein function. Upon perturbation, a function can be reproduced (Functional robustness), lost (Functional failure) or modified (Functional innovation). Thus a protein is robust and adaptable. How can that be?

The function is encoded in a protein sequence, which is translated into a functional structure. The functional robustness is based on the fact that a sequence modification, namely a coding "error", doesn't imply a functional error. Several sequences encode the same structures (e.g. porins) and the function of a protein is resistant to the mutations of most of the protein amino acids. So a sequence "error" does not mean a structural "error". Moreover, several structures fulfill the same function so a structural "error" (structural modification) does not mean a functional "error" (e.g. Hemoglobin). Thus, proteins maintain their function despite sequential and structural "errors", i.e. despite sequence and structural changes.

Functional innovation and adaptability rely on this mechanism of robustness, which introduces differences without functional drawback, because the differences lead the modified proteins to a distinct fate upon subsequent mutations, than the unmodified counterpart, as seen for adaptive mutations. The differences also protect proteins from functional failure by preventing functional damages otherwise occurring on the unmodified version, as observed on rescue mutations [13].

Yet, a sequential error such as a single mutation is sometimes enough to destroy protein structure and function. So how to discriminate functionally "bad" sequential errors from functionally tolerated ones? As a pre-requisite, we first address the question of how to discriminate sequential errors that lead to structural errors from sequential errors that reproduce identical structures. In particular, we investigate what properties the structural design has to cope with sequential errors. Several amino acids encode the structural information so one possibility for structural reproducibility despite sequence errors lies in the structural information redundancy. Evidence supports this, for instance, alpha or beta secondary structures are encoded by different sequences, such

that not every mutation leads to secondary structure changes. Along the same line, we have measured backup interactions within protein structure [1]. A second possibility is that amino acid substitution is structurally tolerated because different amino acids reproduce the same local structure. In other words, there exist alternative amino acid solutions for the same structure. Mutations modify amino acid interactions without necessarily impacting the underling structure, supporting that hypothesis [1]. In fact, the impact of mutations on amino acid interactions is a relevant framework to explain structural robustness, failure and structural rescue/adaptability [74].

In the present work, we investigate this second possibility further and determine to what extent amino acids can be swapped without impacting protein structure. We identify a design rule that conveys high structural integrity. The atomic and amino acid interactions of 736,149 amino acids with their spatial neighbors have been surveyed from a database of 750 protein structures. The result shows that amino acids, regardless of type and position in a structure, interact with their spatial neighbors through a similar moderate average number of atomic interactions. We validate the hypothesis that structural robustness to mutations relies on the reproducibility of such moderate average number of atomic interaction at the site of the mutation by the amino acid substituent. This condition is achieved by adjusting neighbors to the residue and reciprocally, *via* amino acid alternative solutions or compensatory mutations.

3.2 Results and Discussion

The idea is to determine the role of amino acids on structural reproducibility. In order to understand the impact of individual amino acids on amino acid interactions and protein structure, we propose to compare the capacity of interactions of the twenty amino acids.

To do so, a database of 736,149 amino acids is built from the X-ray structures of 750 oligomeric proteins, and the interaction capacity of amino acids is assessed by measuring the number of amino acid neighbors and the number of atomic interactions per amino acid pair and per residue. Protein structures are modeled as a network of amino acids in interaction, and we zoom in that global network onto the 3D-local structure constituted of every single amino acid -i- of the protein and its -k- amino acid neighbors, -j_k- (Methods Section 3.4 Subsection 3.4.2) [1]. Thus, a local amino acid network describes a 3D-local structure. The neighbors -j_k- are the amino acids having at least one atom within a 5 Å distance of at least one atom of -i-. Thus,

the amino acid neighbors are the residues close in space to the residue -i- and not only its neighbors along the sequence. The nodes of the local networks are the amino acids -i- and -j_k-, and the links are the atomic interactions between them. The link between two amino acids is weighted with the exact number of distances between their respective atoms, which falls within 5 Å. Thus, the pairwise weight, referred to as w_{ij}, is the number of atomic interactions between two amino acids. The weight of a residue -i-, referred to as w_i, is the total number of its atomic interactions, namely, the sum of the w_{ij} over all -j_k- neighbors. In summary, each residue -i- is described by its number of amino acid neighbors, referred to as the degree k_i, its total number of atomic interactions, referred to as the weight w_i, and the pairwise atomic interactions (number of atomic interactions between two amino acids), referred to as w_{ij}. The local networks that modeled the 3D-local structures of amino acids in a protein structure at a given position, are also referred to as the amino acids pairs (i, j_k).

3.2.1 Amino acid diversity in terms of amino acid neighbors—Degree statistics

The Figure 3.1 is a statistical percentile representation of the degree versus amino acid type (Subsection 3.4.4). Briefly, for each amino acid type, the degrees are ranked and divided in 100 equal parts and we look at the degree adopted by 5, 50 or 95% of the amino acids as well as the highest (max) and lowest (min) degrees. For example, 95% of Gly adopt degrees equal to 13 or lower while only 5% have degrees equal to 5 or lower. In fact, 90% of the amino acids adopt intermediate degrees between 9 and 19 for the biggest residues (Trp, Phe and Tyr) and between 5 and 17 for the others. Thus, most amino acids are never fully covered with neighbors or depleted of neighbors (Figure 3.1a). This raises the question of the relation between the degree and the burying of residues and the simple assumption that high and low degrees are buried and surface-exposed residues, respectively. We have estimated the percentage of buried residues in the dataset by approximating proteins as a torus, because they are oligomers and cannot be approximated as a sphere like globular protein monomers. Various chain and oligomer diameters are considered in the simulation (Subsection 3.4.5). The result indicates that protein oligomers have at least 17% buried residues, and at most 75%, depending on their respective sizes (Supplementary Figure 3.7). Hence buried (>17%) does not imply maximum number of neighbors (0.3% of max degree), and most buried residues have intermediate degrees. This is confirmed by the box plot of the degrees of buried residues, monitored by

the Accessible Surface Area (ASA) equals to zero, which shows that buried residues cover a large range of degrees (Figure 3.1b).

Consequently, regardless the amino acid types and the position in the structure, residues, even buried, tolerate empty space locally, namely on their surface. There is no simple correlation between the degree of residues and their ASA since every degree covers a range of ASA as illustrated for a Val residue on Figure 3.1c. The same is true for the nineteen other amino acid types (Data not shown). ASA variations for amino acids have been reported previously [58]. Thus, the degree does not reflect the structural position of a residue in terms of ASA, a global measure, but rather provides information on the local void on the surface of the residue. Cavity and voids have been previously described and given several biological roles in enzymatic activities or protein flexibility [16, 36, 53, 54, 59].

On the other hand, measuring ASA and degree, allow assessing the surface distribution of amino acid neighbors (Figure 3.1c, box). For example, two Val residues with a degree 12 and almost identical weights, 109 and 107, respectively, have nevertheless two different ASA. For the ASA equals to zero (buried residue), the twelve neighbors are distributed almost uniformly on the surface of the residue while for the ASA equals to 72 (surface exposed residue), the twelve neighbors are distributed on only half of the surface. Thus, the two residues have different amino acid surface packing and local void, the density of amino acid neighbors is lower for the buried residue than the surface exposed. We are currently investigating such local void measures and their role on structural changes upon mutation.

Regardless of geometrical and chemical properties, all amino acids have degrees ranging from 5 to 17 (Figure 3.1a, solid box). This could mean that amino acids could replace one another and still maintain the same number of neighbors. The biggest residues, Phe, Tyr and Trp adopt degrees outside of the solid box as well, and mutations in such sub-networks might impact the 3D-local structures, decrease the degree and introduce more void locally. More generally, amino acids with local networks having degrees below ten or above fourteen, might be more susceptible to mutations because they can be replaced by less amino acids (Figure 3.1a, dashed box). For instance, replacement of Gly, Asp or Pro residues by Met, Phe, Tyr and Trp residues, is likely to affect the 3D-local structure because generally the former adopt lowest degrees and the latter the highest.

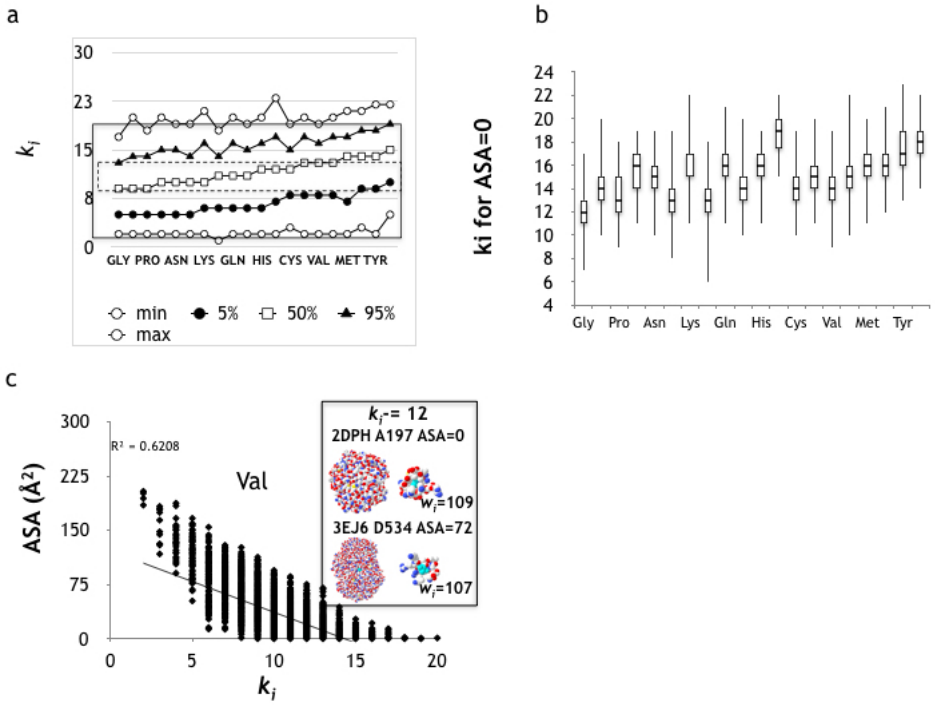


Figure 3.1: (**a**) Statistical percentile representation of the degrees adopted by the twenty amino acids. Min and max degrees correspond to the lowest degree and the highest degree observed for a residue type in the dataset, respectively. The percentage of amino acids adopting these degrees is not indicated on the plot, it remains below 0.3% for all residues. The continuous box represents the maximum overlap between the highest min and the lowest max degrees of the twenty amino acids while the dashed box represents the maximum overlap between the highest degree adopted by 5% of the residues and the lowest degree adopted by 95% of the twenty amino acids. (**b**) Degree for buried amino acids. (**c**) ASA versus degree for Val residue.

3.2.2 Amino acid diversity in terms of number of atomic interactions—Weight statistics

On Figure 3.2, the weights adopted by Gly, Glu and Trp residues are plotted against the degrees, as examples of the smallest, an average size and the biggest residue (Figure 3.2a, 3.2b & 3.2c, respectively). The min, average and max weights of each degree are plotted. The same plots for the other seventeen residues are shown in Supplementary Figures 3.8 to 3.13. The Figures 3.2 and 3.8 to 3.13 show that each degree adopts a range of weights, indicating that amino acids also have many alternative neighborhoods in terms of number of atomic interactions. To illustrate the neighborhood variability, X-ray and local network representations of the 3D-local structures of amino acids are depicted for a min, a mode (most frequent) and a max degree (Figures 3.2 and 3.8 to 3.13). As for the degree, the average weights are significantly lower than the maximum weights, and above the minimum weight, confirming that amino acids tolerate local voids and adopt moderate atomic packing.

To measure the sets of weights and degrees cover by the amino acids, the envelope bordered by the min and max weights of each individual degree is computed (Subsection 3.4.7). On Figure 3.2a, 3.2b and 3.2c, the envelopes for Gly, Glu and Trp, are shown; Gly is reproduced on Figure 3.2b and 3.2c for comparison.

To measure the overlap of weights and degrees between amino acids, the highest weight among the twenty smallest and the lowest weight among the twenty largest are taken (Figure 3.3a). The twenty amino acids cover more degrees and weights than they share (Figure 3.3a, thick line area), yet the intersecting envelope represents a significant overlap, and 58% of the total sampling. The intersecting envelope describes the 3D-local structures having local networks with degrees between 4 and 14, weights between 60 and 140, which can potentially be reproduced by any amino acid.

The amino acids can be classified into four groups based on their envelopes (Figures 3.1 and 3.8 to 3.13). Gly, Ala and Cys have the smallest envelopes. Pro, Val, Thr and Ser are just after, followed by Ile, Leu, Asp, Met, Lys, Asn and Gln. His, Glu, Arg, Phe, Tyr and Trp have the largest envelopes. The envelope classification does not coincide with a chemical or a geometrical classification of the amino acids. The amino acids can also be classified into four groups according to their most frequent degrees (Figures 3.1 and 3.8 to 3.13). Pro, Asp, Gln and Glu have the lowest mode degree ($k_{\text{mode}} \sim 8$–9), followed by Gly, Ser, Thr and Asn ($k_{\text{mode}} \sim 10$–11). Ala, Val, His, Arg and Cys have intermediates modes ($k_{\text{mode}} \sim 12$–13) while

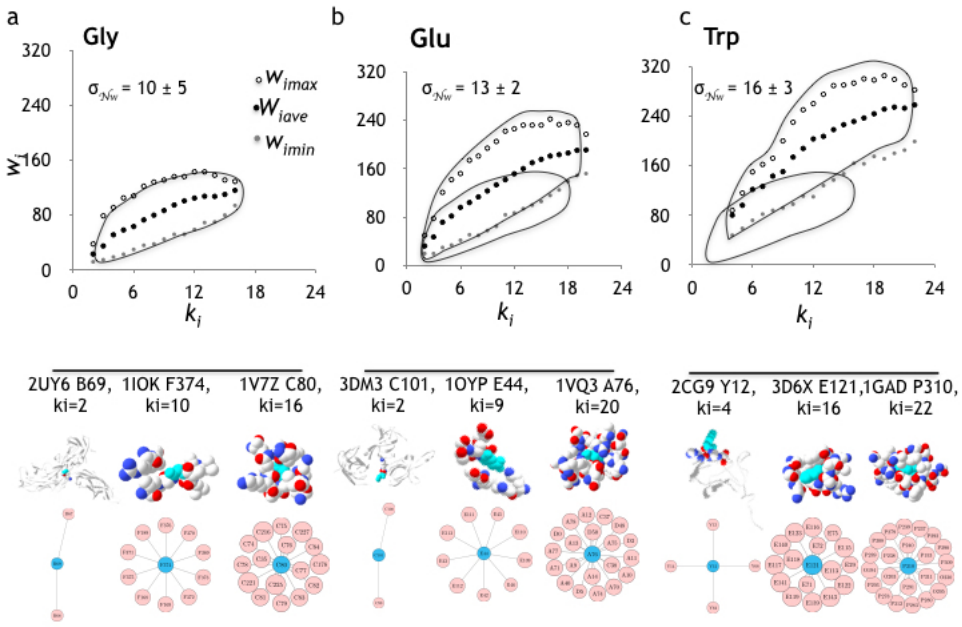


Figure 3.2: Amino acid capacity of interactions: Glycine (**a**), Valine (**b**) and Tryptophan (**c**). Upper panels: Weight versus degree of the amino acids. The continuous line shows the area covered by the set of degrees and weights adopted by the amino acids. The dashed line is the Gly area. Middle panels. Representations of the X-ray structures of Gly, Val and Trp residues for cases of min (left), most frequent (middle) and max (right) degrees. The whole protein is shown for the min degrees but only the 3D-local structures are shown for the most frequent and max degrees. The residue -i- is indicated in cyan and the neighbors -j_k- in CPK. The amino acids -i- and -j_k- are shown in space-fill. The figure is generated with sPDB viewer. The PDB code, the chain, the position of the residue along the sequence and its degree are given. Lower panels. Network representations of the degree cases shown in the middle panels. The residue -i- is indicated in cyan and the neighbors -j_k- in pink. The nodes (circles) are the residues and the links between amino acid pairs (lines) are based on the two residues having at least one atom each within a 5 Å distance.

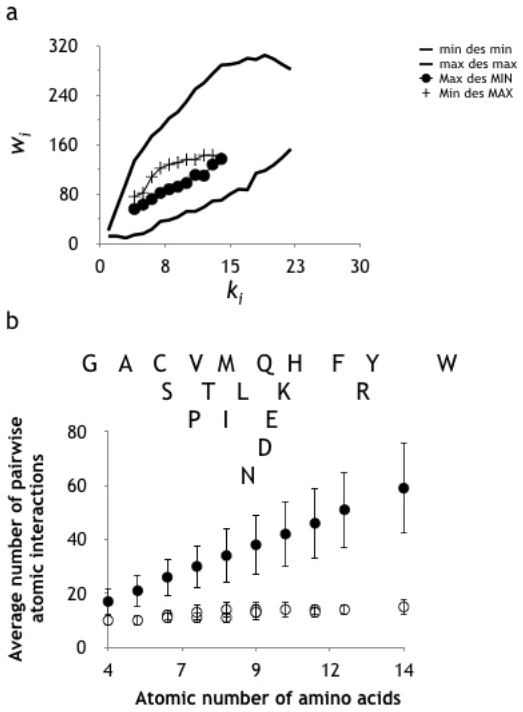


Figure 3.3: Neighborhood survey. (**a**) Overlap of degrees and weights. The area between plus and minus signs describes the degrees and weights shared by the twenty amino acids. The area described by the thick line is the entire set of degrees and weights adopted by the twenty amino acids, for comparison. The degrees and weights within the box are adopted by 48% to 92% of the amino acids (Percentages for Trp and Gly, respectively). (**b**) Supervision of the neighborhood. Average pairwise atomic interactions for observed $\langle w_{ij} \rangle_{obs}$ and theoretical $\langle w_{ij} \rangle_{th}$ data. The average $\langle w_{ij} \rangle$are plotted against the atomic number of the amino acids, for observed (white circle) and theoretical (black circle) data.

Ile, Leu, Met, Phe, Tyr and Trp have the highest mode ($k_{\text{mode}} \sim 14$–16). Again it does not correspond to the chemical or geometrical classification of amino acids. Thus, the interaction capacities of the amino acids do not mirror their individual chemical and geometrical features.

3.2.3 Average Pairwise Weights ($\langle w_{i,j} \rangle$)

Is the redundant interaction capacities shared by the twenty amino acids an indication that a mechanism corrects sequence errors, i.e. sequence changes, such that the local structure may be conserved?

We have reported for the same dataset that the frequency of amino acid pairs differs from the product of the frequencies of individual amino acids, suggesting a supervised pairwise amino acid matching [19]. We have also observed that the mutation of identical amino acid type led to different structural impacts, suggesting a role of amino acid neighbors in maintaining a local 3D structure [1]. A collective role of amino acids on structural regulation has been evidenced by other computational studies [13, 50]. It is possible that amino acid neighbors act as structural error correctors.

To test that possibility, the average number of pairwise atomic interactions $\langle w_{ij} \rangle_{obs}$, ratio of w_i and k_i, is computed over the database. The theoretical average number of atomic interactions, $\langle w_{ij} \rangle_{th}$, is also calculated for comparison based on a model where the interactions between amino acids are assumed to be independent of neighbors but pairwise dependent (k_i=1) and only limited by the number of atoms of each amino acid of the pairs, is used to validate or not a neighborhood supervision (Subsection 3.4.3). Both $\langle w_{ij} \rangle_{obs}$ and $\langle w_{ij} \rangle_{th}$ values are plotted against the atomic number of the amino acids for comparison (Figure 3.3b). The $\langle w_{ij} \rangle_{th}$ values appear far above the $\langle w_{ij} \rangle_{obs}$ values, ascertaining that the model where a residue matching depend only on its own properties is wrong. On the contrary, the results support the role of neighbors in controlling the local interactions of residues, by matching residues and neighbors. Accordingly, the parameter $\langle w_{ij} \rangle_{obs}$ is named the neighborhood watch ($\mathcal{N}_w$) in the rest of the paper.

The $\mathcal{N}_w$ values show remarkably little variability across degrees (Figure 3.2a and Figures 3.8 to 3.13, $\mathcal{N}_w$). Likewise, $\mathcal{N}_w$ values show little variability over the twenty amino acids, $\mathcal{N}_w$ goes from 10 to 15, and is on average 13 with a standard deviation of 2. The moderate $\mathcal{N}_w$ values, regardless the amino acid type or position, suggest that the structures of our dataset are built on amino acid interactions that follow the Goldilocks principle: not too many links, not too few. The Goldilocks principle, already described as a natural selection process, is just the right level of complexity to have a

robust and adaptable system as well as collective behaviors.

Thus, a protein structure is built such that any position can be fulfilled by any amino acid, because positions on average do not involve an extremely low or an extremely high number of atomic interactions, conditions, which would have restricted the numbers of suitable amino acids. Thus, residues and neighbors are not matched randomly but are matched to converge to a moderate common $\mathcal{N}_w$ value at every position in a structure.

This led us to draw the following hypothesis: Structural robustness to mutation would depend on the reproducibility of the $\mathcal{N}_w$ value at the site of mutation by the amino acid substituent. The reproducibility could be achieved with the same -j_k- neighbors, which would be an alternative (i', j_k) amino acid pair solution to the wild-type (i, j_k) amino acid pair solution. The prime indicates mutation. It could also be achieved with some mutations of the neighbors either as alternative (i', j'_k) amino acid pair solutions or as compensatory (i', j'_k) amino acid pair solutions if the substituent -i'- introduced alone some structural default that would be corrected by mutating neighbors. This is consistent with rescue and adaptive mutations [3, 13]. The neighbors would be correcting structural errors.

To test this hypothesis, we use two AB_5 toxins, the cholera toxin B pentamer ($CtxB_5$, PDB ID: 1EEI) and the human heat labile enterotoxin pentamer ($hLTB_5$, PDB ID: 1LTR) which have superimposable atomic structures despite 17 positions out of 103 (per chain) with a different amino acid type [26]. The two pentamers have different stabilities and the two toxins follow different folding and assembly mechanisms to build the same final structure [35, 56]. Thus, the seventeen positions with different amino acid types are structurally and functionally robust to mutations but at least some regulate the folding and assembly paths of the two toxins. The comparison of the two toxins allows us to address the question of how amino acids regulate the structural response of the protein to mutations and the question of how amino acids regulates the conformational changes associated with a protein construction.

We have analyzed every position of the two toxins to determine if mutated and conserved positions reproduce similar degree, weights and $\mathcal{N}_w$ ($\langle w_{ij} \rangle_{obs}$ average number of pairwise atomic interaction per position), despite the mutations and if so, whether that is achieved with the same neighbors, or by mutating the neighbors, or both. In the latter case, we then investigated compensatory mechanisms.

For the sake of simplicity, we consider $CtxB_5$ as the reference and $hLTB_5$ as a mutated version. Again, the prime is used to indicate a mutation that is a position with a different amino acid type in $hLTB_5$. A residue with

"mutated" neighbors in $hLTB_5$, has -j'_k- neighbors whether it has one or more than one "mutated" neighbors.

The degrees and weights of each amino acid of the two toxins are computed from their respective local networks modeled out of the toxin X-ray structures ($CtxB_5$, PDB code: 1EEI; $hLTB_5$, PDB code: 1LTR) [17, 41] (Supplementary Table 3.2). We observe that 49% of the conserved positions have degrees k_i and weights w_i shared by the twenty amino acids (Figure 3.3a, intersecting envelope), suggesting that half of the conserved positions and a fourth of all positions, could be mutated by any other type of residues without much impact on the degree and weight of the 3D-local structures (Table 3.2, *). Among the seventeen mutated positions, 41% have degrees and weights within the intersecting area, such that mutated positions do not seem to be more—or less—susceptible to mutations than conserved positions. Trp88 adopts the highest degree in the structure, with a weight 187 and 217, for $CtxB_5$ and $hLTB_5$, respectively. These weights are below the weight only achievable by the Trp residues and therefore even Trp88 could be mutated because some other amino acids are capable of reproducing similar weight and degree. The highest weights are for positions Arg67 (k_i=17, w_i=214) and Tyr76 (k_i =18,w_i=246) for $CtxB_5$ and $hLTB_5$, respectively, and again such weights and degrees are potentially reproducible by several other amino acid types. Thus, the toxins have no position with degrees and weights that cannot be reproduced by some other amino acids, and as such the toxins have no apparent positions structurally fragile to mutation.

On average over the mutated positions, the degrees vary with Δk_i equals to 1.5 ± 1.4; the weights with Δw_i equals to 14 ± 13 (Table 3.2). The Δk_i maximum is 4, for residues at positions 25, 75 and 80. The Δw_i maxima are 39 and 40 for residues at positions 75 and 80, respectively. There are also little differences between the two toxins $\mathcal{N}_w$ values over the mutated ($\Delta\mathcal{N}_w = 1.2 \pm 1.0$) or conserved ($\Delta\mathcal{N}_w = 1.0 \pm 1.1$) residues, respectively. The maximum $\Delta\mathcal{N}_w$ is 4.3 (position 89) for conserved positions and 2.9 for mutated positions (position 18). The $\mathcal{N}_w$ values varies by a factor of 1.4 across the twenty amino acids and by a factor of 1.3 between the two toxins at any position, indicating that the two toxins do reproduce similar $\mathcal{N}_w$ values at all positions despite the mutations.

This supports our hypothesis, that structural robustness to mutations is related to the maintenance of similar $\mathcal{N}_w$ values at mutated positions but also at conserved positions that could have been impacted by mutated neighbors.

To check the hypothesis further, we compare the $\mathcal{N}_w$ values of the N-terminal domains of $CtxB_5$ and of the verotoxin-1 ($PTXB_5$, PDB code 2XSC), another AB_5 toxin, which has a different N-terminal 2D-structure

(Supplementary Figure 3.14) [65]. Over the first ten amino acids, the maximum $\Delta\mathcal{N}_w$ between $CtxB_5$ and $PTXB_5$ is 9 at position 4. $PTXB_5$ has two smaller amino acids than $CtxB_5$ and $hLTB_5$, at positions 4 and 7 such that the atoms of the two residues are too far apart to have a pair (4, 7) in $PTXB_5$, in contrast to $CtxB_5$ and $hLTB_5$. Hydrogen bonding between residues three to four residues apart is necessary to build α-helices, and the loss of interactions between residues 4 and 7 might therefore explain why the N-terminal of $PTXB_5$ is unable to build one. Failure to reproduce a $\mathcal{N}_w$ value reasonably close to the one observed in $CtxB_5$ at position 4 associated with a change of secondary structure in the N-terminal domain of $PTXB_5$, supports our hypothesis that the neighborhood watch regulates structural robustness and consequently structural transitions. Nevertheless, investigating structural transitions using the $\mathcal{N}_w$ parameter is beyond the scope of the present work.

The next step, is to determine how the $\mathcal{N}_w$ values are reproduced despite the mutation. We investigate whether the $\mathcal{N}_w$ values at mutated positions, are reproduced with the same neighbors in both toxins or with mutated neighbors, or both.

To compare the two toxin neighbors, we use a *Jaccard* measure, which computes for each position, the ratio of the number of identical neighbors to the total number of common and different neighbors (Subsection 3.4.8). The toxin amino acid sequence is plotted against the *Jaccard* measure (Figure 3.4). A *Jaccard* measure of 100% or slightly lower ($> 90\%$) is for residues with identical neighbors or identical neighbors but with different atomic proximity, respectively. *Jaccard* measures below 90% are residues with one or more mutated neighbors and environments significantly different. Over the 102 residues, 91 have mutated neighbors (Figure 3.4, squares) and only 11 have identical neighbors (Figure 3.4, circles). Among the conserved positions, 77 have mutated neighbors (Figure 3.4, white squares) and 8 have identical neighbors (Figure 3.4, white circles). Among the mutated positions, 14 have mutated neighbors (Figure 3.4, black squares) and 3 have identical neighbors (Figure 3.4, black circles). Thus most positions accommodate alternative -j'_k- neighbors as a result of the 17 mutated residues -i'-.

The three mutated positions (31, 38 and 44) which reproduce $\mathcal{N}_w$ with identical -j_k- neighbors can be assumed as amino acid alternative (i', j_k) pair solutions to the (i, j_k) pairs observed in $CtxB_5$ and reproduce $\mathcal{N}_w$ with alternative -j'_k- neighbors (Table 3.1). On the other hand, fourteen mutated positions reproduce $\mathcal{N}_w$ with alternative -j'_k- neighbors (Table 3.1). We wonder whether such multiple pair mutations reflect a compensatory mechanism necessary to reproduce the $\mathcal{N}_w$ values at the positions. In other words, are

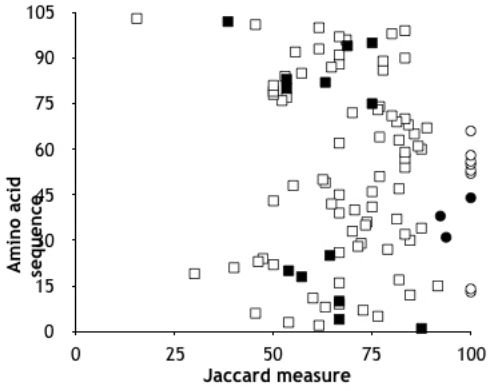


Figure 3.4: *Jaccard* measure. The composition and proximity of the amino acid neighbors of every residue of $CtxB_5$ and $hLTB_5$ are compared with a *Jaccard* measure (Subsection 3.4.8). The amino acid position in the sequence is plotted against the *Jaccard* measure, expressed in percentage. The black symbols indicate positions with two different residues in both the toxins and the white symbols indicate positions with identical residues. Circles are positions with identical neighbors in both the toxins and squares are positions with at least one different neighbor.

the mutations -i'- introducing structural local defaults (a significant change in $\mathcal{N}_w$) that are compensated/corrected by the mutations of some -j'_k- neighbors? Or are the (i', j'_k) pairs just amino acid alternative to the pairs (i, j_k)?

3.2.4 Pairwise network compensation

First, we investigate whether the mutations introduce geometrical and/or chemical perturbation that could have significantly modified $\mathcal{N}_w$, without additional neighbor mutations. The geometrical and chemical properties of (i, j_k) and (i', j'_k) pairs in $CtxB_5$ and $hLTB_5$, respectively, are compared to test this possibility (Table 3.1). The volumes of the pairs account for the geometrical properties while the chemistry of the pairs accounts for the chemical properties. Out of the 14 pairs of mutated residues, only two are geometrically and chemically equivalent in both toxins, indicating that no geometrical or chemical pairwise compensatory mechanism is necessary for structural robustness. Even, structural robustness appear largely tolerant to chemical changes as illustrated by the introduction of two histidines at positions 18 and 94 in $CtxB_5$. The chemical perturbation at positions 18 and 94 is consistent with the experimental observation that $CtxB_5$ assembly is inhibited at pH 6.0, a pH close to the histidine pKa whereas $hLTB_5$ assembly is inhibited at pH 7.0, a pH close to a N-terminal pKa [56,80]. It is reasonable to assume that at low pH, His 18 and His 94 are protonated and that electrostatic repulsion prevents the two residues from interacting. In $hLTB_5$, there are a Tyr and an Asn at positions 18 and 94, respectively, residues with no susceptibility at low pH. Thus, the chemical perturbation would be structurally robust and would rather impact the folding and assembly paths.

Second, we investigate if the modifications of some pairwise atomic interactions (-w_{ij}-) introduced by the mutations -i'- and altering the $\mathcal{N}_w$ could be compensated by additional pairwise atomic interactions through the mutations of neighbors. To test that possibility, the pairwise weights -w_{ij}-, -$w_{i'j}$- and -$w_{i'j'}$- of the fourteen mutated positions having mutated neighbors are computed and compared between the two toxins (Supplementary Table 3.3). For the positions 10, 18, 20 and 82 (Figure 3.5, *), the $\mathcal{N}_w$ is reproduced because there are little differences in the two toxin pairwise atomic interactions (Δw_{ij} <6), and the multiple pair mutations are considered amino acid alternative (i', j'_k) solution.

The $\mathcal{N}_w$ is an average value, that is the sum of every pairwise atomic interactions at the position divided by the degree of the position. Thus, differences in atomic pairwise interactions introduced by a mutation -i'- can be compensated by differences in the atomic pairwise interactions of other

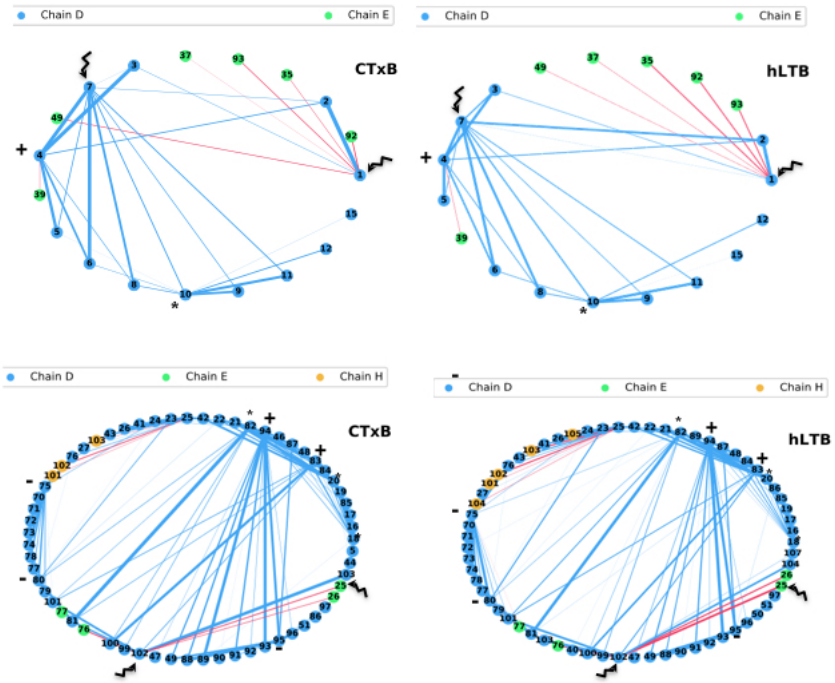


Figure 3.5: Graph representation of mutated networks. The networks are composed only of mutated positions and their neighbors. The circles are the nodes with the residue position indicated in it, the lines are the links between two amino acids measured by the atomic interactions. The thickness of the line correlates with the -w_{ij}- values. Circles of different colors indicate residues from different chains. The blue lines are intramolecular amino acid pairs, whereas the red lines are intermolecular pairs. Stars indicate positions with alternative pair solutions, thunder are positions with weight-compensatory mechanism, positions marked with a minus ('-') have a degree-compensatory mechanism. The positions not compensated but with -w_{ij}- variability not sufficient to jeopardize the structure are indicated by a plus symbol ('+').

pairs through -j'_k- mutations. Alternatively the differences in atomic pairwise interactions introduced by the mutation -i'- can be compensated by modifying the degree at the position. Accordingly, the positions 1, 7, 25 and 102 (Figure 3.5, thunder) are found to have pairwise atomic interaction modifications that compensate each other over all the pairs such that the $\mathcal{N}_w$ value is reproduced (weight-compensatory mechanism). This is illustrated on the graph representation of the networks of mutated positions -i'- and their neighbors (Subsection 3.4.9). The positions are indicated by circles with their residue number in them, and the atomic pairwise links by lines, whose thickness correlate with the number of atomic interactions. Altogether these positions in $CtxB_5$ have less atomic interactions in intermolecular pairs (pairs of residues that belongs to two different chains) (Figure 3.5, red color lines, positions 1, 25 and 102) or in pairs involving residues at the interface (Figure 3.5, position 7, pairs 1.7 and 7, 2).

The positions 75, 80 and 95 (Figure 3.5, –) have significantly less atomic interactions per pair in $CtxB_5$, compensated on average by a lower degree of the position (Table 3.3 and Figure 3.5).

Consequently, the positions might be less stable in $CtxB_5$. The positions 4, 83 and 94 are neither weight-compensated or degree-compensated, and have significantly more atomic interactions per pair in $CtxB_5$ as such they might be more stable in this toxin (Figure 3.5, +). Nevertheless the differences in the two toxins -w_{ij}- are not sufficient to modify the structure.

The pairwise network analysis highlights that $CtxB_5$ has less intermolecular interactions at the interface involving the N-terminal of the toxin, suggesting a lower stability of that pentameric interface. To check this hypothesis, $CtxB_5$ was treated at pH 8.4 to deprotonate the N-terminus of the toxin, and perturb it, and the pentamer stability was measured by SDS-PAGE analysis. Only at pH 8.4, $CtxB_5$ pentamer becomes SDS-sensitive in contrast to $CtxB_5$ at pH 7.4, which resists SDS-treatment (Figure 3.6a, lanes 2 and 4, respectively). At pH 8.4 as at pH 7.4, $CtxB_5$ is a pentamer as can be seen by the toxin cross-linking prior SDS-PAGE analysis (Figure 3.6a, lines 3). $hLTB_5$ is SDS-resistant up to pH 1031.

The pairwise network analysis also suggests a weaker $CtxB_5$ main interface, involving the interface residues 25 and 102 as well as weaker intramolecular positions 75, 80 and 95. The pair 75 and 80 constitutes a hinge that positions the beta strand 80 to 89 to the helix 75 to 70 and to the beta strand 25 to 18 through the amino acid pairs (25, 76), (24, 75), (22, 80), (23, 80) and (24, 80) (Figure 3.6b). A weaker interaction between residues 75 and 80, would make these 2D structure elements more mobile and consequently make their 3D-folding more difficult. In turn, the alignment of the main in-

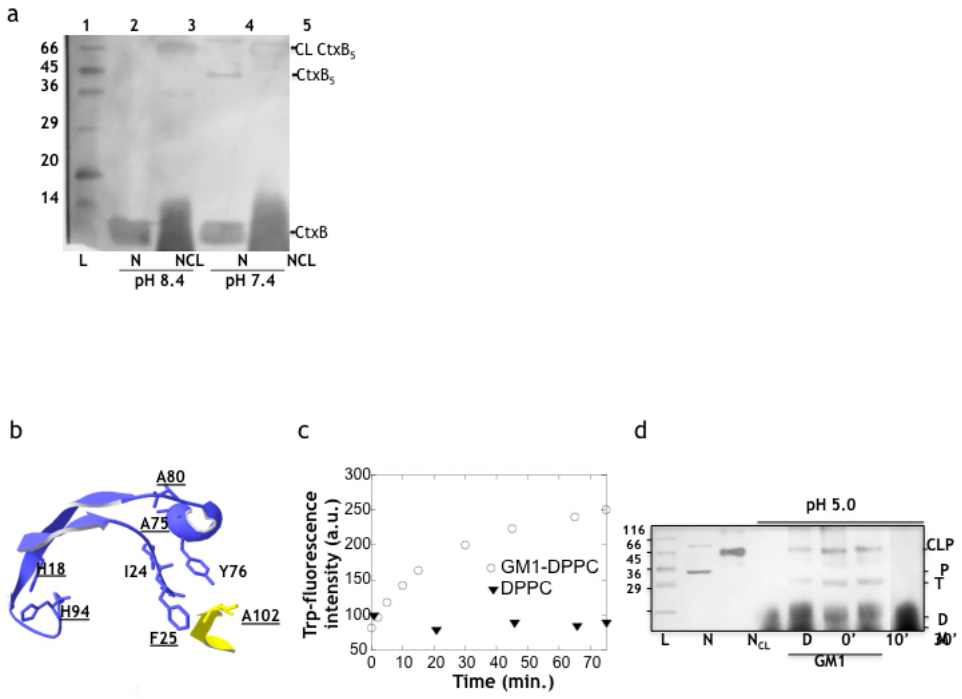


Figure 3.6: Mutations impact on stability and assembly mechanisms but not on structures. (**a**) Mutations impair $CtxB_5$ stability. SDS-PAGE analysis (Subsection 3.4.10). Treatment of $CtxB_5$ at pH 8.4, compared to pH 7.4, weaken the pentamer which becomes SDS-sensitive (lanes 2 and 4, respectively). The pH 8.4 does not however dissociate the pentamer, still observed on the gel if cross-linked before SDS treatment (lanes 3 and 5, respectively). CL and NCL stand for cross-linked and non cross-linked, respectively. L and N are low molecular weight marker and native $CtxB_5$, respectively. (**b**) Weak network pairs on $CtxB_5$ structure. A part of the $CtxB_5$ monomer is shown (PDB 1EEI). The residues 18 to 25, 75 to 94 of chain D are shown in blue ribbons and the residues 94 to 103 of chain E are shown in yellow ribbons. The residues 18, 94, 24, 25, 75, 76, 80 and 102 shown in sticks. (**c**) Mutations impair $CtxB_5$ assembly measured by fluorescence spectroscopy. Trp-fluorescence intensity signals are measured against time, after $CtxB_5$ dissociation at pH 1.0, return to pH 5.0 and addition of GM1-DPPC liposomes (white circle) or DPPC alone (black triangle). (**d**) Mutations impair $CtxB_5$ assembly measured by SDS-PAGE and cross-linking. As in c. with samples of $CtxB_5$ cross-linked immediately after returning to pH 5.0 (D), just after addition of GM1 (0'), ten (10') and thirty minutes (30') later. Same condition ten minutes after addition of DPPC alone (DPPC). Native (N) and cross-linked pentamers (NCL) are also shown on the gel. On the left is the molecular weight marker with the apparent molecular weights indicated. D, T, P and CLP stand for $CtxB_5$ dimers, trimers, pentamers and cross-linked pentamers. All samples at pH 5.0 are cross-linked prior SDS-Page analysis.

Table 3.1: Mutated pair features

Pair position	i, j	i', j'	V_i		$V_{i'}$	V_j		$V_{j'}$	$\lvert\Delta V_i\rvert$	χ^a	$w_{ij}, w_{i'j'}$	Structure
1, 7	> 5Å	Ala, Glu	-	-	92	-		155	-	-	0, 2	Secondary
4, 7	Asn, Asp	Ser, Glu	135	$>$	99	125	$<$	155	6	1	11, 25	Secondary
7, 10	Asp, Ala	Glu, Ser	125	$<$	155	92	$<$	99	37	0	11, 14	Secondary
18, 94	His, His	Tyr, Asn	167	$<$	203	167	$>$	135	4	0	6, 5	Tertiary
18, 20	His, Leu	Tyr, Ile	167	$<$	203	168	$\sim$	169	37	0	15, 20	Secondary
18, 83	His, Glu	Tyr, Asp	167	$<$	203	155	$>$	125	6	0	3, 3	Tertiary
20, 83	Leu, Glu	Ile, Asp	168	$\sim$	169	155	$>$	125	29	1	11, 12	Tertiary
83, 82	Glu, Val	Asp, Ile	155	$>$	125	142	$<$	169	3	1	26, 26	Primary
83,102	Glu, Ala	Asp, Glu	155	$>$	125	92	$<$	155	33	0	8, 11	Tertiary
20, 82	Leu, Val	Ile, Ile	168	$\sim$	169	142	$<$	169	28	1	11, 15	Tertiary
82, 80	Val, Ala	Ile, Thr	142	$<$	169	92	$<$	122	57	0	6, 5	Secondary
80,75	Ala, Ala	Thr, Thr	92	$<$	122	92	$<$	122	60	0	15, 24	Tertiary
95, 94	Ala, His	Ser, Asn	92	$<$	99	167	$>$	135	25	0	22, 24	Primary
102, 25	Ala, Phe	Glu, Leu	92	$<$	155	203	$>$	168	28	0	9, 12	Quaternary

[a] χ stands for chemical properties, 1 identical 0 different.

terface strands (25–31 and 96–102) would also be more difficult, even more so because of a weaker intermolecular interaction between residues 25 and 102. This suggests that $CtxB_5$ assembly is slower than $hLTB_5$ assembly because the 3D-folding of the beta strands 25 to 18 and 80 to 89 is more difficult. To test this hypothesis experimentally, we use two experimental results. First, the 3D-folding of the beta strands 80 to 89 and the beta strand 25 to 18 involves the pair (18, 94) which are two histidines in $CtxB_5$ (Figure 3.6b). As mentioned $CtxB_5$ assembly is inhibited at pH 5.0 certainly by the protonation of these two histidines. Second, GM1, the toxin cellular receptor binds to the toxin residues Trp88, Lys91, Tyr 12 and Asn14, cross-linking upon binding similar areas than the interaction between residues 18 and 94 and residues 22 and 80. If the amino acid pairing (His18, His94) is necessary for the interface formation but not for the 3D folding of the strand 80 to 94 and the strand 18 to 25, then addition of the GM1 in the reassembly solution at pH 5.0 would have no impact on the toxin assembly.

But on the contrary, despite pH 5.0, $CtxB_5$ reassembly resumes after addition of the GM1, but not after addition of liposomes containing no GM1 (Subsection 3.4.10). After addition of GM1, there is an increase of the Trp-fluorescence signal indicating reassembly (Figure 3.6c), confirmed by the presence of bands at the apparent molecular weights of $CtxB_5$ trimer and pentamer on a SDS-PAGE (Figure 3.6d). The result supports the hypothesis that $CtxB_5$ assembly is inhibited by the folding of the toxin monomer into

an assembly-competent state, as proposed from the network analysis.

The combination of theoretical and experimental investigations highlight an incredible collective behavior of the amino acid and their neighbors to regulate the role of a position in protein folding and in response to mutations.

3.3 Conclusion

Three network parameters monitored from the amino acid spatial position, the degree, the weight and the neighborhood watch, appear relevant to assess the structural robustness of a protein to mutation.

The pairwise geometrical, chemical and atomic interaction properties of a position appears a new reasonable tool to investigate the folding mechanisms of a protein.

Finally, the structural susceptibility of a protein to mutation (structural robustness, fragility and rescue) is based not on the individual properties of the amino acids but on the properties of the amino acids and their neighbors.

3.4 Methods

3.4.1 Database

A database of 736,149 amino acids is built from the atomic structures of 750 protein oligomers. All the amino acids within one type have an identical number of atoms, and therefore the same interaction capacity. All atoms are considered except for hydrogens, which are not detected by X-ray crystallography. Degrees observed in less than three proteins are not considered. The database is accessible online at the https://github.com/rodogi/biographs.

3.4.2 Amino Acid Network

For a given protein, we compute the Amino Acid Network (AAN) where the nodes are amino acids of the protein and links connect pairs of amino acids at distance less than 5 Å(Section). More precisely, a link between two amino acids at position -i- and -j- exists, if at least one atom of the amino acid -i- is at distance less than 5 Å from an atom of the amino acid -j-. For an amino acid at position -i-, its set of neighbors is noted -j_k- and is comprised of all the all amino acids with at least one atom at distance less than 5 Åfrom atoms of the amino acid -i-. The neighbors also refer to as the environment. Given an amino acid -i-, the environment of -i- is a

sub-network of AAN containing the node -i- with all its connections to -j_k-. A sub-network describes a 3D-local structure composed of any amino acid and its neighbors in the AAN.

We have constructed a database of the amino acid environments from a dataset of 750 proteins for statistical analysis of all environments (list in https://github.com/rodogi/biographs).

The number of links in the sub-network of an amino acid at position -i- is the degree of -i-, i.e. the degree k_i of -i- is the number of amino acids -j_k- in the environment of -i-. The weight, w_{ij}, of a link between two amino acids -i- and -j-, refers to the pairwise atomic interaction: it is the number of pairs of atoms, one in the amino acid -i- and the other in the amino acid -j-, that constitutes the link. The weight wi of a residue -i-, is the sum of all pairwise links incident to amino acid -i-.

The degrees and weights probably overestimate the number of amino acid neighbors and the atomic packing of amino acids because the radius of van der walls of atoms is ignored. Yet, because amino acids are composed of the same atoms, carbon, hydrogen (not included here), oxygen, nitrogen and sulfur (Met and Cys), and because in the dataset the residues have identical number of atoms within each amino acid type, the range of degrees and weights within one type and their comparison over the twenty amino acids is reasonable with such approximation.

3.4.3 Pairwise theoretical average number of atomic interactions

The theoretical average weight, referred to as $\langle w_{ij} \rangle_{th}$, is calculated only for a degree $k = 1$. For each type of amino acid, $\langle w_{ij} \rangle_{th}$ is calculated in two steps:

1. $\langle w_{ij} \rangle = \frac{n(n+1)/2}{n} = \frac{n+1}{2}$ with -n- equals to the product of the number of atoms of each amino acid of the amino acid pair (i, j). The hydrogen atoms are excluded.

2. The $\langle w \rangle_{th}$ is the average $\langle w_{ij} \rangle$ over the twenty possible pairs adopted by each amino.

3.4.4 Degree statistics

For each of the 20 amino acid types, we compute the sequence of degrees of their respective environments. The sequence is ordered from the smallest degree to the largest degree, “min” is the minimal value in the ordered

sequence and "max", the maximal. The median is the value in the middle of the ordered sequence (Figure 3.1, white square). In order to define the percentile, the ordered sequence is divided into one hundred equal parts.

The k-th percentile (k% value in Figure 3.1) is the value at the separation between the k-th and the $(k+1)$-th parts. The 5th percentile (Figure 3.1, black circle) is the value at the separation between the 5-th and the 6-th parts. The median is thus the 50-th percentile. In the preceding definition according to the splitting in a hundred equal parts, if the separation of the k-th and the $(k+1)$-th parts is not between one but two values, then the k-th percentile is the average of these two values.

3.4.5 Torus

In order to compare the number of amino acids on the surface of a protein and the number of amino acids inside the protein (called buried amino acids), we made a theoretical model. As proteins in the dataset are polymers their topology is a torus (a donut-shaped object). In order to define a torus, we need two quantities: the whole diameter $2R$ of the donut (from the two most opposite outside points) and the diameter $2r$ of the "tube" of the donut (from an outside point to its closest opposite point inside point on the tube). The area (that is the contact surface of the donut) is calculated with the usual formula for a torus, namely $4\pi^2 Rr \times 0.9$ where 0.9 is the density of spherical packing on the plane, because as a first approximation an amino acid is a sphere on the surface. The volume is computed with the usual volume of a torus namely $2\pi^2 Rr^2 \times 0.74$ where 0.74 is the spherical packing in space. With this computation the ratio of the number of amino acids of the protein and the number of amino acids on the surface of the protein is between 0.2 and 2 when r varies from 3 to 8 Å. This means that the donut-shaped model gives a large possibility of ranges: from a number of amino acids twice as large as at the surface to a number of amino acids 5 times bigger on the inside of the protein.

3.4.6 Accessibility Surface Area (ASA)

The Accessibility Surface Area was calculated using the server http://cib.cf.ocha.ac.jp/bitool/ASA/, which calculates the ASA of a protein given its PDB file. The calculation is done based on the algorithm of Shrake and Rupley [64].

3.4.7 Degree and weight Envelopes

Each amino acid is investigated in terms of degree and weight in order to assess the overlap between the twenty amino acids. First, we compute an array of all degrees and weights to obtain, for a given amino acid and a given degree, and we give, the minimal and maximal value of weights. Finally, for each type of amino acid, we consider the surface in the plot between the minimal and maximal weights. Second, for each pair of amino acids, we consider the intersection of their two surfaces obtained with their respective minimal and maximal weights. This intersection gives the range of possible weights that are common to both amino acids.

3.4.8 *Jaccard* measure

We made an algorithm to compare the environment of the two toxins. We start with a vector of 20 counters associated with the 20 amino acid types, and we initialize each counter with a value equals to 0. Given an amino acid -*i*-, the vector gives the number of occurrences of each amino acid type in the environment of -*i*-, e.g. if Val is three times in the environment of -*i*-, then the entry corresponding to Val in the vector is equaled to 3.

In order to compare the two environments, we calculate a *Jaccard* similarity measure on the pair of vectors. The *Jaccard* similarity is computed using the environment vectors as follows: The intersection of each entry of the vector, that is the number of occurrences in common in the two proteins for each amino acid type, e.g. if there are five Val in the environment of amino acid -*i*- in protein 1 and three in protein 2, then the intersection of the entry Val in the vectors is equal to three. There is one intersection value per amino acid type and the sum of the twenty intersection values is noted inter(-*i*-). Likewise, we compute the union of each entry in the two vectors and the sum of the union is noted union(-*i*-). The *Jaccard* measure for amino acid -*i*- is the ratio inter(-*i*-) to union (-*i*-). Note that the *Jaccard* measure is a value in the interval [0, 1] because inter(-*i*-) $<$ equal union (-*i*-).

If *Jaccard*(-*i*-) equals to 0, this means that inter(-*i*-) equals to 0 and the environments of -*i*- of the two proteins are either composed of 0 or do not share an amino acid type in common. If on the other hand, *Jaccard*(-*i*-) equals to 1, then the two environments are identical. The X-ray structure of $hLTB_5$ (1LTR) contains a longer C-terminus than the $hLTB_5$ used for experimental studies. The extra amino acids modify the residue 103 degree and weight, and the *Jaccard* measure compared to the residue 103 in $CtxB_5$, but the two toxins have identical C-terminal ends in the experimental studies.

Because of this difference, the amino acid at position 103 is ignored in the theoretical analysis.

3.4.9 Mutated networks

The amino acids having a different composition in both toxins constitutes a sub-network refers to as a mutated network. It is represented as a graph where the nodes are mutated positions and their neighbors and the links are the atomic pairwise interactions -w_{ij}-. Python is used to represent the graphs with a color per chain and a link thickness correlating with the -w_{ij}-.

3.4.10 Experimental methods

Reagents and buffers—Cholera toxin B pentamer ($CtxB_5$) and all other chemicals were obtained from Sigma. GM1 and DPPC were bought from Avanti. McIlvaine buffer (0.2M disodium hydrogen phosphate, 0.1 M citric acid, pH 5.0–8.0), PBS and 0.1 M KCl/HCl at pH 1.0 were used. All buffers were filtered through sterile 0.22 μm filter before use.

SDS-PAGE analysis—SDS-PAGE (15%were performed with a Bio-Rad mini-Protean 3 system using the Laemmli method [31]. The gels were silver stained. 1 mg of sample was loaded on each lane of the gel.

$CtxB_5$ pH-pentamer stability. Briefly, lyophilized native $CtxB_5$ was dissolved in PBS at a concentration of 344 μM and was diluted in micropore water at pH 8.4 at a final concentration of 8.6 μM. The pentamer stability was measured by SDS-PAGE combined with cross-linking analysis.

Reassembly of $CtxB_5$—The conditions used for reassembly were adapted from elsewhere [35]. Briefly, native $CtxB_5$ was acidified in 0.1 M HCl/KCl at pH 1.0 for 15 min at a final toxin concentration of 86 μM. The toxin was subsequently diluted to a final concentration of 8.6 μM in McIlVaine buffer at pH 5.0. The sample was incubated for 30 min at 23°C before addition of GM1-DPPC (10% w/v GM1) liposome or of DPPC (Dipalmytoil phosphatidyle choline) liposome. Liposomes were prepared by reverse phase separation and used at a final concentration of 10 mM [73]. The samples were analyzed by SDS-PAGE combined with cross-linking after 30 min at pH 5.0 and at different times after addition of the GM1-DPPC liposome. In addition the Trp-fluorescence intensity of the samples was measured just after addition of the liposomes.

Chemical cross-linking and SDS-PAGE analysis of the oligomeric state of $CtxB_5$—The cross-linking conditions were adapted from elsewhere [35]. The

cross-linking of $CtxB_5$ samples allows detecting $CtxB_5$ assembly intermediates (dimer, trimer, tetramer and non SDS-stable pentamer).

Trp-fluorescence—The Trp-fluorescence method was adapted from elsewhere [35]. Fluorescence measurements were performed using a Cary eclipse Varian spectrofluorimeter. Excitation was at 295 nm, with emission recorded at 349 nm and slit widths of 2.5 and 10 nm for excitation and emission, respectively.

3.5 Supplementary material

3.5.1 Supplementary Tables

p_i	1EEI	1LTR	k_i^{1EEI}	k_i^{1LTR}	k_i	w_i^{1EEI}	w_i^{1LTR}	w_i	$\langle wij \rangle_{obs}^{1EEI}$	$\langle wij \rangle_{obs}^{1LTR}$	$\langle wij \rangle_{obs}^{2XSC}$
1*	T	A	7	7	0	81	68	13	12.0	10.0	12.0
2*	P	P	10	10	0	118	123	5	12.0	12.0	14.0
3*	Q	Q	9	9	0	108	95	13	12.0	11.0	12.0
4*	N	S	7	7	0	131	115	16	19.0	16.0	10.0
5	I	I	17	14	3	151	151	0	9.0	11.0	10.0
6*	T	T	8	8	0	120	116	4	15.0	15.0	11.0
7*	D	E	9	9	0	128	125	3	14.0	14.0	11.0
8	L	L	16	15	1	147	144	3	9.0	10.0	8.0
9*	C	C	12	12	0	135	141	6	11.0	12.0	9.0
10*	A	S	8	7	1	84	88	4	11.0	13.0	12.0
11*	E	E	8	8	0	141	126	15	NaN	NaN	NaN
12	Y	Y	12	12	0	168	175	7			
13*	H	H	5	5	0	77	80	3			
14*	N	N	8	8	0	120	119	1			
15	T	T	11	12	1	153	151	2			
16	Q	Q	11	10	1	151	138	13			
17*	I	I	10	10	0	124	133	9			
18	H	Y	10	11	1	154	162	8			
19*	T	T	6	6	0	99	93	6			
20	L	I	12	10	2	145	152	7			
21*	N	N	7	7	0	123	102	21			
22	D	D	9	9	0	142	148	6			
23*	K	K	10	10	0	129	139	10			
24	I	I	14	15	1	140	139	1			
25	F	L	11	15	4	163	165	2			
26	S	S	10	10	0	145	140	5			
27	Y	Y	17	17	0	209	220	11			
28	T	T	11	12	1	157	161	4			
29	E	E	15	16	1	170	179	9			
30*	S	S	12	12	0	137	137	0			
31	L	M	15	16	1	144	145	1			
32*	A	A	11	11	0	134	129	5			

p_i	1EEI	1LTR	k_i^{1EEI}	k_i^{1LTR}	k_i	w_i^{1EEI}	w_i^{1LTR}	w_i	$\langle wij \rangle_{obs}^{1EEI}$	$\langle wij \rangle_{obs}^{1LTR}$	$\langle wij \rangle_{obs}^{2XSC}$
33*	G	G	8	9	1	90	90	0			
34*	K	K	8	8	0	91	92	1			
35	R	R	13	13	0	183	182	1			
36	E	E	17	17	0	186	191	5			
37	M	M	15	15	0	137	156	19			
38	A	V	11	13	2	110	144	34			
39	I	I	16	15	1	149	160	11			
40	I	I	14	16	2	155	154	1			
41	T	T	9	10	1	151	156	5			
42	F	F	14	14	0	212	213	1			
43*	K	K	6	9	3	83	99	16			
44*	N	S	5	4	1	95	79	16			
45*	G	G	5	5	0	64	64	0			
46*	A	A	8	7	1	112	106	6			
47*	T	T	10	10	0	127	129	2			
48	F	F	15	15	0	205	208	3			
49	Q	Q	16	16	0	200	204	4			
50*	V	V	12	14	2	123	128	5			
51*	E	E	12	12	0	129	134	5			
52*	V	V	11	10	1	113	123	10			
53*	P	P	11	9	2	89	82	7			
54*	G	G	5	5	0	90	58	32			
55*	S	S	4	5	1	60	56	4			
56*	Q	Q	8	8	0	127	101	26			
57	H	H	10	11	1	185	174	11			
58*	I	I	9	7	2	126	88	38			
59*	D	D	6	6	0	81	81	0			
60*	S	S	8	8	0	108	86	22			
61	Q	Q	15	14	1	200	186	14			
62*	K	K	9	8	1	108	104	4			
63*	K	K	9	11	2	91	110	19			
64*	A	A	12	12	0	114	126	12			
65	I	I	14	13	1	174	175	1			
66	E	E	12	11	1	172	161	11			
67	R	R	17	18	1	214	224	10			
68	M	M	17	17	0	182	177	5			
69	K	K	16	15	1	190	186	4			
70	D	D	10	11	1	160	163	3			
71	T	T	14	14	0	154	167	13			
72	L	L	15	18	3	145	151	6			
73	R	R	15	15	0	187	199	12			
74*	I	I	11	12	1	125	137	12			
75	A	T	11	15	4	123	162	39			
76	Y	Y	16	18	2	202	246	44			
77*	L	L	10	13	3	122	132	10			
78*	T	T	7	8	1	119	116	3			
79*	E	E	8	10	2	107	138	31			

p_i	1EEI	1LTR	k_i^{1EEI}	k_i^{1LTR}	k_i	w_i^{1EEI}	w_i^{1LTR}	w_i	$\langle wij \rangle_{obs}^{1EEI}$	$\langle wij \rangle_{obs}^{1LTR}$	$\langle wij \rangle_{obs}^{2XSC}$
80*	A	T	10	14	4	88	128	40			
81	K	K	12	10	2	172	119	53			
82	V	I	14	17	3	147	152	5			
83	E	D	12	11	1	174	152	22			
84	K	K	13	13	0	166	153	13			
85	L	L	16	16	0	145	148	3			
86	C	C	15	16	1	153	149	4			
87	V	V	14	14	0	182	175	7			
88	W	W	19	19	0	187	217	30			
89	N	N	8	8	0	144	138	6			
90*	N	N	5	6	1	104	109	5			
91*	K	K	10	10	0	126	140	14			
92*	T	T	7	7	0	93	94	1			
93	P	P	11	11	0	165	170	5			
94	H	N	13	14	1	169	169	0			
95	A	S	10	11	1	126	155	29			
96	I	I	16	16	0	138	139	1			
97*	A	A	12	13	1	131	133	2			
98*	A	A	13	13	0	129	130	1			
99	I	I	16	17	1	142	143	1			
100	S	S	12	11	1	149	140	9			
101	M	M	17	16	1	140	161	21			
102*	A	E	8	9	1	106	120	14			
103	N	N	5	10	5	74	153	79			

Table 3.3: Pairwise atomic interactions of the fourteen mutated positions with mutated neighbors.

mutation i	Neighbor j	w_{ij} (1EEI)	w_{ij} (1LTR)	Δw_{ij}	mutation i	Neighbor j	1EEI	1LTR	Δw_{ij}
1	35	5	11	-6	80	22	2	1	1
	37	2	4	-2		23	8	12	-4
	2	37	30	7		24	1	1	0
	3	5	6	-1		75	15	24	-9
	7	0	2	-2		76	1	1	0
	49	7	4	3		78	14	20	-6
	92	10	11	-1		79	21	26	-5
	93	8	10	-2		81	22	27	-5
4	2	10	10	0		82	6	5	1
	3	34	25	9		101	1	4	-3
	5	27	26	1		77	1	5	-4
	6	19	17	2		103	0	2	-2
	7	33	25	8		74	0	1	-1
	8	7	9	-2		76	0	2	-2
	39	4	4	0	82	20	11	15	-4
7	1	0	2	-2		21	13	8	5
	2	5	20	-15		22	19	18	1
	3	2	14	-12		23	2	2	0
	4	33	25	8		24	3	4	-1
	5	9	8	1		42	6	8	-2
	6	31	24	7		80	6	5	1
	8	21	22	1		81	32	28	4
	9	8	8	0		83	26	26	0
	10	11	14	-3		84	8	6	2
	11	11	10	1		85	3	3	0
10	5	1	0	1		99	6	8	-2
	6	5	8	-3		100	6	5	1
	7	11	14	-3		101	5	10	-5
	8	8	9	-1		40	0	1	-1
	9	21	23	-2		72	0	1	-1
	11	24	24	0		75	04	-4	
	12	11	10	1	83	5	1	0	1
	15	2	1	1		18	3	3	0
18	16	20	24	-4		19	9	9	0

mutation i	Neighbor j	w_{ij} (1EEI)	w_{ij} (1LTR)	Δw_{ij}	mutation i	Neighbor j	1EEI	1LTR	Δw_{ij}
	17	36	34	2		20	11	12	-1
	19	25	25	0		21	27	14	13
	20	15	13	2		81	6	2	4
	85	23	22	1		82	26	26	0
	84	10	10	0		84	47	43	4
	83	3	3	0		85	2	2	0
	87	5	6	-1		100	24	18	6
	89	1	2	-1		101	8	10	-2
	94	6	5	1		102	8	11	-3
	48	13	9	4	94	16	4	0	4
	86	0	1	-1		18	6	5	1
20	18	15	13	2		47	3	6	-3
	19	31	31	0		48	15	22	-7
	21	28	26	2		49	14	14	0
	22	12	17	-5		87	14	12	2
	82	11	15	-4		88	15	11	4
	83	11	12	-1		89	27	15	12
	84	4	5	-1		90	1	1	0
	85	7	8	-1		91	15	14	1
	42	20	16	4		92	5	4	1
	44	1	0	1		93	27	28	-1
	46	1	0	1		95	22	24	2
	48	6	3	3		96	4	4	0
25	23	4	5	-1	95	49	18	23	-5
	24	24	25	-1		51	8	14	-6
	26	35	29	6		86	3	2	1
	27	2	2	0		87	13	13	0
	41	21	15	6		88	30	33	-3
	42	19	17	2		91	6	7	-1
	43	15	13	2		93	5	5	0
	76	6	6	0		94	22	24	-2
	101	3	3	0		96	25	29	-4
	102	9	12	-3		97	4	4	0
	103	3	13	-10		50	0	1	-1
75	24	5	4	1	102	25	9	20	-11
	70	1	1	0		26	6	13	-7
	71	7	14	-7		76	11	11	0
	72	12	18	-6		81	16	2	14

mutation i	Neighbor j	w_{ij} (1EEI)	w_{ij} (1LTR)	Δw_{ij}	mutation i	Neighbor j	1EEI	1LTR	Δw_{ij}
	73	8	9	-1		83	8	11	-3
	74	24	29	-5		100	2	2	0
	76	21	24	-3		101	20	21	-1
	77	8	8	0		103	29	27	2
	78	17	15	2					
	79	8	6	2					
	80	15	24	-9					
	10	1	4	-3					
	82	0	4	-4					
	77	0	1	-1					

3.5.2 Supplementary Figures

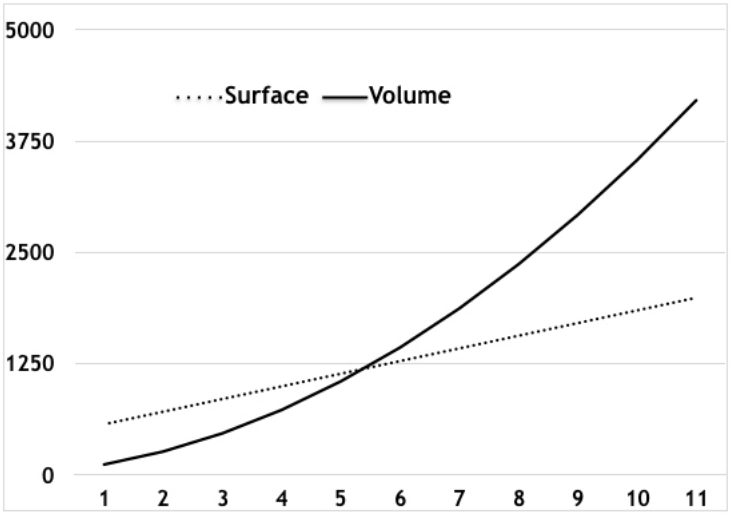


Figure 3.7: Torus Simulation

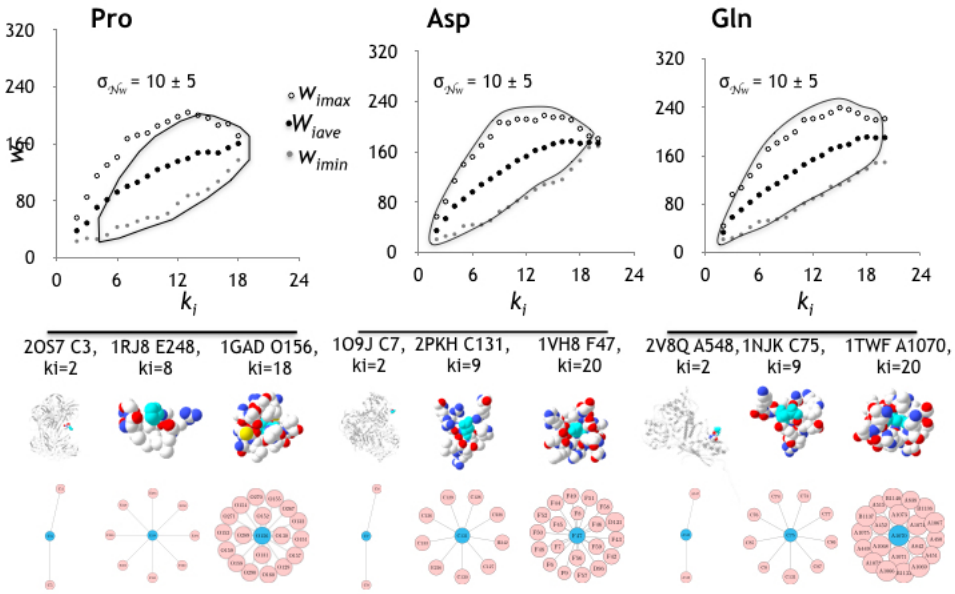


Figure 3.8: Amino acid capacity of interaction: Pro, Asp, and Gln.

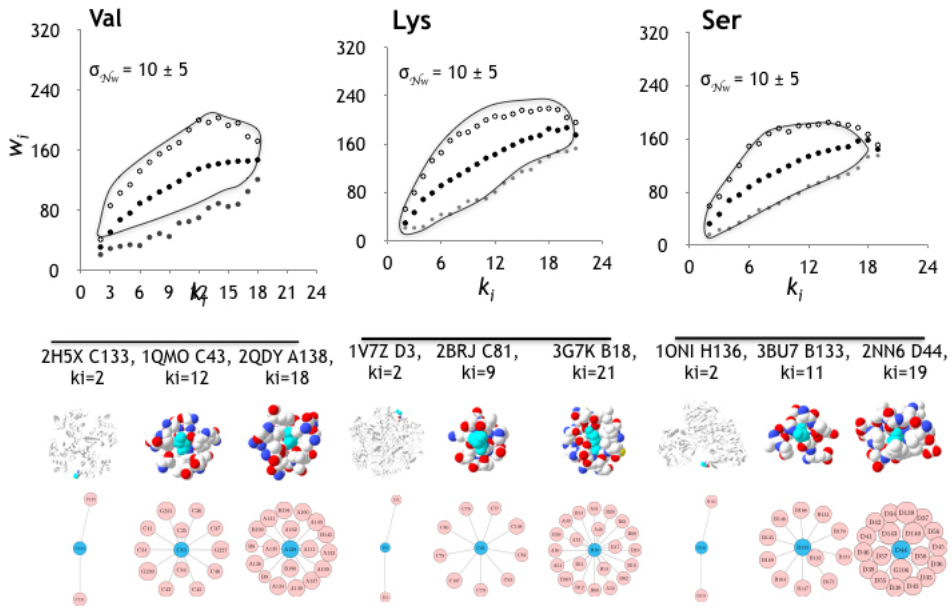


Figure 3.9: Amino acid capacity of interaction: Val, Lys, and Ser.

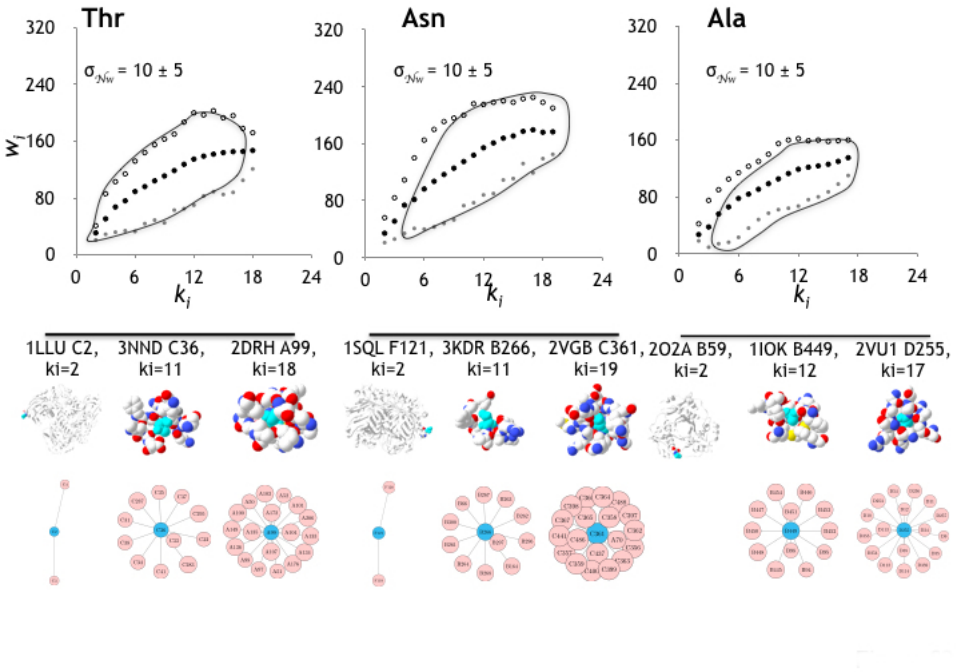


Figure 3.10: Amino acid capacity of interaction: Thr, Asn, and Ala.

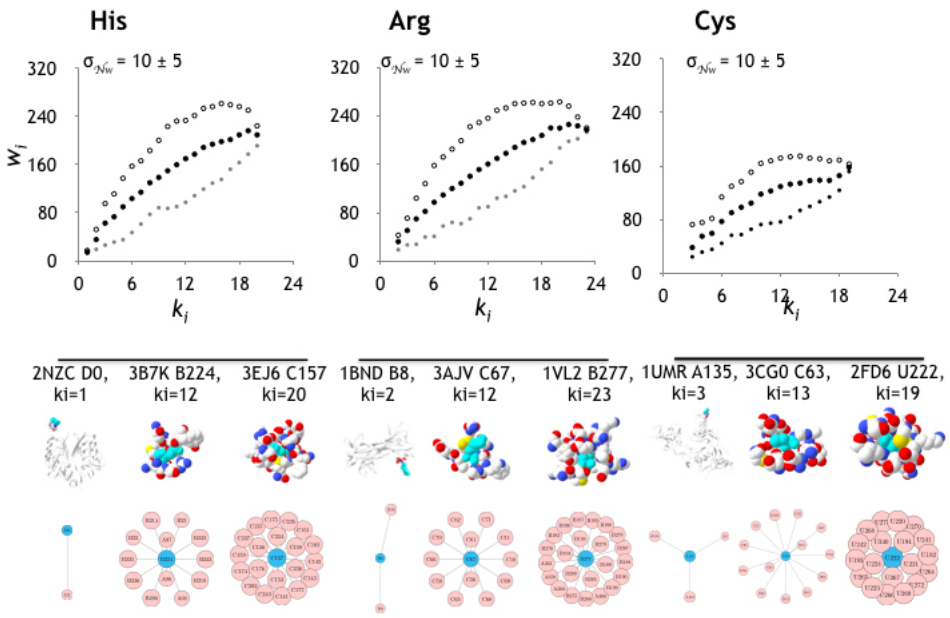


Figure 3.11: Amino acid capacity of interaction: His, Arg, and Cys.

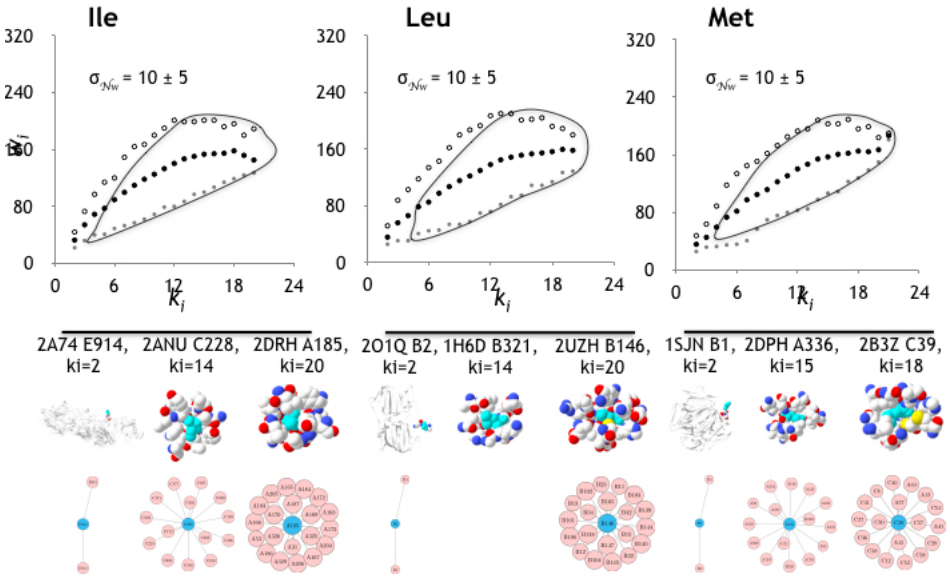


Figure 3.12: Amino acid capacity of interaction: Ile, Leu, and Met.

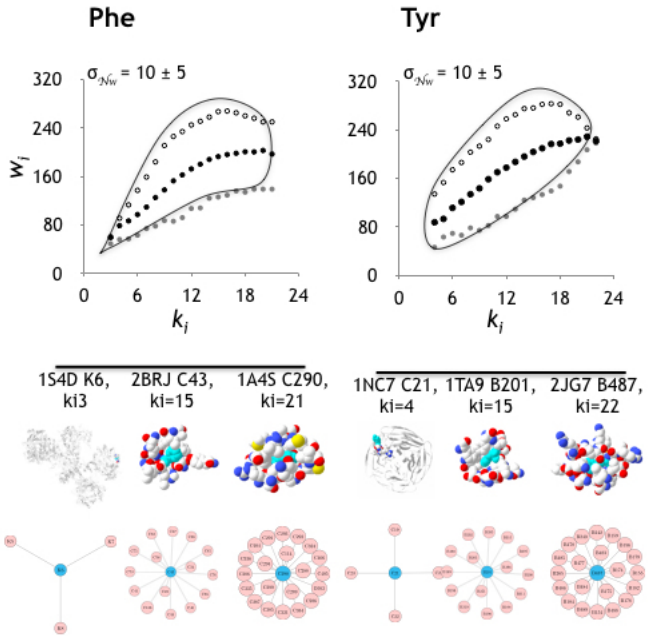


Figure 3.13: Amino acid capacity of interaction: Phe and Tyr.

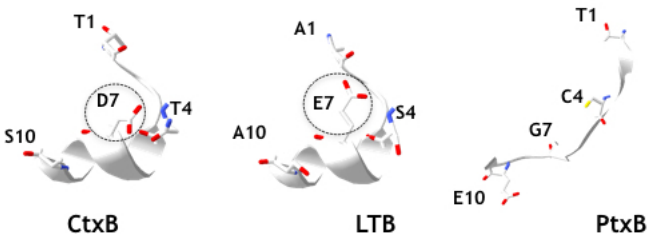


Figure 3.14: N-terminal of verotoxin-1 (PDB code 2XCS)

Chapter 4

Perturbation of amino acid networks: A statistical study of the defects introduced in protein interactions by mutations

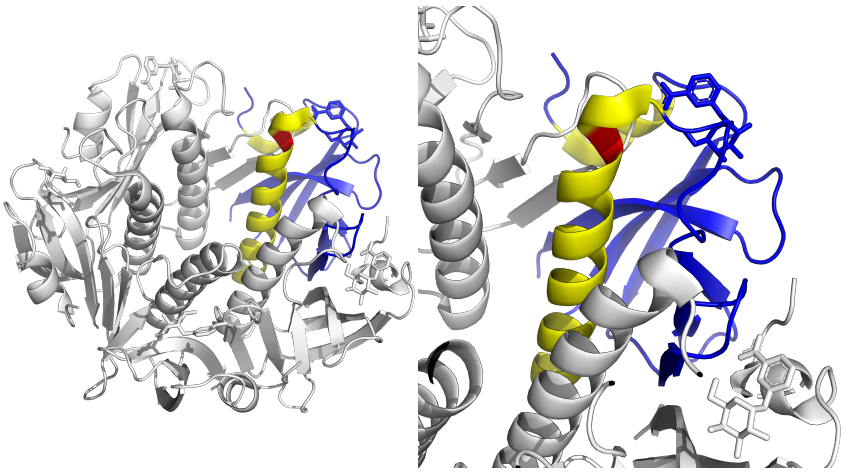
4.1 Introduction

According to the thermodynamical hypothesis, the amino acid sequence, also called the primary structure of a protein, is the one defining the native structural configuration of proteins [4]. The event of an amino acid replacing another in the sequence—a mutation—produces a structural change in the protein which can have a subsequent effect on the protein function. In reality, most mutations don't produce any change on the function [68]: Two amino acid sequences differing in one or more amino acids yield proteins having the same function, this is known as functional *robustness*. A small set of mutations, however, has a deleterious effect on the native function of the protein, which causes *fragility*. In some cases, these mutations can yield a second protein function and which allows for *adaptability* [75]. The final consequence of a mutation on the protein function depends, in part, on the effect of the mutation on the structure of the protein. Large throughput analyses of the effects of mutations on protein structure are rarely investigated [28].

A single mutation is a variation of the amino acid sequence at one position. McLaughlin et al. studied the effects relative to functional change of a set of 1577 single mutations in the third PDZ domain of the PSD-95 protein [42]; where 83 positions were mutated by the other 19 amino acids ($83 \times 19 = 1577$) and for each mutation, the functional change was measured experimentally. To be able to compare the functional change across positions, they annotated for each position the average value of the change caused by a mutation at the position. A subset of 20 positions out of 83 was shown to have average values of functional change that deviated more than two standard deviations from the total mean [42]. These 20 positions are here called *functionally sensitive*, where single mutations happening at functionally sensitive positions (FSP) cause fragility and/or adaptation remarkably more often than in the rest of the positions [42]. Moreover, sequence subsets of FSP appear to be modulating functional change within a protein family, as studies on protein sequence alignment suggest [24, 42].

Our question is: Do FSP feature some structural characteristic that makes them more functional susceptible to mutation? Indeed, their functional sensitivity could imply that FSP have a privileged place in the protein structure, which do not depend on the geometrical structure of the positions, but of the spatial distribution of the atoms belonging to their neighborhood.

Studies on the evolution of proteins show that protein structure is robust to mutations: mutations can affect the structure and function of the protein without destroying the protein phenotype (structure) to promote evolvability [76], where the changes in the structure can modulate the protein function



(a) (b)

Figure 4.1: (a) The structure of the Cholera Toxin B pentamer is shown together with a residue at sequence position 61 in chain E in red on its quaternary structure (white, the whole pentamer), tertiary structure (blue, chain E), and secondary structure (yellow, alpha helix). (b) Zoom on the residue E61.

towards a new evolutionary direction [52]. However, the systematic study of the effects of mutations on protein structure outside of the context of evolution, are challenging by the complexity of the structure. One can, for instance, study directly the rearrangement of atoms in space provoked by a mutation (e.g. by root-mean-square deviation), but doing so fails to consider the distinct inherent structures of a protein: The protein structure consists of four different types of substructures, namely, secondary, tertiary, and quaternary structures (Section 1.1). These underlying substructures are close from each other in the atomic arrangement of the protein, but account for different structural hierarchies in the protein. A measure of structural change should therefore take into account the intrinsic relation between substructures connected within the 3D structure of the protein. The protein primary, secondary, tertiary and quaternary structures refer to an onion like structural system, where each amino acid is part of a secondary, tertiary and quaternary structure, but its neighbors could belong to distinct structures (Figure 4.1).

In this Chapter, we look at the interactions between amino acids and atoms to model protein structure. We produce the same set of mutations *in*

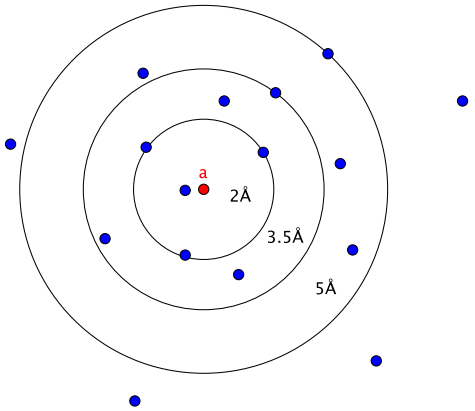
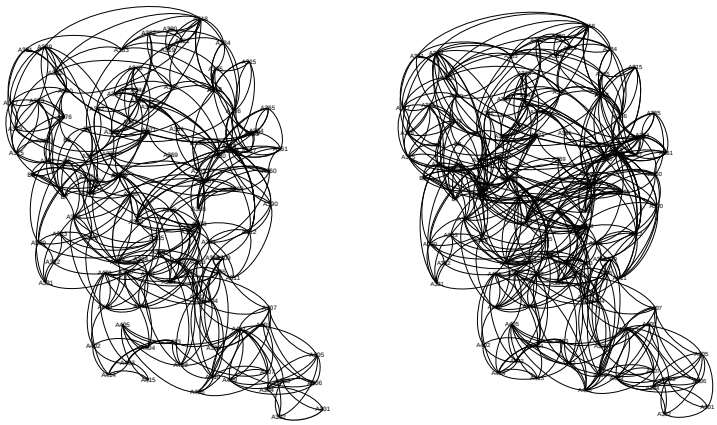


Figure 4.2: Different cutoffs—2 Å, 3.5 Å, and 5 Å—represented by concentric circles around an atom 'a' define the number of interactions of 'a'.

silico as McLaughlin [42], used to study experimentally the functional change of proteins due to the complete set of single mutations on 83 positions of the PSD-95 protein for a total of 1577 mutations. Our goal is to answer the question whether functional impact of mutations can be explained by structural change, more precisely by atomic and amino-acid interaction changes.

To model the global protein structure, we propose a network approach in which amino acids represent nodes, and links between them, interactions. We say that two amino acids are interacting if they share a link in the network and *vice versa*. For each link, we consider the number of underlying atomic interactions between two amino acids to define the *weight* of the link. In order to define the atomic interactions, we use a distance cutoff measured in Ångströms (Å), such that two atoms at distance smaller than the cutoff, are said to be interacting; only considering atoms in different amino acids to be interacting. The distance cutoff (or simply, cutoff), modulates the number of links in the network: larger cutoffs allow longer-distance and therefore create more connections between amino acids within the network (Figures 4.2 and 4.3).

A mutation, by modifying the number of atoms and their spatial configuration in space impacts on the amino acid interactions and consequently might impact the structure. The structure of a given position in the se-



(a) $c_1 = 4$Å (b) $c_2 = 5$Å

Figure 4.3: Two amino acid networks of the third PDZ domain of the PSD-95 protein constructed using a different cutoff distances c_1 and c_2. A larger cutoff creates more links between nodes. In (b) the network connects amino acids otherwise not connected in (a).

quence can then be studied as a function of the structural change caused by mutations at the position, monitored by the change in the set of interactions within the protein. The interactions are defined by a range of 71 distance cutoffs to study interaction changes upon mutation and include the chemical interactions (3–5 Å) and above (6–10 Å).

The *local structure* of an amino acid is the region of the protein including the amino acid and its amino acid neighborhood. There is an intrinsic relation between the interactions of amino acids and their neighbors, with their local structures. Globally, amino acid interactions control and maintain the protein structure, whereas locally, they define the neighborhood of amino acids.

On one hand, any perturbation of the protein structure implies a movement in at least two local structures: the one of the sequence position structurally modified and the local structure of at least one neighbor. On the other hand, a change in the amino acid interactions consequence of a mutation, causes a perturbation in the global structure[1]. This reciprocal implication between amino acid interactions and protein structure, supports a model of

[1]A change in the structure is understood here as a displacement of atoms in space.

protein structure in terms of amino acid interactions.

4.2 Methods

4.2.1 Amino Acid Network

The structural PDB files of 1577 mutations were obtained using the software FoldX version 3b6 [61]. The set of mutations, corresponds to all possible single-mutations over 83 positions of the third domain of the PSD-95 protein, PDB code 1BE9 [15] (83 positions, each mutated by all other 19 amino acids). For each mutation, an amino acid network (AAN) was constructed to model the amino acid interactions of the protein. Throughout this Section, we will define different measures accounting for the impact or *perturbation* of a mutation on said interactions. Any such measure, is called a *perturbation measure*, and assigns a real number to each mutation. That is, a perturbation measure f is defined on the set of mutations and mapped to the set of real numbers $\mathbb{R}$. Its values are obtained by comparison of the resulting amino acid network of a mutation, the mutation network, and the wild type amino acid network, the WT network (See also Section 3.4.2). The set of perturbation measures used in this Chapter were established and tested for their relevance in monitoring structural changes upon mutation in Chapter 2.

Given a distance distance cutoff c, the 83×83 adjacency matrix A of the AAN of a protein structure, has the following construction properties:

1. Each row and each column corresponds to one of the 83 positions in the sequence of amino acids.

2. For each pair of amino acids (v_1, v_2) in the protein, we count the number of atomic pairs (a_1, a_2), such that $a_1 \in v_1$ and $a_2 \in v_2$, and $dist(a_1, a_2) \leq c$. Where $dist$ is the Euclidean distance.

3. The entry $A_{i,j}$ is equal to the number of atomic pairs between amino acids corresponding to nodes i and j. If $A_{i,j} = 0$, then amino acids corresponding to nodes i and j are said not to be interacting.

The amino acid network $G = (V, E, c, w)$ is the network with adjacency matrix A. Nodes are labeled by their positions in the amino sequence, and two nodes $i, j \in V$ share a link if $A_{i,j} > 0$. The function w defined on the set of links E, assigns to each link in $ij \in E$, a real number, called the weight of ij and equal to $A_{i,j}$. The weight of a link connecting two nodes represents the number of shared atomic interactions between the two amino

acids (Figure 4.4). For convenience, we will use interchangeably the terms node, position, and amino acid; as well as link, and amino acid interaction, depending on the context.

4.2.2 Perturbation network $\mathcal{P}$

The *perturbation network*, noted $\mathcal{P}$, captures the change in atomic and amino acid interactions caused by a mutation. Given a mutation m, the network $\mathcal{P} = (V_{\mathcal{P}}, E_{\mathcal{P}}, m)$ is obtained by comparing the set of connections within the network WT (AAN of the PSD-95 protein) and the mutation network (AAN of mutation m). The sets $V_{\mathcal{P}}$ and $E_{\mathcal{P}}$, correspond to set of nodes and links of the perturbation network, respectively. Let M and A be the adjacency matrices of the networks mutation and WT, respectively. The adjacency matrix P of $\mathcal{P}$, is based on the absolute difference of matrices A and M, noted P':

$$P' = abs(M - A) \tag{4.1}$$

Where $abs(M - A)$ is the absolute value matrix of the difference of A and M. The adjacency matrix P is obtained by removing any all-zeroes row and column from P'.

The entry $P_{i,j}$ is equal to the absolute value of the number of atomic interactions between amino acids i and j in the mutation AAN, minus the number of atomic interactions in the WT AAN. In this Chapter, the perturbation network is the cornerstone of the quantification of the impact of a mutation on the original set of atomic and amino acid interactions.

4.2.3 Sphere of influence

The sphere of influence (Section 3.4.2), represents the extent of the changes in the amino acid interactions produced by a mutation, changes that spread from the mutated position to the rest of the structure. These changes happen in a cascade mechanism, where the mutation first affects interactions on the neighboring amino acids of the mutated position, which in turn affect interactions with their neighbors, and so on. This cascade mechanism can be thought of as a map of the interaction changes around the mutated position.

Perturbation measure $\mathcal{W}_{\mathcal{P}}$

The first perturbation measure we consider, is the sum of weights of links of the perturbation network $\mathcal{P}$, noted $\mathcal{W}_{\mathcal{P}}$. Given a mutation m, the number

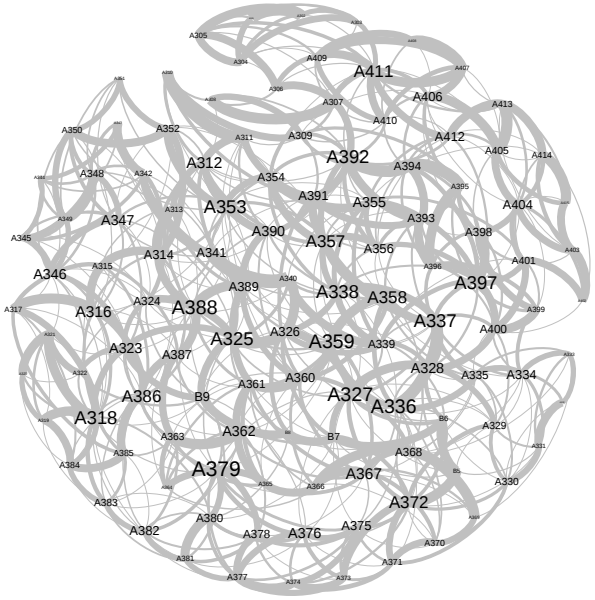


Figure 4.4: The amino acid network of the third PDZ domain of the protein PSD-95 (PDB code 1BE9). The set of nodes is equivalent to the set of positions in the amino acid sequence, and the set of links represents the interactions between different amino acids. The interactions are defined by the distance between amino acids in the protein 3-D structure. In this network, two amino acids are considered to be interacting if they share an atomic pair at distance less than 5 Ångströms (Å). Nodes are labeled based on the format "Chain + position". The size of the labels in the display are proportional to their number of contacts. Two nodes are connected by a link if they are "interacting", where the width of the link is proportional to its weight, representing the number of interacting atomic pairs shared by the two nodes.

$\mathcal{W}_\mathcal{P}$ is defined as follows:

$$\mathcal{W}_\mathcal{P}(m) = \sum_{e \in E_\mathcal{P}} w(e) \tag{4.2}$$

Where $\mathcal{P} = (V_\mathcal{P}, E_\mathcal{P}, m)$ is the perturbation network of mutation m. The measure $\mathcal{W}_\mathcal{P}$ was introduced as the amino acid rank in Chapter 2, and was not part of the original concept of sphere of influence of a mutation, which was conceived as a quality assessment of the interaction changes (Section 2.4).

The weight of a link in $\mathcal{P}$, represents the absolute difference in the number of atomic interactions between the two incident nodes in the WT and mutation networks. Therefore, $\mathcal{W}_\mathcal{P}(m)$ is the total change in atomic interactions captured by the perturbation network, i.e., $\mathcal{W}_\mathcal{P}$ counts the number of atomic interactions perturbed by mutation m.

The importance of the perturbation measure $\mathcal{W}_\mathcal{P}$ lies in the fact that it accounts for a perturbation in the original set of interactions, at an atomic level. The rest of the measures obtained using the perturbation network consider changes produced at an amino acid level.

Perturbation measure $\mathcal{D}_\mathcal{P}$

Measure $\mathcal{D}_\mathcal{P}$is similar to $\mathcal{W}_\mathcal{P}$as it counts the number of links of the perturbation network (but does not consider the weight of each link). If $\mathcal{P} = (V_\mathcal{P}, E_\mathcal{P})$, then $\mathcal{D}_\mathcal{P} = |E|$, where '$|E|$' denotes the cardinality of the set E.

Perturbation measure $ord_\mathcal{P}$

Another perturbation measure taken from the perturbation network $\mathcal{P}$, is the number of nodes in $\mathcal{P}$ (the order of $\mathcal{P}$), noted $ord_\mathcal{P}$. The measure $ord_\mathcal{P}$ represents the number of nodes incident to a link whose weight differs in the amino acid networks corresponding to the mutation and WT. This is measure was included in Chapter 2, where it was used to define the sphere of influence of mutations happening at the protein interface. Given a mutation m and its perturbation network $\mathcal{P} = (V_\mathcal{P}, E_\mathcal{P}, m)$, $ord_\mathcal{P}$ is the number of amino acids perturbed by m:

$$ord_\mathcal{P}(m) = |V_\mathcal{P}| \tag{4.3}$$

Where $|V_\mathcal{P}|$ is the cardinality of $V_\mathcal{P}$.

Perturbation measure $\mathcal{G}_{\mathcal{P}}$

A third perturbation measure included in the sphere of influence of a mutation, is the length $\mathcal{G}_{\mathcal{P}}$ of the *maximal shortest path* in the perturbation network $\mathcal{P}$ starting at the mutated node. Let $\mathcal{P} = (V_{\mathcal{P}}, E_{\mathcal{P}}, m_p)$ be the perturbation network of mutation m_p at position of the amino acid sequence $p \in V_{\mathcal{P}}$. A path in $\mathcal{P}$ starting at node p, is a non-empty network $Q = (V, E)$ of the form

$$V = \{x_0, x_1, \ldots, x_k\} \qquad E = \{x_0x_1, x_1x_2, \ldots, x_{k-1}x_k\},$$

where the x_i are all distinct and $x_0 = p$. We say that Q is a *shortest path* if for every other path C starting at p and ending at x_k, we have $|Q| \leq |C|$, where the cardinality of a path is equal to its number of links.

A shortest path Q in $\mathcal{P}$ starting at p, is said to be *maximal*, if for any shortest path Q' in $\mathcal{P}$ starting at p, we have $|Q'| \leq |Q|$. The length of a maximal shortest path starting at the mutated position, accounts for a measure of "depth" of the perturbation by the mutation. It is an approximation of the extent of the sphere of influence of the mutation.

The maximal shortest path itself is a traceability-measure of the changes from the site of the mutation to elsewhere in the structure, in which the perturbations occur in the cascade mechanism of the sphere of influence. The length of the maximal shortest path starting at the mutated position p, $\mathcal{G}_{\mathcal{P}}(m_p)$, is called the *geodesic distance*[2] of a mutation in Chapter 2 (Section 2.3).

Perturbation measure $\mathcal{E}_{\mathcal{P}}$

A third measure—in addition to $\mathcal{W}_{\mathcal{P}}(m_p)$ and $ord_{\mathcal{P}}(m_p)$—included in the sphere of influence of a mutation m_p, is the *maximal Euclidean distance*, noted $\mathcal{E}_{\mathcal{P}}(m_p)$, measured from the mutated position p, to an amino acid in the perturbation network. Let $dist(p, v)$ be the Euclidean distance from p to amino acid $v \in V_{\mathcal{P}}$, we say that $dist(p, v)$ is maximal if for any other distance $dist(p, u)$ from p to amino acid $u \in V_{\mathcal{P}}$ we have

$$dist(p, u) \leq dist(p, v),$$

in which case $\mathcal{E}_{\mathcal{P}}(m_p) = dist(p, v)$.

[2]The term *geodesic distance* on its original sense was defined as the shortest route between two points on the Earth's surface.

4.2.4 The matrix $\mathcal{M}$

We have introduced so far a several measures of the structural perturbation of a mutation using an amino acid network approach (perturbation measures). These measures are used to compare the average impact of all mutations on a same position, in the amino acid sequence. Furthermore, given the values of a perturbation measure for all possible mutations at one position, we can calculate the average impact of a mutation at that position, and subsequently, compare two different positions in the sequence.

Given a distance cutoff c, and a perturbation measure f^c, one tool to visualize the values of a perturbation measure f of all possible mutations at the 83 positions of the protein sequence, is the 20×83 matrix, $\mathcal{M}(f^c)$. Each column corresponds to an amino acid sequence position, and each row to one of the 20 amino acids (Figure 4.6). For convenience, we order columns in increasing order of sequence position, and rows in alphabetical order from top to bottom. In this manner, column j corresponds to the j-th position in the amino acid sequence, and row i to the i-th amino acid by alphabetical order. Given a perturbation measure f defined on the complete set of 1577 mutations, the matrix $\mathcal{M}(f^c)$ contains the values of $f(m)$ for any single mutation m (Figure 4.6).

In order to compare the values of impact of mutations across sequence positions using a cutoff c and a perturbation measure f^c, we define the value of $f^c(p)$ of sequence position p, as the average value of f^c of mutations at p:

$$f^c(p) = \frac{1}{20} \sum_{i=1}^{20} \mathcal{M}(f^c(m = ip)) \tag{4.4}$$

Where i is any of the twenty amino acid types and $m = ip$ is the mutation at position p by amino acid type i.

4.2.5 Cutoffs

Perturbation measures described so far measure the change in the *rewiring* of the set of atomic and amino acid interactions produced by mutations. They are obtained by comparing the protein wild type and mutation amino acid networks. Therefore, they vary depending on the cutoff used to define the distance at which interactions are defined to take place. As previously noted, the larger the cutoff, the larger the possible distance and strength between two amino acids to interact: For a given amino acid, a greater cutoff means a larger "interaction scope".

However, its number of interactions depend on the structure surrounding that amino acid, too (Figure 4.2). It is ultimately the local structure of the amino acid that explains how differently mutations affect interactions. The cutoff used, defining the local structure taken into account, is consequently critical for the relevance of the output of a perturbation measure: The amino acid network considers interactions to happen only at a distance bounded by the cutoff. The use of different cutoffs (therefore different local structures), results in different perturbation values. For the analysis of the values of a perturbation measure, we underline the importance to use a large set of cutoffs, and to analyze values across cutoffs. For the purpose of this work, we use 71 cutoffs within the interval 3–10 Å, with a separation of 0.1 Å from each other. This includes chemical distances in the range 3–5 Å and above for cutoffs 5–10 Å, all smaller than the diameter of the protein.

4.2.6 The boolean matrix $\mathcal{R}$

The central question of this Chapter deals with the comparison of structural measures called perturbation measures, across different types of positions. Specifically, we want to know if the functionally fragility of positions obtained experimentally [42], is related to their structural fragility. In order to answer this question, we use different distance cutoffs to define the local structures of amino acids in the 3D structure (Subsection 4.2.5).

Here, we propose a method to compare values of a given perturbation measure across different cutoffs, while mainly focusing on the functionally sensitive positions. Given a perturbation measure f, we compare the distinct values of sequence positions across cutoffs. To do that, we note $f(p, c)$ as the value in f of sequence position p, using cutoff c; and consider the ordered array r^c

$$r^c = (r_1^c, r_2^c, \ldots, r_{83}^c),$$

such that:

$$f(r_1^c, c) \geq f(r_2^c, c) \geq \cdots \geq f(r_{83}^c, c).$$

In other words, r^c is a ranking of the 83 sequence positions given the perturbation measure f and a cutoff c. The position r_1^c has the larger f-value when using cutoff c and is therefore ranked 1st in r^c.

The method consists on computing the ranking r^c for each cutoff in the

$$\mathcal{R}(f) = \begin{array}{c} \\ c_1 \\ c_2 \\ \vdots \\ c_{70} \\ c_{71} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{1st rank} & \text{2nd rank} & \cdots & \text{82th rank} & \text{83th rank} \end{array} \\ \begin{pmatrix} \mathcal{R}_{1,1} & \mathcal{R}_{1,2} & \cdots & \mathcal{R}_{1,82} & \mathcal{R}_{1,83} \\ \mathcal{R}_{2,1} & \mathcal{R}_{2,2} & \cdots & \mathcal{R}_{2,82} & \mathcal{R}_{2,83} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathcal{R}_{70,1} & \mathcal{R}_{70,2} & \cdots & \mathcal{R}_{70,82} & \mathcal{R}_{70,83} \\ \mathcal{R}_{71,1} & \mathcal{R}_{71,2} & \cdots & \mathcal{R}_{71,82} & \mathcal{R}_{71,83} \end{pmatrix} \end{array}$$

Figure 4.5: The 71×83 matrix $\mathcal{R}(f)$ of the perturbation measure f. Here, $c_1, c_2, \ldots, c_{71}$ are the cutoffs equal to 3 Å, 3.1 Å, ..., 10 Å, respectively. The value of $\mathcal{R}_{i,j}$ depends on the position p ranked jth by decreasing order of f. If the position p is a functionally sensitive position then $\mathcal{R}_{i,j} = 1$, otherwise $\mathcal{R}_{i,j} = 0$.

interval 3–10 Ångströms with a step of 0.1 Ångströms (71 cutoffs total).

$$\begin{aligned} r^3 &= (r_1^3, r_2^3, \ldots, r_{83}^3) \\ r^{3.1} &= (r_1^{3.1}, r_2^{3.1}, \ldots, r_{83}^{3.1}) \\ &\vdots \\ r^{10} &= (r_1^{10}, r_2^{10}, \ldots, r_{83}^{10}) \end{aligned}$$

Given a perturbation measure f, to each position there corresponds 71 different rankings, one for each cutoff (Figure 4.5). The use of rankings facilitates locating and highlighting ranks belonging to a particular subset of positions across multiple cutoffs. In order to highlight a subset of sequence positions S over all ranks and compare their distribution across cutoffs, we construct a 71×83 boolean matrix $\mathcal{R}(f)$ such that:

$$\mathcal{R}_{i,j} = \begin{cases} 1 & r_j^{c_i} \in \text{S} \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

Where $r_j^{c_i}$ is the sequence position $j \in \{1, \ldots, 83\}$ is the j-th rank using cutoff c_i, for $i \in \{1, \ldots, 71\}$.

4.2.7 Buriedness

A convex polyhedron $\mathcal{P}$ is a solid figure with flat faces that has the property that any line between two points of $\mathcal{P}$ is contained within $\mathcal{P}$. The *convex*

hull $conv(S)$ of a set of points $S \in \mathbb{R}^3$ is the smallest convex polyhedron containing S. The surface of a protein is modeled here as the convex hull of the atomic coordinates.

The *buriedness* is a measure of the proximity of an amino acid to the boundary of the convex hull of the set S of atomic coordinates of the protein. First, the shortest Euclidean distance of each atom $a \in S$ to the boundary of $conv(S)$ is computed. The distance from a to the protein surface is modeled as the distance from a to the nearest face of $conv(S)$. The buriedness of a residue $r \subset S$, is defined as the average atomic distance from an atom in r to $conv(S)$.

This model has the advantage of assigning to each residue, a positive buriedness, as each residue has at least one atom in the interior of conv(S). The similar measure of accessible surface area (ASA), assigns a positive value only to residues in contact with the surface of the protein, and a zero value, otherwise [64]. Positions with no contact to the surface but close to it, are not mutually differentiable from buried positions. The modeling of the surface as a convex hull turns the protein surface into a collection of polygons, this makes the distance of the distance from a point in the protein to the surface easy to calculate.

The model cedes in sensitivity in terms of the shape of the surface, as concavities and holes are not considered. In our model, a residue in an actual concavity of the protein, will probably be detected as a position potentially buried, depending on the angle of the concavity. Nevertheless, our case study is based on a globular protein with no inner holes, limiting the possibility of falsely buried positions concave angles. The use of buriedness in this Chapter is meant for assessing whether buried and surface positions have differentiable local structures. The local structure of a residue in a pocket, for instance, would be expected to be somewhere in between the one of a buried and a surface position, even if the residue is technically in the surface of the protein. Reason why using buriedness to categorize residue positions in the 3D structure, is congruent with our objective of the study of different local structures.

4.3 Results and Discussion

In order to model the atomic and amino acid interactions occurring in a protein, we use amino acid networks (Subsection 4.2.1). An amino acid network is composed of a set of nodes, which represent the positions of amino acids in the 3D structure but are labeled by their position in the amino acid

sequence; and a set of links, which represent amino acid interactions. A weight is given to each link based on the number of atomic interactions between the two nodes. Atomic interactions are modeled using a distance cutoff: Two atoms are said to be in interaction if their distance is less than a given cutoff. Two amino acids, in turn, are connected by a link if they share an interacting atomic pair. The cutoff allows for several distinct *rewires* of the amino acid interactions of one structure. The use of a large range of cutoffs is useful to capture amino acid interactions when the characteristic-scale of protein structures/perturbations is not known.

We can quantify the difference of two structures in terms of their distinct sets of connections, or *wirings*. If two 3D structures yield the exact same *wiring*, they are considered to be topologically equal. Thus, given the structure corresponding to a mutated sequence, we can measure the *perturbation* of a mutation on the original set of amino acid interactions. The measure considers the different atomic and amino acid interactions between the mutation and the wild type proteins, as well as the furthest point perturbed in the 3D structure by the mutation (Subsection 4.2.3), and are obtained using the perturbation network of the mutation (Subsection 4.2.2).

In this Chapter, we analyze the consequences of mutations upon the structure surrounding amino acids in the third PDZ domain of the PSD-95 synaptic protein (PDB code 1BE9) [15]. Furthermore, we study the relation of the consequences of mutations in the amino acid interactions with the function and structure of the protein. The consequences on the amino acid interactions of a protein, are investigated using different *perturbation measures* (Subsection 4.2.3). Where a perturbation measure of a mutation quantifies the change of one structural property, using the amino acid networks of the mutation and the wild type as models of both structures.

The third PDZ domain of the PSD-95 synaptic protein was used to experimentally quantify the functional change, ΔE, of all single mutations across 83 sequence positions [42]. The functional change was shown in a 20×83 matrix $\mathcal{M}(\Delta E)$ (Figure 4.6) where each column j represents the j-th position by increasing order, and a row i represents the i-th amino acid by alphabetical order. The value of functional change of a position $\Delta E(p)$ was defined as the average functional change at position p: $\Delta E(p) = \frac{1}{20} \sum_{i=1}^{20} \mathcal{M}_{i,p}$. Functionally sensitive positions (FSP) are amino acid positions where average functional change by mutations are deviating more than two standard deviations from the total mean [42]. Do FSP have unique structural properties that are distinguishable from the rest of positions? To answer this question we will compare interaction changes in the neighborhood of FSP to the rest of positions. If FSP happen to have characteristic local structures, do the lo-

$$\mathcal{M}(f^c) = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \vdots \\ \text{W} \\ \text{Y} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{pos 1} & \text{pos 2} & \cdots & \text{pos 82} & \text{pos 83} \end{array} \\ \begin{pmatrix} f^c(1\text{A}) & f^c(2\text{A}) & \cdots & f^c(82\text{A}) & f^c(83\text{A}) \\ f^c(1\text{C}) & f^c(2\text{C}) & \cdots & f^c(82\text{C}) & f^c(83\text{C}) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f^c(1\text{W}) & f^c(2\text{W}) & \cdots & f^c(82\text{W}) & f^c(83\text{W}) \\ f^c(1\text{Y}) & f^c(2\text{Y}) & \cdots & f^c(82\text{Y}) & f^c(83\text{Y}) \end{pmatrix} \end{array}$$

Figure 4.6: The 20×83 matrix $\mathcal{M}$ of perturbation measure f using cutoff c. Entry $\mathcal{M}_{i,j} = f^c(m = ij)$ is the value of the perturbation measure $f^c(m)$ of mutation $m = ij$. Where m is replacing amino acid at position j by amino acid type i. Note that at least one value per column is equal to zero, namely when the amino acid type at the sequence position (column) coincides with the mutant amino acid type (row). This matrix was used by McLaughlin et al. to show the functional landscape of the effects of mutations for the third PDZ domain of the PSD-95 protein [42], where the function f was equal to functional change (Without the need of a distance cutoff).

cation in the 3D structure of positions (buried or surface exposed) influences the local structure of FSP? and the rest of the positions?

Based on the assumption the distinct structural properties of FSP, if existent, can be assessed from their PDB files, we reproduce the mutational set *in silico* from where the functionally sensitive positions were obtained (Subsection 4.2.1). Then, we model each mutated structure as an amino acid network to be compared to the wild type network. We then modulate the local structure of positions using different cutoff distances (Subsection 4.2.5). Moreover, we compare the structural perturbation values to the global structural value *buriedness* (Subsection 4.2.7), to study different local structures relative to their position in the 3D structure.

4.3.1 Sphere of Influence

Total weight of $\mathcal{P}$ and buriedness

We first consider the perturbation measure of the absolute change in atomic interactions $\mathcal{W}_\mathcal{P}$ consequence of a mutation (Equations 4.1 and 4.2). The perturbation measure $\mathcal{W}_\mathcal{P}$ counts, for each pair of amino acids, the post mutational difference in atomic interactions between the wild type and the mutation networks defined by a distance cutoff. It considers the difference

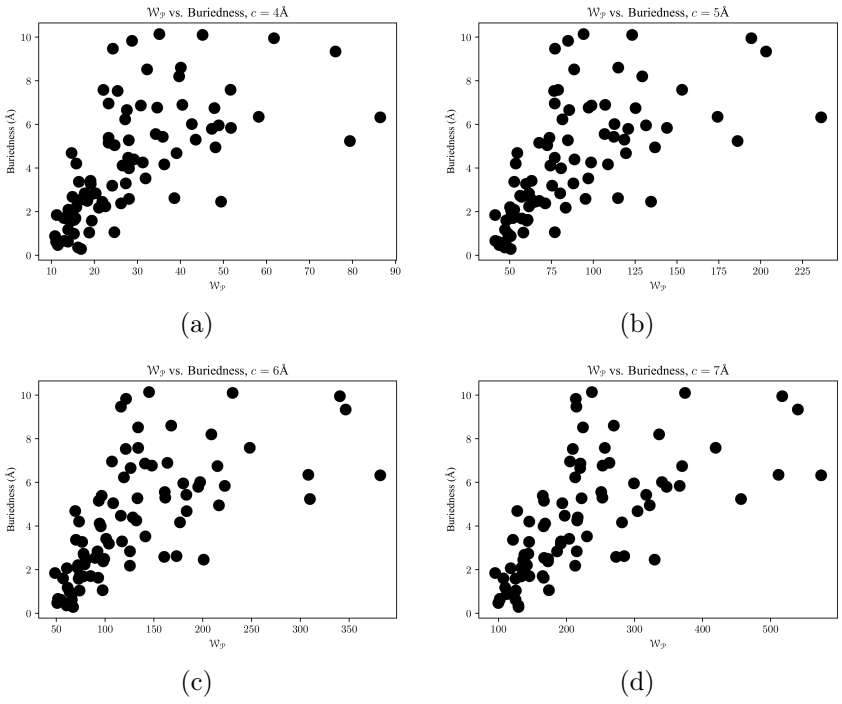


Figure 4.7: Perturbation measure $\mathcal{W}_\mathcal{P}$ vs. Buriedness, using four distinct cutoffs 4, 5, 6, and 7 Ångströms, in (a), (b), (c), and (d), respectively.

in the weight of each link in both networks (Subsection 4.2.2), as well as links that only exist in one network. We calculated the values of $\mathcal{W}_\mathcal{P}$ for 83 positions in the amino acid sequence. A comparison between $\mathcal{W}_\mathcal{P}$ and buriedness of positions—how buried the position is in the 3D structure (Subsection 4.2.7)—shows that for greater cutoffs, buried positions are accountable for a larger change in the set of interactions (Figure 4.7), i.e, mutations happening at buried positions perturb more atomic interactions for larger cutoffs. To calculate the correlation between buriedness and $\mathcal{W}_\mathcal{P}$, we used Pearson's correlation coefficient[3].

The buriedness of positions, accounts for 62% of the number of perturbed atomic interaction values for a cutoff of 4 Ångströms (Å). When using a cutoff of 10 Å, this percentage goes to 69% (Figure 4.8). On the other hand, the total difference in atomic interactions ($\mathcal{W}_\mathcal{P}$) is explained by the buriedness

[3]Pearson's correlation $-1 \leq r \leq 1$ is a measure of linear correlation between two variables X and Y, where $r = 1$ if there is total correlation and $r = -1$ if there is total negative correlation.

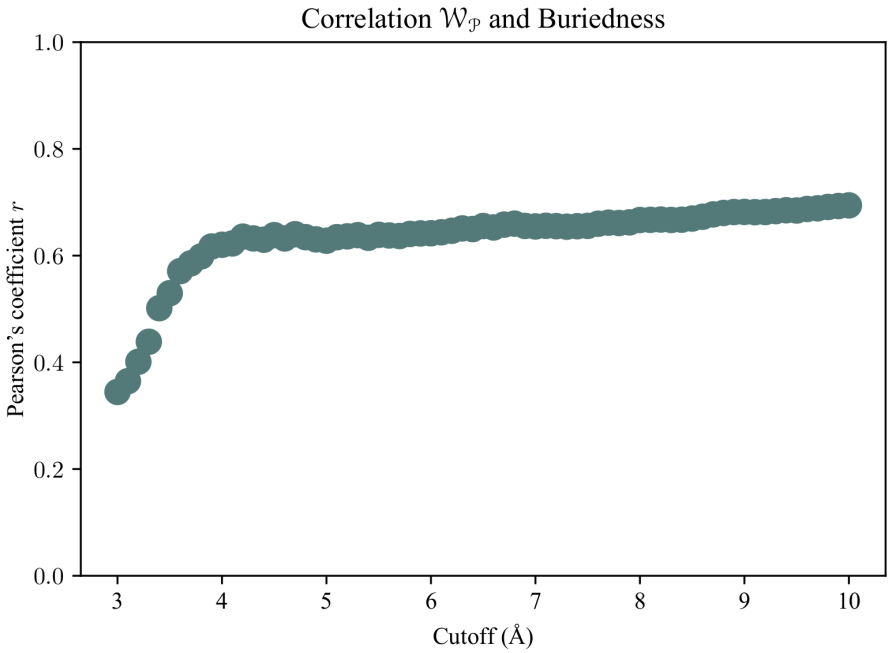


Figure 4.8: Pearson correlation r between the perturbation measure $\mathcal{W}_\mathcal{P}$ and buriedness. Each circle in the figure represents a correlation value between $\mathcal{W}_\mathcal{P}$ and buriedness for a given cutoff. The correlations are calculated for the 71 cutoffs in the interval 3–10 Å. Based on the buriedness of an amino acid, we can predict its approximate $\mathcal{W}_\mathcal{P}$ value for 62% to 69% of cases, depending on the cutoff used.

for only 34% of the cases when using a cutoff of 3 Å. This is explained by the fact that the larger the distance defining atomic interactions, the larger the connectivity for buried amino acids compared to surface-exposed amino acids, in terms of number of contacts.

The number of atomic interactions of a residue, being strongly correlated to its buriedness, supports this suggestion. The weighted degree of amino acids (their number of atomic interactions) in the amino acid network, increases more for buried positions as the cutoff increments than non-buried or surface positions. At 4 Å, the correlation between buriedness and weighted degree is equal to 0.47, significantly increasing the relation between weighted degree and buriedness of each position compared to smaller cutoffs. Any greater cutoff considered enhances the correlation; for a cutoff of 5 Å it is

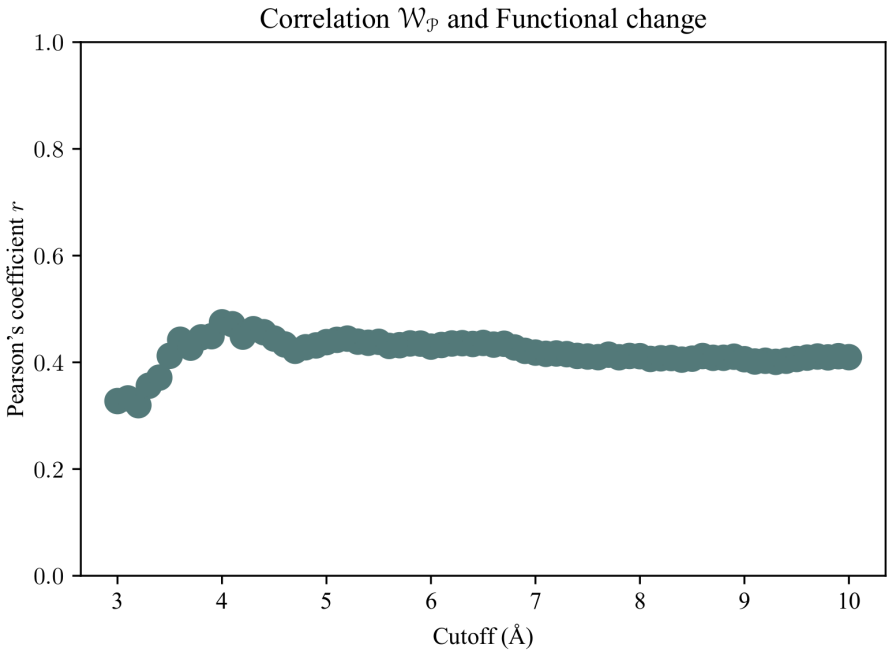


Figure 4.9: Pearson correlation coefficient r between perturbation measure $\mathcal{W}_{\mathcal{P}}$ and functional change. Each circle represents a correlation value between $\mathcal{W}_{\mathcal{P}}$ and functional change. The correlations are calculated for 71 cutoffs in the interval 3–10 Å.

0.6, and it peaks at 0.73 for a cutoff of 10 Å.

This suggests that local structures around 4 Å in the protein structure, are sufficient to discriminate between surface and buried positions (depending on their distance to the boundary of the convex hull), in terms of post mutational structural effects, i.e., the neighborhood of a buried amino acid is distinguishable from one of a surface amino acid at around 4 Å.

However, this correlation is non existent for cutoffs too small within the range of 3.0–3.9 Ångströms ($r = 0.117 \pm 0.19$). This fact could be explained by the *empty space* surrounding amino acids, making the local structures of small cutoffs unrecognizable from each other (Chapter 5).

Total weight of $\mathcal{P}$ and functional change

The set of atomic interactions in a protein modulate the protein activity. Atomic interactions occur between the active site and the ligand, and main-

tain stable the rest of the structure. We suggested in Chapter 2, that a "rewiring" of the original set of atomic interactions is not necessarily detrimental to the protein well functioning (Section 2.5). For example, a *rescue mutation*, that is, a mutation counterbalancing the negative effects of a second mutation [13], is necessarily changing the atomic composition of the original structure, and therefore, the atomic interactions. This suggests that there should be no strong correlation between functional change and $\mathcal{W}_{\mathcal{P}}$.

We compared the average number of perturbed atomic interactions of the 83 positions with their values of functional change, obtained experimentally [42]. The average number of perturbed atomic interactions, explains no more than 50% of the functional change values for any cutoff used (Figure 4.9). The change in atomic interactions being the most representative of the functional change when a cutoff of 4 Å is used (Pearson correlation coefficient $r = 0.47$). The average correlation on all cutoffs is 0.42 and the standard deviation 0.027.

The weak correlation between $\mathcal{W}_{\mathcal{P}}$ and functional change, implies that there are some mutations affecting the function of the protein, but not the atomic interactions. *Vice versa*, some mutations impact the atomic interactions but not the protein function. Gross change in atomic interactions is not sufficient to explain functional change, a more qualitative approach to atomic interactions is therefore needed.

Euclidean distance and buriedness

The perturbation network of a mutation is composed by the set of amino acids with at least one interaction impacted by the mutation. The set of links is the set of the affected amino acid interactions (Equation 4.3 in Subsection 4.2.3). The sphere of influence of a mutation, takes into account two parameters of the perturbation network of the mutation: The number of links, and the largest Euclidean distance from the mutated position to another amino acid in the perturbation network (Section 2.4.

One parameter determining the sphere of influence of a mutation is the maximal Euclidean distance $\mathcal{E}_{\mathcal{P}}$, in Ångströms, between the position mutated and any other amino acid in the perturbation network (Subsection 4.2.3). This measure combines information from the perturbation network and from the protein 3D structure. The correlation between the Euclidean distance from the mutation to the farthermost perturbed amino acid, and the functional change, varies very little across cutoffs (standard deviation = 0.05), and is on average -0.23. There is a weak correlation between the functional change produced by the mutation and the furthest distance in the 3D struc-

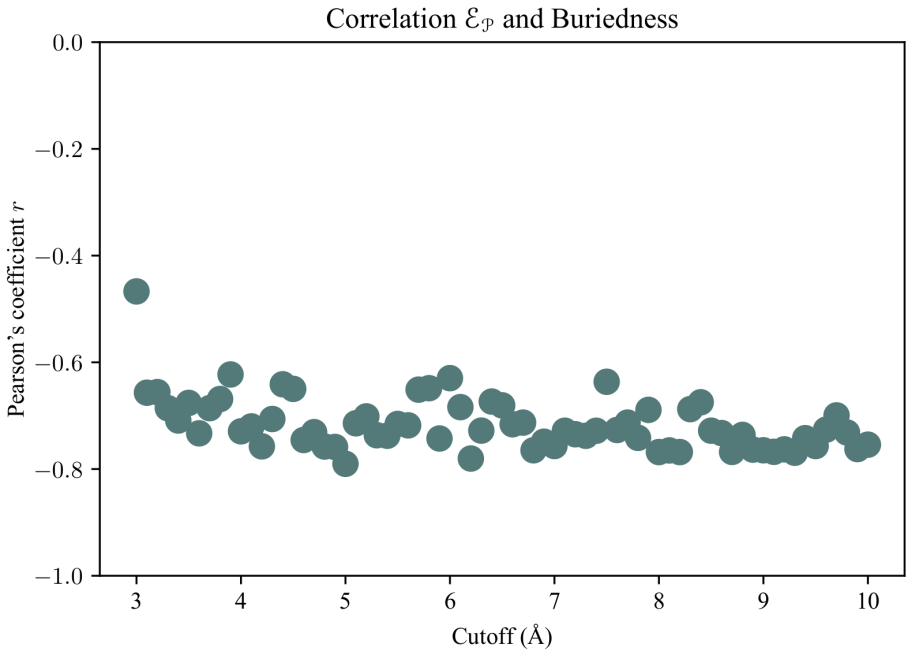


Figure 4.10: Correlation r between buriedness of a position and the maximal Euclidean distance $\mathcal{E}_\mathcal{P}$ perturbed by a mutation in the protein structure. Each circle represents the correlation between $\mathcal{E}_\mathcal{P}$ and buriedness for a given cutoff. Values are calculated for 71 cutoffs in the interval 3–10 Å.

ture affected by it.

On the other hand, the correlation between $\mathcal{E}_\mathcal{P}$ and buriedness is very strong (Figure 4.10). The buriedness of an amino acid given the set of atomic coordinates S of a protein, we remind, is defined as the distance from the amino acid to the nearest face of the *convex hull* of S, roughly representing the surface contact area of the protein (Subsection 4.2.7). The Pearson correlation coefficient between buriedness and the Euclidean distance of a mutation is less than or equal to -0.6 for any cutoff, with the exception of 3 Å. Two important observations of this result are worth being underlined. The first is that buried positions, when mutated, perturb less far in the structure than less buried positions. The second is that this is the true for all cutoffs in the range of 3 to 10 Ångströms.

The local structure of an amino acid, as previously mentioned, is the set of atomic and amino acid interactions of the amino acid. The corre-

lation between buriedness and $\mathcal{E}_{\mathcal{P}}$, implies that the local structure of even small cutoffs is representative of the place of the amino acid in the protein structure. In other words, the structure surrounding an amino acid around more than 3 Ångströms, carries information of its position in the 3D protein structure.

Furthermore, the fact that the correlation between maximal Euclidean distance and buriedness is true for all cutoffs, implies that the perturbation of a mutation in the structure "runs" further in the structure when it happens at surface positions. This reflects a real mechanism of the protein called allostery, in which the perturbation made by an effector at one site of the protein has a functional effect at the peptide binding site trough structural alteration [70].

For a same number of neighbors, the distribution of neighbors is more dense in surface positions than in buried positions, therefore, a same size perturbation is distributed across more amino acids when happening at a buried positions, in contrast to a less buried position where the propagation of the perturbation is more constricted structurally.

4.3.2 Number of perturbed amino acids and functional change

The number of nodes of the perturbation network, is the set of nodes incident to an edge whose weight differs in the mutation network, relative to the wild type network (Subsection 4.2.3). It is interpreted as the number of amino acids whose group of interactions has been "perturbed" by the mutation. The relation of this measure with the change in function, is much stronger than the $\mathcal{W}_{\mathcal{P}}$ (Figure 4.9). The number of perturbed nodes, explains up to 70% of the functional change of all 83 positions. All cutoffs considered, the number of perturbed nodes explains, on average, for 57% of the functional change (Figure 4.11).

To further test the relation between the number of nodes of the perturbation network (size of the network) and functional change, we computed the boolean matrix $\mathcal{R}$ (Equation 4.5 in Subsection 4.2.6). The 71×83 boolean matrix $\mathcal{R}$ has for each row a different cutoff, and the entry at column j and row i is equal to 1 if the j-th rank of the range of measure op using cutoff i, in decreasing order belongs to a functionally sensitive position (FSP), and equal to 0, otherwise: If the j-th rank of $ord_{\mathcal{P}}$ using cutoff i does not belong to a FSP, then the value for that entry in $\mathcal{R}$ is 0. This is done to strongly contrast ranks of the FSP with those corresponding to the rest of the positions.

Considering the 71 cutoffs, each having 83 ranks, there is a total of 5893

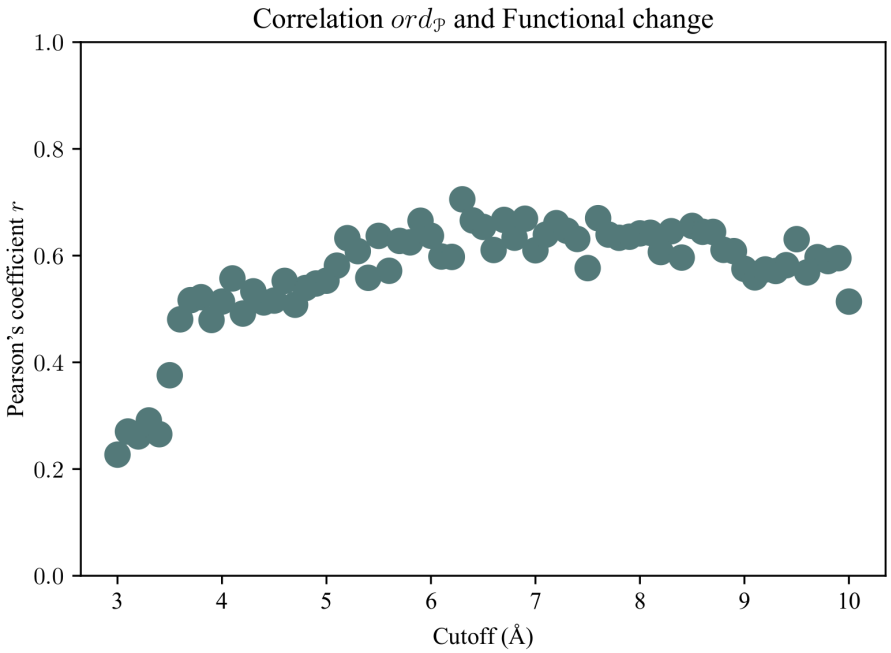


Figure 4.11: Pearson correlation r between number of nodes incident to a link in the perturbation network and functional change. Each circle corresponds to the correlation between $ord_{\mathcal{P}}$ and functional change. The correlations are calculated using 71 cutoffs in the interval 3–10 Å.

ranks (Subsection 4.2.6), from which $20 \times 71 = 1420$ belong to FSP. The matrix $\mathcal{R}$ shows a concentration of FSP having large $ord_{\mathcal{P}}$ values, all cutoffs combined (Figure 4.12). A total of 94% of the ranks (1334) corresponding to FSP, all cutoffs combined, are within the first half of the ranking, that is, within the first 42 ranks.

In other words, a rank belonging to a functionally sensitive position, is in the top half with a probability of 0.94. The probability of the rank of a FSP being in the top 25 ranks is 0.7 (1002 ranks). If we constraint the matrix only to cutoffs between 4–10 Ångströms, this probability is equal to 0.75 (915 ranks). Finally, the first 20 ranks, all cutoffs considered, are shared between FSP (60%, 844 ranks) and the rest of the positions (40%, 576 ranks), however only 24% (20 positions out of 83) of the total number of the positions are FSP (Figure 4.12).

Moreover, under cutoffs 4–10 Ångströms, the mutations perturbing the

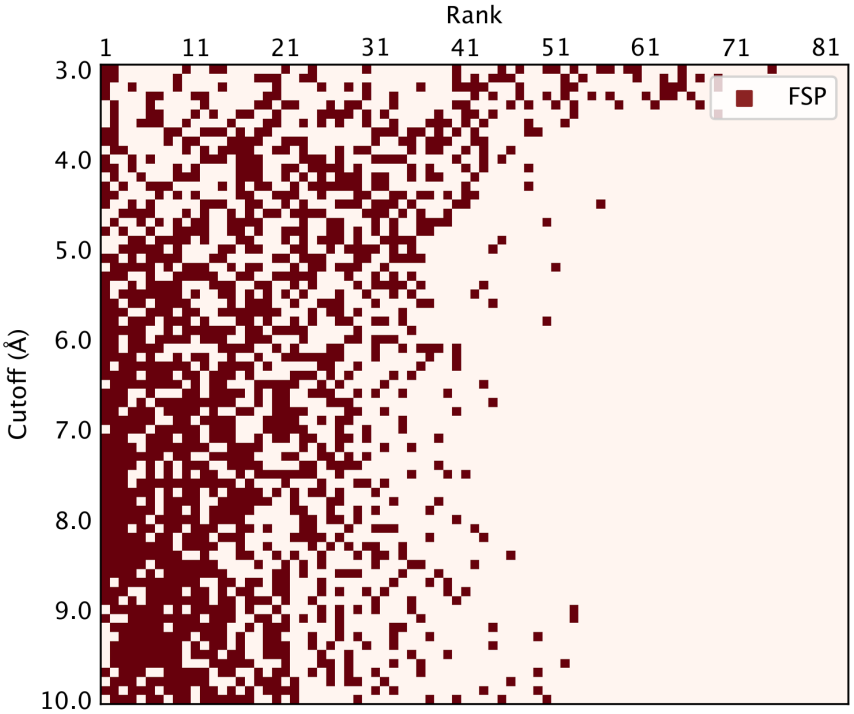


Figure 4.12: The $71, 83$ boolean matrix $\mathcal{R}$ highlighting the ranks of functionally sensitive positions by decreasing $ord_{\mathcal{P}}$ (Subsection 4.2.6). Entry $\mathcal{R}_{i,j}$ corresponds to 1 if the j-th rank at cutoff i corresponds to a functionally sensitive position and 0, otherwise.

interactions of a large number of amino acids, also impact a large number of interactions among FSP. This suggests a compact perturbation network, in which paths to go from one node to any other have few links. We have seen, however, that these perturbation networks do not correlate better with functional change by comparing the values of the total change in atomic interactions $\mathcal{W}_{\mathcal{P}}$ to functional change (Figure 4.9). Positions functionally sensitive to mutations, even though supposing compact perturbation networks, are differentiable from the rest of the positions with compact perturbation networks by the number of amino acids in their perturbation networks. Specifically, when interactions are modeled using cutoffs ranging from

5–7 Ångströms, compact perturbation networks decrease to be correlated to functional change, and the size of the perturbation network, however, stands for more than 60% of the functional change (Figures 4.9 & 4.11).

We have seen that a change in a large number of atomic interactions, does not necessarily translates to a functional change as shown by the values obtained for all positions in terms of $\mathcal{W}_{\mathcal{P}}$. Moreover, the number of nodes in the perturbation network ($ord_{\mathcal{P}}$) is correlated with functional change. This implies that functional change is mostly explained by changes in amino acid interactions, as opposed to atomic interactions. In other words, the change in atomic interactions alone, fails to explain the functional change produced by a mutation: It is of greater relevance to functional change, a larger number of perturbed amino acids by the mutation, independently of the number of perturbed atomic interaction between them.

The measure $\mathcal{W}_{\mathcal{P}}$ calculates the gross difference in atomic interactions, without discriminating the number of amino acids perturbed. The number of amino acids perturbed, however, is more relevant to the functional change of a protein after a mutation. A mutation affecting a large number of atomic interactions on a small number of amino acids, is likely to be irrelevant to the functional change. On the other hand, a mutation affecting a large number of amino acid interactions, independently of the number of atomic interactions perturbed, is likely to be relevant to the function.

A possible cause is that interactions between amino-acid pairs is done in more than one path in the protein 3D structure. If the number of amino acids perturbed by the mutation is small, even supposing that the number of atomic interactions perturbed is large, a small number of interaction paths are perturbed by the mutation. In this case, interaction between amino acids can still be maintained through other interaction paths. Moreover, the interaction of amino acids through different paths would be a sort of error correction mechanism, in which two amino acids are able to communicate using more than one pathway. Under this logic, most mutations would be tolerated through the interaction of amino acids on alternative interaction pathways.

4.4 Conclusion

Proteins contain functionally-sensitive positions, where mutations significantly undermine the performance of the protein on the native function. These functional sensitivity is in general independent of the particular amino acid replacing the original, suggesting a standalone importance of the po-

sition. Here, we address the question of whether these positions present a characteristic local structure differing from the rest of the positions in the sequence. Moreover, if there is such a local structure, we seek to know what is the distance at which local structures start to diverge depending on the type of positions.

A local structure can be thought as the distribution of atomic coordinates around an amino acid at a given distance. Because the placement of the atoms in space is far from been regular The use of several cutoff distances to define the atomic interactions is

We used the number of perturbed atomic and amino acid interactions as a comparison point between functionally sensitive positions and the rest, we also considered the furthest position in the 3D structure perturbed by the mutation as a measure of structural change. Results showed a characteristic local structure for buried and surface positions using the three perturbation measures. A local structure starting at around 4 Ångströms was shown to increment the correlation between the measures and the position of the amino acids in the 3D structure (Figure 4.13). Correlation between the values of the amino acids of al cutoffs and their buriedness increased steadily after four Ångströms up to ten Ångströms (Figure 4.7). A local structure defined by a distance of less than four Ångströms didn't present any correlation with the amino acid positions in the 3D structure suggesting that such distances fell short to capture a characteristic neighborhood of surface or buried positions (Figures 4.10 & 4.8).

Moreover, the number of perturbed amino acids showed a good correlation with functional change for atomic interactions starting at 4 Ångströms. This correlation peaked when a distance cutoff of 6 Ångströms was used to define the atomic interactions (Figure 4.11). The rest of the perturbation measures did not show any relation with functional change, suggesting that the number of perturbed atomic interactions or the length in the 3D structure of the perturbation were not good indicators of functional change.

Indeed, we found that the number of perturbed atomic or amino acid interactions (Figures 4.9 & 4.14), are not characteristic to functionally sensitive positions, as they affect similarly the rest of the functionally robust positions. Figure shows the correlation between functional change and change in amino acid interactions ($\mathcal{D}_{\mathcal{P}}$). Where measure $\mathcal{D}_{\mathcal{P}}$ is defined as the number of links changed between before and after a mutation (Paragraph 4.2.3). On the other hand, a larger number of perturbed amino acids was found to be linked to functionally sensitive positions. We observe that this result could underlie several communication paths between amino acids as a method of error-correction. Further work needs to be done to test this hypothesis how-

ever.

Finally, the threshold of 4 Ångströms for characteristic local structures was made clear by the gap in the correlation between all perturbation measures and the position of amino acids in the 3D structure.

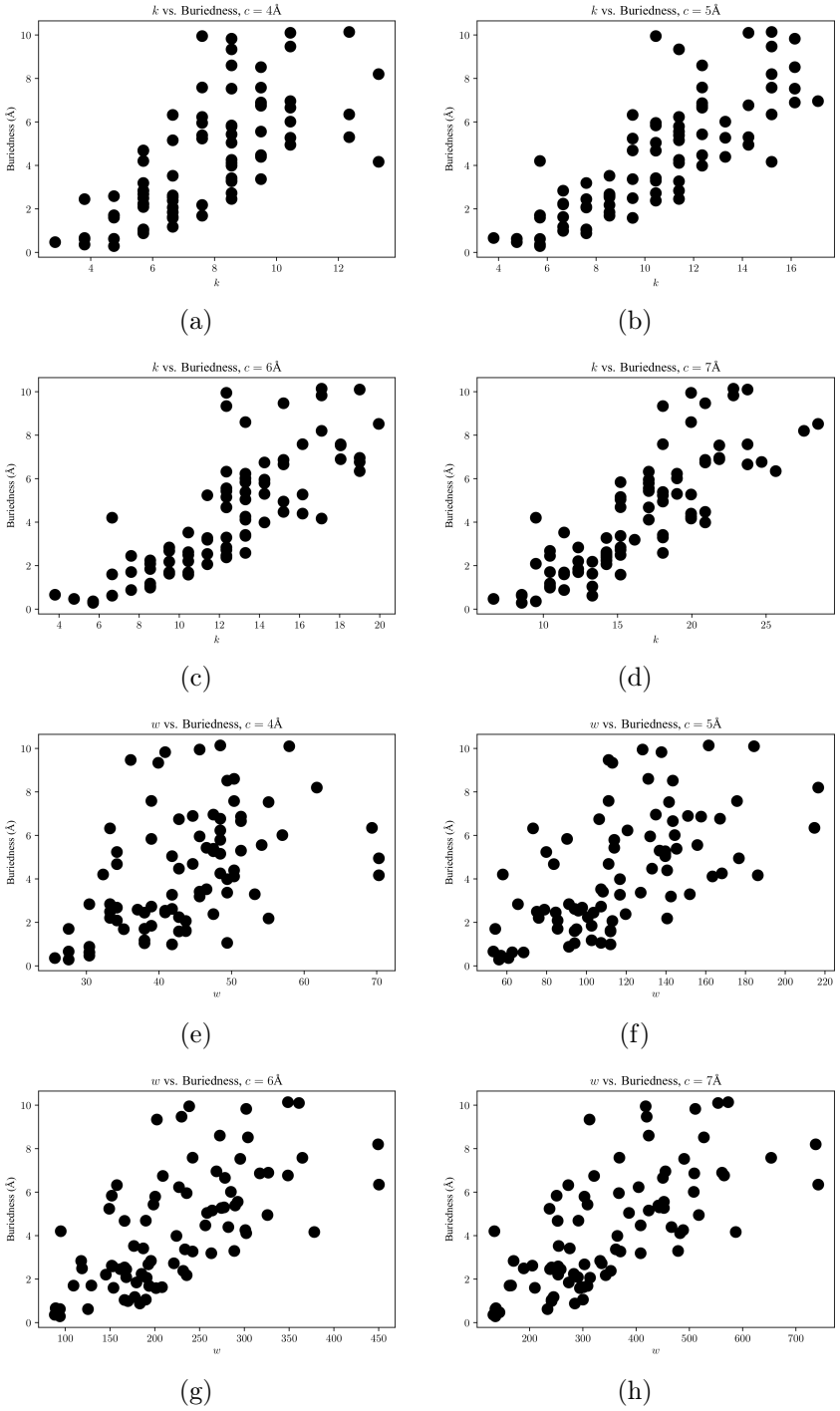


Figure 4.13: Degree k vs. Buriedness, using four distinct cutoffs 4, 5, 6, and 7 Ångströms, in (a), (b), (c), and (d), respectively. Weight w vs. Buriedness, using the same four distinct cutoffs is shown in (e), (f), (g), and (h), respectively.

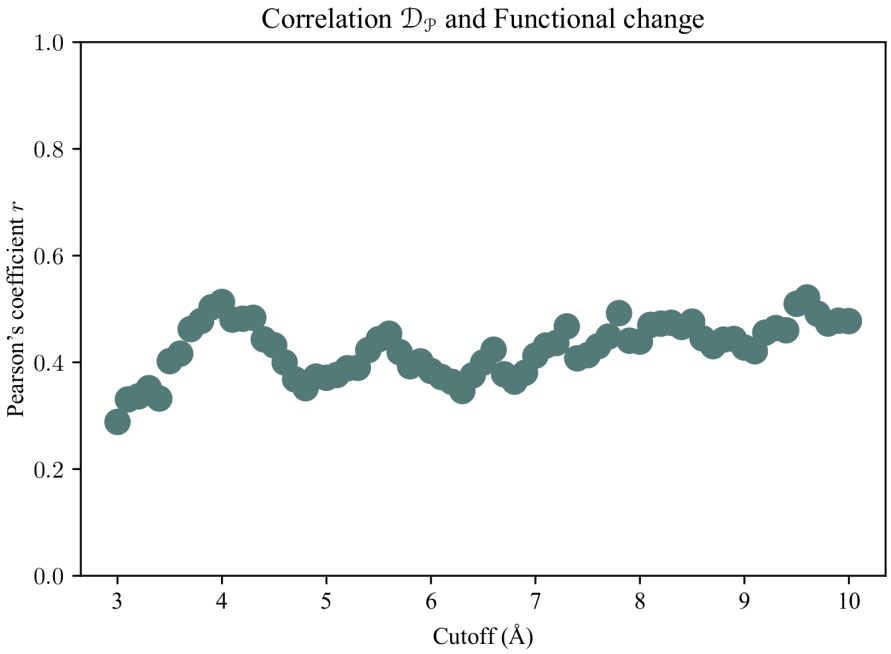


Figure 4.14: Pearson correlation coefficient r between perturbation measure $\mathcal{D}_\mathcal{P}$ and functional change. Each circle represents a correlation value between $\mathcal{D}_\mathcal{P}$ and functional change. The correlations are calculated for 71 cutoffs in the interval 3–10 Å.

Chapter 5

Void around amino acids: An algorithmic approach

5.1 Introduction

A protein is a biological object which faces a paradigm: on one hand it needs to be robust to exterior perturbations like in the case of mutations, and at the same time, be able to adapt to new biological conditions and functions.

Evidence shows that a minority of sequence positions are responsible for the fluctuations in protein function happening as a consequence of single mutations or during evolution. During the latter, a small set of co-evolving positions was found using multiple sequence alignment, in which subsets of co-evolving positions formed a functional motif called a functional sector [24]. When compared to the effects of mutations per position on the same protein, the so-called sectors were shown to be strongly correlated to the most sensitive positions to mutations [42].

Importantly, functional positions, when mutated, have an effect on the resulting function, generally when replaced by most amino acids; independently of the chemical properties of the amino acid type [42]. This gives a relevant place in functional changes to the sequence position, that is, positions are relevant to functional change by themselves when it comes to the consequences of mutations and evolution (Subsection 4.2.6).

This interesting result led us to study the protein structure with a focus solely on the positions of the amino acid sequence. In Chapter 2, we first introduced the idea that two structures could be structurally alternative, showing that structural change could perturb many interactions between regions in the structure but leave the the original groups connected, while other perturbations could affect the connectivity of a small number of interactions necessary for the connection between regions in the protein.

Subsequently, we measured the structural perturbation under a network approach, comparing across different *perturbation measures* the change on the atomic and amino acid surroundings of sequence positions (Chapter 4). Only one measure was found to be correlated to functional change of mutations: the number of perturbed amino acids by the mutation (Section 4.3). Functional change was not related to the number of interactions, results showed; whether it be atomic or amino acid interactions perturbed in the network. Both measures are strongly correlated to the buriedness of a position, that is, to the distance from the position to the protein surface (Section 4.3).

Based on the results introduced by McLaughlin et al. [42], showing as previously mentioned that some positions conduct the functional change due to mutations, and on our own results showing different structural consequences depending on what position of the sequence is mutated (Chapter 4), in this Chapter, we moved to a more general question keeping in mind the relevance

of positions in the sequence towards a functional or structural consequence by studying the *local void* of positions in the structure. In the previous Chapter, we were interested in the analysis of the atomic distribution around positions depending on their functional relevance to mutations, but also on their buriedness. Here, we study another structural property surrounding positions in the structure: their surrounding empty space.

The packing density of proteins, i.e. the study of the relative amount of free and non-free space inside the protein, has been previously studied. A number of techniques have been used, based on the study of cavities with a space-filling model, using the Voronoi diagram or Delaunay triangulation of the atomic coordinates of the structure to capture the free space within in the protein [36, 51, 54]. However, here we are interested in the systematic study of local void around amino acids, without considering the overall packing of the protein, particularly to help shed light on the question of how void is distributed across positions in the protein structure. More specifically, we seek to study the void across distinct types of positions; e.g. are buried positions surrounded by more or less void than surface positions?

Let us recall that void around residues is not mathematically well-defined, therefore, we use three different approaches to measure void around amino acids, each one based on a different algorithm. We used well defined geometrical objects like the convex hull or the Delaunay tessellation of a set of points in $\mathbb{R}^3$ to calculate volumes and voids. The two first methods presented here depend on a distance cutoff for the definition of void, with the idea of ignoring mere bulk space within or around the protein. The third method calculates void that is "trapped" inside the protein and uses no cutoff.

In summary, the first method constructs an *envelope set* of points around the target residue to compute void as the volume of the envelope set minus the volume of the convex hull of the target residue. The second method, uses the Delaunay tessellation of the atomic coordinates of the protein to measure the void of a residue as the sum of the volumes of adjacent tetrahedra to the residue. Finally, the third method is based on a concept of "trapped" empty space, that is, constrained inside the protein, ignoring the empty space that has direct access to the solvent of the protein. Atom radii are considered as well in the third method, indeed, the overlap between atoms is used for the location of void trapped inside the protein, inspired on a previously used method [36].

5.2 The protein as a discrete mathematical object

Given a protein structure, we want to measure the void or empty space surrounding amino acids inside the protein. In this Chapter, we propose three different methods that calculate the empty space from a PDB file containing the atomic coordinates of a protein. The calculation of the void is based on a model of the protein structure as a discrete object which is characterized by its set of 3D atomic coordinates $S \subset \mathbb{R}^3$. Henceforth, we will use the term protein structure referring actually to the set S, and *vice versa*. In the same fashion, an atom in the protein is a point $a \in S$, and a residue r is a subset of S.

The collection of residues $\mathcal{R}$, is a partition of S such that:

1. $\bigcup_{r \in \mathcal{R}} r = S$, and
2. $\forall r_1, r_2 \in \mathcal{R}; r_1 \cap r_2 = \emptyset$.

The union of the all residues in $\mathcal{R}$ amounts to the complete set of atoms, and not two residues have atoms in common in the protein.

A discrete representation of a protein structure is shown in Figure 5.1, where the set of atomic coordinates S is shown in grey, and atoms are spheres. A residue is shown within S in the structure (Figure 5.1b, blue), with neighboring atoms at distance 3 Ångströms or less (Figure 5.1a, orange).

5.3 Convex Hull Method

In geometry, a polyhedron is a solid in the 3-dimensional Euclidean space bounded by planar polygonal faces. The corners of the polyhedron, and the edges joining the faces are, respectively, the vertices and the edges of the polyhedron. A polyhedron is said to be convex if any line segment between two points within the polyhedron is also in the polyhedron (Figure 5.2 for an example of a convex polyhedron).

A central concept in computational geometry is the *convex hull*. The convex hull of a set of points is the minimal convex polyhedron containing them (Figure 5.3). The convex hull of a set of two-dimensional points is a polygon, while the convex hull of a three-dimensional set is a polyhedron (Figure 5.3 and Figure 5.2, respectively).

The point set S which is a subset of $\mathbb{R}^3$, is in general position, meaning that there are no 4 points of S on a same plane. The point set S in general position is true for our complete database of atomic coordinates, but this

(a) (b)

Figure 5.1: Representation of the protein structure of activity-regulated, cytoskeleton-associated repressor protein (PDB code 1BAZ). (a) The residue Asparagine in chain A at position 11 (blue) is shown with its neighborhood at distance 3 Å (orange). The neighboring amino acids include residues leucine, glycine, valine, and asparagine (LEU, GLN, VAL, and ASN respectively). (b) Zoom out of the residue showing the rest of the atomic coordinates of the protein (grey).

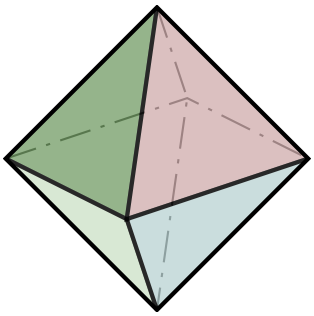


Figure 5.2: An Octahedron is a convex polyhedron with eight faces, twelve edges and six vertices.

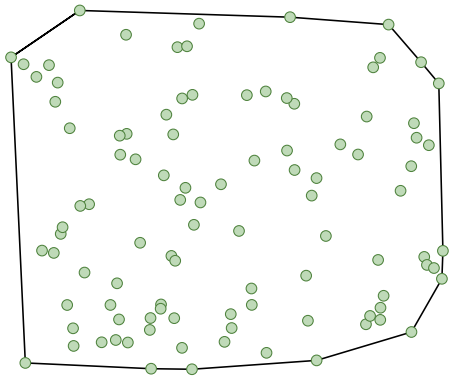


Figure 5.3: A set of points in 2D bounded by its convex hull.

condition is not necessarily respected in all pdb files; indeed, this condition is a consequence of the form of our data. Therefore each face of conv(S) is triangular. The Convex hull of the subset of atoms of the protein (a residue) is therefore a convex polyhedron with triangular faces. In this method, we model residues as the convex hull using their sets of atomic coordinates, and the volume of the residue as the volume of its convex hull.

5.3.1 Envelope set

We model the volume of a residue $r \in \mathcal{R}$ as the volume of its convex hull, noted conv(r); an idea to measure the void around r is to construct a set E_r, whose convex hull "wraps" or "envelops" r, and take the difference between the volumes of conv(r) and conv(E_r). The void of a residue, is thus defined by its atomic coordinates together with the set of points E_r (Figure 5.4). For convenience, the set $E_r \subset \mathbb{R}^3$, is called here the envelope set of r[1]. The set E_r is said to be an envelope set of r, if conv(r) $\subseteq$ conv(E_r). In order to measure the void around a residue r, we select an envelope set E_r that respects the following three conditions. Call V the set of vertices of conv(E_r), and let $p \in \mathbb{R}^3$:

1. If $p \in$ conv(E_r), then $dist(p, \text{conv}(r)) \leq 5$Å; where $dist(p, \text{conv}(r))$ is the Euclidean distance from p to conv(r)[2].

[1]This term 'envelope' is used differently and broadly in mathematics, usually referring to objects unconnected to the one used here.

[2]The distance from a point $p \in \mathbb{R}^3$ to a convex hull is the minimal orthogonal distance

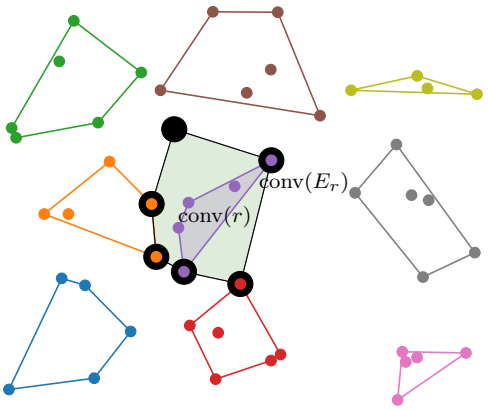


Figure 5.4: Representation of the void of a residue $r \in \mathbb{R}^2$ in a hypothetical protein structure $S \subset \mathbb{R}^2$. Each color corresponds to a different residue, i.e., points of a same color are atoms of the same residue. The convex hull of each residue is shown together with the one of the envelope set E_r of r with vertices filled in black. The envelope set E_r of r is the set of vertices of conv(E_r) together with the vertices of conv(r). The void of r (green) is defined by the difference between the area of conv(E_r) and of conv(r) (purple). Note that one point of E_r is not an atom of the protein (black point).

2. If $p \in S$ and $p \notin r \subset S$, then $p \in V$.

3. If $p \in r$ then $p \in \text{conv}(E_r)$.

Note that, if $p \in V$ and $p \in r$ then the point p is in E_r by the definition of the convex hull. So the set r is an envelope set of itself.

It is important to note that a point of E_r does not necessarily belong to S, i.e., is not necessarily an atom of the protein (Figure 5.4) and any point of E_r, is at chemical distance (5 Å [21]) from the residue r according to Condition 1. It ensures that only chemically relevant points are taken into account to measure the void, to exclude the general "bulk" space in the molecule from the computation of the empty space around the residue. This is because not all empty space between two atoms should count as void. Moreover, Condition 2 says that any point of $S \setminus r$ in conv(E_r), is a vertex of conv(E_r). This ensures that the region between the boundaries of conv(r) and conv(E_r) is indeed empty space around residue r and does not contain

from the point p to one plane spanned by a face of the convex hull.

parts of another residue of S. Condition 3 makes sure that void is at the exterior of points of residue r.

5.3.2 Basic idea

Void or free space, can be thought of as a potentially interactive region in the molecule with an absence of atoms. Here, 5 Ångströms (Å) is the interaction distance between pairs of atoms, as most atomic interactions happen under this threshold [21]. The idea of computing the void surrounding each residue of the protein, the *local void* of a residue, is to find the neighboring empty regions at chemical reach from each residue within the protein. To do so, we consider the atomic coordinates of each residue and, we then select a subset of atoms to create an *envelope set* "wrapping" the residue.

Let S to be the atomic coordinates of a protein, and $r \subset S$ the atomic coordinates of a residue, or simply a residue of S. The spatial difference between the convex hulls of the envelope set E_r of r and r, is the local void of r (Figure 5.4). Let's recall that the envelope set does not necessarily have all its points in S; but it is conditioned to exclude the bulk space of the molecule by selecting only points at chemical distance from r.

A useful tool to model the boundary of a set of points r in $\mathbb{R}^3$ is the so-called convex hull of r, noted $\text{conv}(r)$. The convex hull, is an approximation of the shape of the residue, and is used here to compute its volume. If we compute the volume of a residue, we do the same for the volume of the convex hull of the envelope set E_r, to obtain the volume of the local void of r as:

$$\text{Void}(r) = V_{E_r} - V_r.$$

Where V_{E_r} and V_r are the volumes of $\text{conv}(E_r)$ and $\text{conv}(r)$, respectively. Note that the volume of r does not take into account the radii of its atoms, usually given by the van der Waals radii. However, this is also true for the points of the envelope set, removing a bias from the method. For our purpose the method is sufficient for its congruence with other methods used to calculate the volumes of amino acids [20, 40]. The correlation between volumes of 151,045 residues using the Pearson correlation coefficient yields 0.74 obtained using the convex hull and method [40], respectively.

The set E_r being the envelope set of r, is conditioned to contain the set r (Condition 3 in the previous Subsection):

$$r \subseteq E_r.$$

Cases when $r = E_r$ allow non-existent local void, which happens only rarely depending on the properties of the faces of conv(r) relative to the placement of the points in $S \setminus r$, as explained by the construction of E_r in Subsection 5.3.3. By definition, void is empty space, therefore the only atoms of the molecule inside conv(E_r), should belong to r. In other words, atoms of the structure in the neighborhood of r belonging to the envelope set E_r, should be on the boundary of conv(E_r) (Condition 2). Finally, an atom of r in the envelope set E_r, is at chemical reach, less than 5 Ångströms, from r (Condition 1). The construction of the envelope set of each residue in the structure is explained in the following Subsection.

5.3.3 Algorithm

The local void of a residue $r \subset S$, where S is the set of atomic coordinates of a protein structure, is obtained considering the volume of a larger set acting as an "envelope set" of r, from which we subtract the volume of r. The volume of both sets is calculated using their convex hulls. The envelope set is noted E_r and its convex hull, conv(E_r) contains conv(r). The empty space between conv(r) and conv(E_r) is defined as the void of r. The construction of E_r is done by iteration on the faces of conv(r), where each face is defined by three points of r. The algorithm selects all the points in r and for each face at most one point of $\mathbb{R}^3 \setminus r$ gradually as follows:

1. For each face Δ of conv(r), we select all points of S whose orthogonal projection points on the plane spanned by Δ are inside the Δ (Subsection 5.3.4). In Figure 5.5a, the projection of all four points to the plane lie on Δ (triangle).

2. If the projection of more than one point lies on Δ, like in Figure 5.5a, then we select the point with the shortest distance to the face Δ (a_1, in the case of Figure 5.5a). If, on the other hand, the orthogonal projection of no point lies on Δ, we select none.

3. If a point p is selected by Step 2 above, then we add p to E_r but *only if* its normal distance to Δ, is at most 5Å. In Figure 5.5b, the point a_1 is projected within the face of conv(r), however, a_1 is not added to E_r because is too far (14.9Å) from the face.

4. If the point p is the closest point to conv(r) selected in Step 2 is too far, we simply approach discretely p across the normal vector of Δ until the resulting point p' is at distance less or equal to 5 Å from Δ; then we

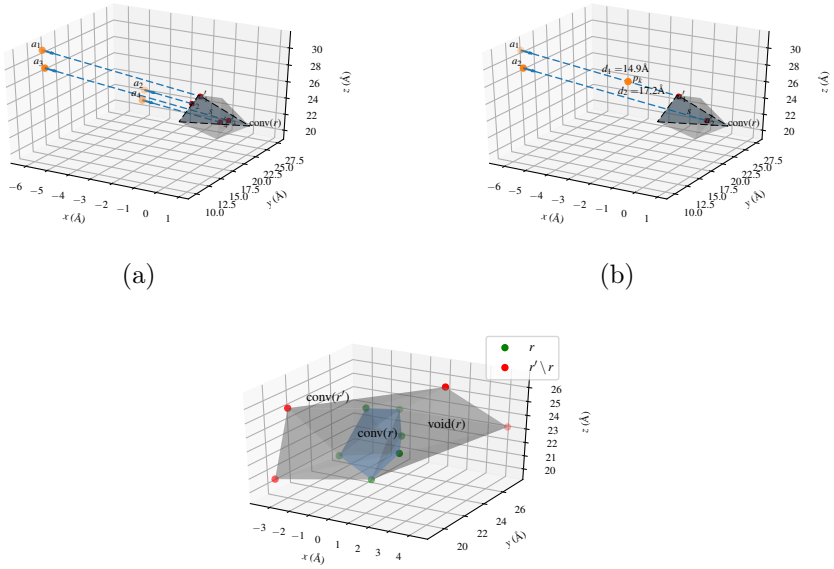


Figure 5.5: (a) The points a_1, a_2, a_3, and a_4 are projected orthogonally to the plane spanned by face s (blue triangle) of convex hull conv(r) of r. The projection vector is represented by dashed blue segments, and the projected points a'_1, a'_2, a'_3, and a'_4 are shown in red. (b) The points a_1 and a_2 are projected orthogonally within the face s of conv(r). The projection vectors are represented by dashed blue segments, and the projection points a'_1 and a'_2 are shown in red. The distances between the points a_1, a_2 and s are noted d_1 and d_2, respectively. The point p_k lies between a_1 and face s, and is at distance 5 Å from s. (c) Example of the void of a residue r. The atomic coordinates of r (green) define conv(r), and the set r' defines conv(r'). The envelop set r' of r was computed using the surrounding atoms from r (Steps 1 to 5). The void of r, noted void(r), is defined as the difference between the volumes of conv(r) and conv(r'). This example is based on the Cholera Toxin protein with PDB code 1EEI used in Chapter 2; where r is residue at position 103 in chain H.

add p' to E_r. In Figure 5.5b, the point p_k is obtained by approaching a_1 across the normal vector of the face until its distance is 5 Å from the r.

5. Finally, we add the three vertices of the face Δ of conv(r) to the envelope set E_r.

It follows from Step 5, that $r \subset \text{conv}(E_r)$ respecting Condition 3. Moreover, if a point $p \in S \setminus r$ is in E_r, we know from Steps 3 and 4, that p is a vertex of conv(E_r), respecting Condition 2. Finally, Condition 1 stating that any point in E_r is at chemical reach from conv(r), follows from Steps 3 and 4.

An example of the 3D convex hulls of r and the envelope set of r noted r', are depicted in Figure 5.5c. The void of r is the difference in the volumes of the convex hulls of r and r', and the envelope set r' is obtained using Steps 1 to 5 above.

The convex hulls are computed using the Scipy library for the programming language Python [30], and the void is part of the biographs package written for the purpose of this work and can found at `https://github.com/rodogi/biographs`. Finally, the pseudo-Algorithm 1 formally describes the previous process, and an explanation of how to decide whether an orthogonal projection of a point to a plane lies within a triangle in the plane is explained in Subsection 5.3.4, using the barycentric coordinates of the triangle.

5.3.4 Barycentric coordinates

To know whether the orthogonal projection of a point in the structure S to the plane defined by a face of conv(r) lies on the face, we can use the system of barycentric coordinates of the triangular face. The process of finding the barycentric coordinates system of a triangle is explained in more detail in [63].

We can think of the barycentric coordinates system as a non-orthogonal system of coordinates having as a basis the vectors defined by two edges of the triangle. For example if a, b, and c are the vertices of the triangle, then we can choose $\vec{u} = b - a$ and $\vec{v} = c - a$ as the basis, and a as the origin of the coordinates system (Figure 5.6). This way, any point $p \in \mathbb{R}^3$ lying on the plane of the triangle abc can be written as

$$p = a + \beta\vec{u} + \gamma\vec{v}.$$

We can rewrite the previous equation to get:

$$p = (1 - \beta - \gamma)a + \beta b + \gamma c,$$

Algorithm 1 Return the envelope set set E_r of residue $r \subset S$.

procedure VoidConvexHulls(r, S)
  $E_r = \emptyset$
  for Each face Δ of conv(r) **do**
    Let P be the set points of $S \setminus r$ orthogonally projected to Δ, and remove from P the points on the half-space containing conv(r)
    Let $\vec{n}$ be the unit normal vector of Δ
    if P is non empty **then**
      Let $p \in P$ be the closest point to Δ, and
      let $d = dist(p, \Delta)$ be that distance.
      if $d \leq 5$ **then**
        Add p to E_r
      else
        $\lambda = d - 5$
        Define $p' := p - \lambda\vec{n}$
        Add p' to E_r
      end if
    end if
    Add the vertices of Δ to E_r
  end for
  return E_r
end procedure

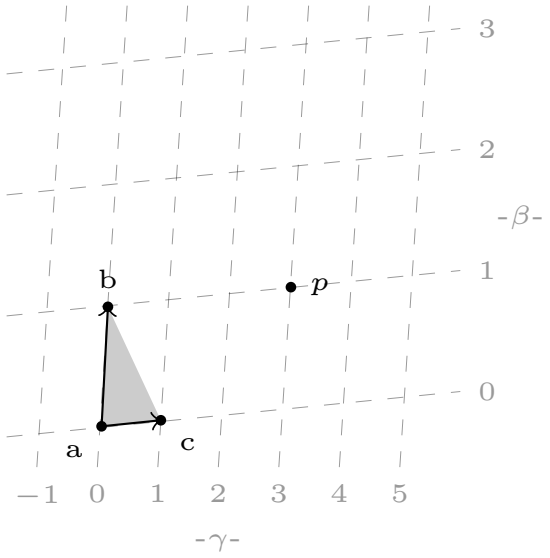


Figure 5.6: A triangle can be used to obtain a non-orthogonal coordinate system of the plane. The origin of this coordinate system is point a and vectors $\vec{u} = c - a$ and $\vec{v} = b - a$ can be used as its basis. Any point in the plane can be then represented by an ordered pair (γ, β). For example $p = a + 3\vec{u} + \vec{v}$, where $\gamma = 3$ and $\beta = 1$ in such a coordinate system.

and define a new variable $\alpha = 1 - \beta - \gamma$ to have:

$$p = \alpha a + \beta b + \gamma c.$$

With the constraint that $\alpha + \beta + \gamma = 1$. Another way to compute the barycentric coordinates is to compute the areas A_a, A_b, and A_c, of the sub-triangles obtained by partitioning the triangle into three sub-triangles joining each vertex a, b, c to a fourth point in the plane. Barycentric coordinates respect the rule:

$$\alpha = A_a/A, \beta = A_b/A, \gamma = A_c/A,$$

where A is the area of the triangle.

For a triangle in $\mathbb{R}^3$, its normal vector $\vec{n}$ is the vector perpendicular to any vector lying on the plane defined by the triangle. This includes of course the edges of the triangle. One way to obtain $\vec{n}$ is by taking the cross product of two vectors on the plane; in our case, we can take two edges of the triangle:

$$\vec{n} = (b - a) \times (c - a).$$

The area of the triangle can be found by taking the length of the cross product:

$$\text{area}_A = \frac{1}{2}\|\vec{n}\|.$$

This is not a signed (positive or negative) area so it cannot be used directly to calculate the barycentric coordinates (as we need to calculate the other areas with a normal vector on the same direction). For this goal, we consider the dot product of two parallel vectors to know if both vectors have the same direction:

$$\vec{u} \cdot \vec{v} = \|u\|\|v\|\cos(\theta),$$

where $\cos(\theta) = 1$ if both have the same direction and $\cos(\theta) = -1$, otherwise. Using the previous formulas and a point p on the same plane as the triangle abc, we can calculate the barycentric coordinate α as,

$$\alpha = \frac{A_a}{A} = \frac{\|(c - b) \times (p - b)\|}{\|\vec{n}\|} = \frac{\|(c - b) \times (p - b)\|\|\vec{n}\|}{\|\vec{n}\|^2}.$$

Given that $\vec{n_a} = (c - b) \times (p - b)$ and $\vec{n}$ are parallel and have the same direction, we have

$$\alpha = \frac{\vec{n_a} \cdot \vec{n}}{\|\vec{n}\|^2}.$$

In a similar way we get

$$\beta = \frac{\vec{n_b} \cdot \vec{n}}{\|\vec{n}\|^2}, \gamma = \frac{\vec{n_c} \cdot \vec{n}}{\|\vec{n}\|^2},$$

where

$$\vec{n_b} = (a - c) \times (p - c) \text{and } \vec{n_c} = (b - a) \times (p - a).$$

Finally, if

$$0 \leq \gamma < 1 \qquad 0 \leq \beta < 1,$$

the projection of the point $p = \alpha a + \beta b + \gamma c$ to the plane defined by triangle abc lies inside of the triangle.

5.4 Delaunay Method

The following method to compute the local void of residues is based on a triangulation or tessellation of the set of atomic coordinates S. More specifically, we use the *Delaunay triangulation* of S, noted $\mathcal{D}(S)$. The Delaunay triangulation $\mathcal{D}(S)$ is derived from the *Voronoi* diagram of S, so in order to define $\mathcal{D}(S)$, we first need to introduce the Voronoi diagram of S.

Let each point $p \in S$ be expanded into a region where every point in the region is closer to the point p than to any other point in S. Such a region is called a cell of the Voronoi diagram of S. Each cell contains all the points that are closer to exactly one point of S. The border or edge between two cells contains the points equidistant from the two adjacent points in the Voronoi diagram. In Figure 5.7, edges are equidistant to the points of S (bluepoints) inside the two adjacent cells separated by each edge (cells). Vertices of the Voronoi diagram are by extension equidistant to the points whose cells meet at each vertex. When the point set used to create the diagram is in general position, every vertex is equidistant to exactly three points (A point set is not in general position if four points lie on the same circle). In that case, the *dual graph* of the Voronoi diagram is called the Delaunay triangulation, and it's unique.

In graph theory, the dual graph of a partition of the plane is a graph which has one vertex for each face (cell) of the plane and two vertices are connected if their corresponding faces are adjacent. In figure 5.8, the points defining the Voronoi diagram are the vertices of the dual graph and two vertices are connected by an edge in the dual graph (blue segments), if the corresponding cells of each incident vertex are adjacent in the Voronoi diagram.

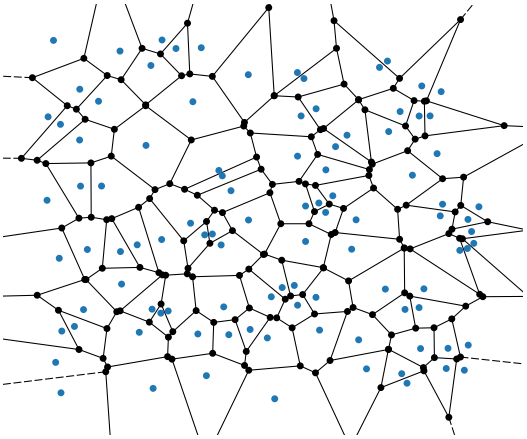


Figure 5.7: The Voronoi diagram of 100 points in general position in the plane. Each blue point in a cell is closer to any other point in that cell than to any other blue point. The vertices of the Voronoi diagram (black points) are equidistant to the blue points in the adjacent cells.

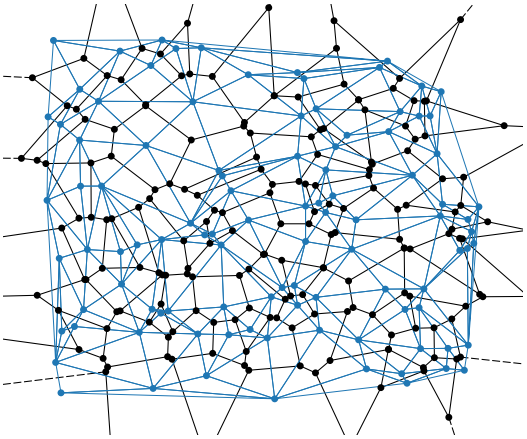


Figure 5.8: The dual graph of the Voronoi diagram. Each vertex of the graph corresponds to one cell of the diagram. Two vertices of the dual graph are connected by an edge if the two corresponding cells are adjacent in the Voronoi diagram. It may seem that some non-bounded cells are not adjacent but their respective edges intersect beyond the limits of the figure.

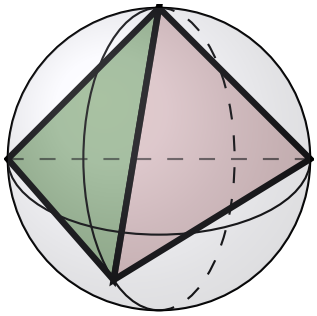


Figure 5.9: A tetrahedron and its circumscribed sphere (gray): the vertices of the tetrahedron are on the surface of the sphere.

The concept of the Voronoi diagram in 2D can be extended to a set $S \subset \mathbb{R}^3$. It suffices to think of each cell as a polyhedron instead of a polygon containing the points closer to exactly one point in S. Two adjacent cells share a face of their polyhedron, and three adjacent cells of the diagram share one edge. If the set S is in general position, a vertex of the Voronoi diagram is equidistant to at most 4 points in S.

The dual-graph of the Voronoi diagram of a 3D set, is a tessellation where adjacent cells form a tetrahedron. Here the circumscribed sphere centered at the vertex in the Voronoi diagram is equidistant to the four vertices of the dual graph only (Figure 5.9).

If $S \subset \mathbb{R}^3$ is in general position (there is no five points on a same sphere), a Delaunay tessellation $\mathcal{D}(S)$ of S, is the dual-graph of the Voronoi diagram of S, and it's unique. Each tetrahedron in $\mathcal{D}(S)$ satisfies the empty-sphere property, where its circumsphere contains no point of S in its interior.

5.4.1 Basic idea

We aim at using the tetrahedra produced by the Delaunay triangulation of the set of atomic coordinates to calculate the void around residues. In particular, we use the empty sphere property of the tetrahedra: to each simplex in the Delaunay triangulation there is one circumscribed empty sphere (with no points in its interior), which implies that the length of the edges in the triangulation is in principle minimized. The basic idea is to select a set of tetrahedra adjacent to one residue and think of their volume as the local void of that residue. To compute this void it then suffices to calculate the volume of each tetrahedron and take the sum of all to be the local void the residue.

Algorithm 2 Return local void of residue $r \subset S$ for a protein structure S, and a cutoff $c \in \mathbb{R}$.

procedure VOIDDELAUNAYTRIANGULATION(r, S, c)
  Let $\mathcal{D}(S)$ be the Delaunay tessellation of S.
  $\mathcal{T}_r = \emptyset$
  for $x \in r$ **do**
    $N_x = \{v : v, x \in V(\Delta)$ for some $\Delta \in \mathcal{D}(s)\}$,
    where $V(\Delta)$ is the set of vertices of Δ.
    $E = \{n_x \in N_x : dist(n_x, x) < c$ and $n_x \neq x\}$
    for $y \in E$ **do**
      $N_y = \{v : v, y \in \Delta$ for some $\Delta \in \mathcal{D}(s)\}$
      $T = \{(y, n_y) \in E \times N_y : dist(n_y, y) < c\}$
    end for
    for $(y, z) \in T$ and $w \in E$ **do**
      if Both (w, y) and (w, z) are in T **then**
        Δ^{wxyz} is the tetrahedron with vertices w, x, y, z
        if $V(\Delta^{wxyz}) \cap r \neq V(\Delta^{wxyz})$ **then**
          Add Δ^{wxyz} to $\mathcal{T}_r$
        end if
      end if
    end for
  end for
  return $\mathcal{T}_r$
end procedure

For that purpose, let $\mathcal{D}(S)$ be the Delaunay tessellation of the set of atomic coordinates $S \subset \mathbb{R}^3$. We know that the circumscribed sphere of any tetrahedron $\Delta \in \mathcal{D}(S)$, contains no atomic coordinates in its interior. Moreover, because all points of S are included in $\mathcal{D}(S)$, the vertices of the set of tetrahedra surrounding a residue contain the entire atomic neighborhood of the residue. In order to compute the void, we first select all tetrahedra having at least one vertex in the residue (each vertex is an atomic coordinate) and at most three. The volume of each tetrahedron represents the empty space within its four vertices. Therefore, we can measure the empty space around a residue by considering the tetrahedra that are adjacent to the residue. This excludes any tetrahedron having all four vertices in the residue as its volume would account as empty space inside the residue instead of space around it.

Let $r \subset S$ be a residue, and $\mathcal{T}_r$ be the set of tetrahedra in $\mathcal{D}(S)$, with at

least one vertex in r and at most three vertices in r, that is:

$$\mathcal{T}_r := \{\Delta \in \mathcal{D}(S) : 1 < |V(\Delta) \cap r| < 4\},$$

where '||' denotes the cardinality of the set and $V(\Delta)$ is the set of vertices of tetrahedron Δ. In other words, the tetrahedra having at least one vertex in the residue and at least one in a different residue. Note that tetrahedra with no vertex in r are not considered in the computation of the local void of r, as it would account as empty space not adjacent to the residue. Finally, we compute the void of r as

$$\text{Void}(r) = \sum_{\Delta \in \mathcal{T}_r} V_\Delta,$$

where V_Δ is the volume of tetrahedron Δ.

As with the previous method, we restrict the void to be within chemical distance: We constraint $\mathcal{T}_r$ to have tetrahedra with only vertices at chemical reach from r. For this purpose, wee state the condition that a tetrahedron $\Delta \in \mathcal{T}_r$, if and only if, for any edge e of Δ, we have:

1. length of $e < c \leq 5$ Å

Where c is a positive real number. Usually we will let c to be equal to 5 Ångströms, as this distance respects most chemical interactions happening between atoms, and the distribution of the length of the edges of tetrahedra in $\mathcal{D}(S)$ for 750 proteins shows a considerable change around 5 Ångströms.

5.4.2 Algorithm

In order to obtain the set $\mathcal{T}_r$ from residue $r \subset S$, we implement an algorithm iterating on the atoms of r. This algorithm first finds for each atom a, the set of tetrahedra adjacent to atom in $a \in r$. From this set of tetrahedra, we select the tetrahedra whose edges are shorter than a given cutoff $c \in \mathbb{R}$ (c being usually equal to 5). Finally, we select from that set, the tetrahedra with at least one vertex in a residue in $S \setminus r$. In other words, the tetrahedra in $\mathcal{T}_r$ must have at least one vertex outside of r (Algorithm 2). For each atom $x \in r$ in residue r, we select all 3-tuples of atoms (w, y, z) not all in r, such that there exist a tetrahedron in $\mathcal{D}(S)$ with vertices x, w, y, and z. The resulting tetrahedra set is one step away from $\mathcal{T}_r$, which results from removing all tetrahedra with at least one edge whose length is greater than c.

5.5 Empty tetrahedra method

The previous method defined the void of a residue in terms of the volume of the tetrahedra around the residue, taken from the Delaunay triangulation $\mathcal{D}(S)$ of the protein structure S. A restriction on the length of edges of the tetrahedra is imposed to discriminate between void and the more general bulk of the molecule. In this method, we redefine the void of a residue $r \subset S$ considering the radii of the atoms composing the residue. These radii are the atomic van der Waals radii taken from [20].

Moreover, we use the concept of *empty tetrahedra*; where a tetrahedron $\Delta \in \mathcal{D}(S)$ is empty if the length of every edge of Δ is greater than the sum of the radii of its endpoints (based on the concept of empty tetrahedra in [36]). Note that a non-empty tetrahedron has at least one pair of two-body overlapping atoms.

The empty space within Δ is defined by:

$$\text{void}(\Delta) = V_\Delta - \sum_{v \in \Delta} V^v_{\text{overlap}}$$

Where V^v_{overlap} is the volume of the overlap between Δ and the spheres centered at vertices $V(\Delta)$ and radius equal to the van der Waals radius of their respective atoms. The method to measure the overlap volume between atoms and the tetrahedron is explained in Subsection 5.5.1.

The use of empty tetrahedra allows for the possibility of defining bounded and non-bounded empty space. In this method, void is only considered if it is *bounded* empty space in the interior of the molecule, as opposed to empty space outside of the surface. An empty tetrahedron Δ is bounded, if there exist no path of adjacent empty tetrahedra $\Delta, \Delta_1, \Delta_2, \ldots, \Delta_n$, from Δ to a empty tetrahedron Δ_n, such that Δ_n is adjacent to a tetrahedron at the boundary or surface of $\mathcal{D}(S)$. The surface is modeled with the convex hull $\text{conv}(S)$ of the atomic coordinates of the protein S, and an adjacent tetrahedron to the surface shares one face with $\text{conv}(S)$. In the same fashion, two tetrahedra are adjacent if they share a face (i.e. three vertices). This approach guarantees that only empty tetrahedra containing void "trapped" inside the molecule are considered as opposed to empty space found in pockets and depressions near the surface of proteins.

5.5.1 Overlap between a sphere and a tetrahedron

The volume of the overlap between a sphere and a tetrahedron has no direct analytical calculation. However, the volume can be decomposed into

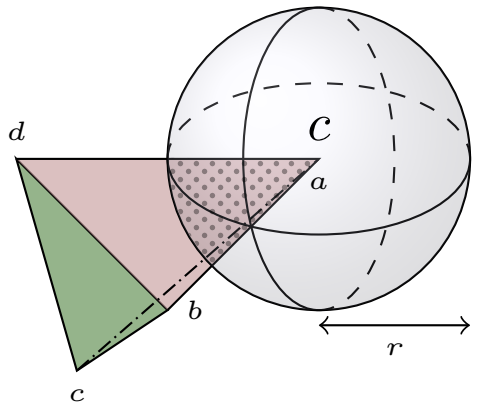


Figure 5.10: Overlap representation between a sphere centered at point c and radius r, and a tetrahedron with vertices a, b, c and d. Vertex a coincides with the center of the sphere and is also the only vertex of the tetrahedron to be inside the sphere. The overlap region between the sphere and the tetrahedron is represented by a doted pattern. Here, an atom is represented by a sphere using its van der Waals radius.

multiple basic parts which are analytically computable [66]. In our case, the overlapped sphere is centered at one of the vertices of the empty tetrahedron with the other three vertices located outside of the sphere (Figure 5.10). This simplifies considerably the overlap compared to cases where the center of the sphere is not a vertex of the tetrahedron and/or more than one vertex is inside the sphere [66].

A spherical cap is the part of a sphere that is cut by a plane (Figure 5.11). When a plane cuts the sphere through its center the cap is called a hemisphere. In our case, three faces of the tetrahedron intersect the sphere. Moreover, the faces intersect the sphere at its center, therefore each cap produced by a plane spanned by its faces is a hemisphere (Figure 5.10).

A spherical wedge is defined by the intersection of two planes through the center of the sphere (Figure 5.11). The angle of the wedge is equal to the dihedral angle of the planes' intersection. In our case, the dihedral angle of each wedge coincides with the dihedral angle of the intersecting faces of the tetrahedron.

The volume of the overlap between the tetrahedron and the sphere can be calculated starting with the volume of the sphere and removing each sub-volume not necessary for the calculation in an inclusion-exclusion procedure [66].

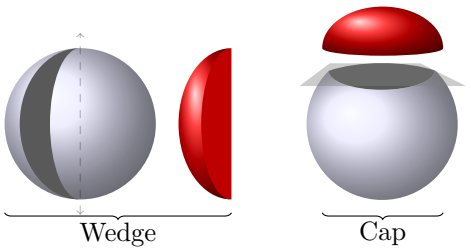


Figure 5.11: A wedge and a cap of the sphere.

Let V_{sphere} be the volume of the sphere, to compute the volume of the overlap, we first subtract from V_{sphere}, the volume of the caps defined by each face of the tetrahedron (exclusion). After the subtraction of the three caps from V_{sphere}, we remove two times each pairwise intersection of the caps. The next step is then to add back these volumes, i.e., the volume of each wedge defined by the intersecting edges of the tetrahedron with the sphere (inclusion). Finally, we need to subtract the three-wise intersection of these wedges (exclusion), conveniently, this volume coincides with the volume we seek to measure. We have:

$$V_{\text{overlap}} = V_{\text{sphere}} - \sum_{\text{faces}} V_{\text{cap}} + \sum_{\text{edges}} V_{\text{wedge}} - V_{\text{overlap}}.$$

Where V_{overlap} is the volume of the overlap of the tetrahedron and the sphere.

The only volume that is non trivial to calculate is the volume of the wedges. The volume of a wedge is defined by the dihedral angle of the intersecting planes forming it. If ϕ is such an angle, then

$$V_{\text{wedge}} = \frac{2\phi r^3}{3},$$

where r is the radius of the sphere. The dihedral angle of two planes can be obtained by computing the arc cosine of the dot product of their normal vectors. Suppose ϕ_1, ϕ_2, and ϕ_3 are the dihedral angles of the wedges defined by the three intersecting edges of the tetrahedron with the sphere. Then we can rewrite the previous formula to obtain the volume of the overlap as:

$$V_{\text{overlap}} = \frac{r^3}{3}\big[-\pi + \phi_1 + \phi_2 + \phi_3\big].$$

5.5.2 Bounded empty tetrahedra

As previously mentioned, an empty tetrahedron is characterized by having its edges of a length larger than the sum of the corresponding radii of the

Algorithm 3 Given a starting tetrahedron $\Delta_s \in \mathcal{D}(S)$, find its connected component in the graph where vertices are open tetrahedra which are connected by an edge if they share a face in $\mathcal{D}(S)$. Return the connected component $\mathcal{C}$, and a list of the "checked" or visited vertices by the depth first search algorithm.

```
1: procedure DepthFirstSearch(Δ_s)
2:     checked = ∅
3:     stack = Δ_s
4:     C = ∅
5:     while First element Δ in stack is empty do
6:         Append Δ to checked
7:         N_Δ = {Δ_n : Δ_n is adjacent to Δ}
8:         if A tetrahedron in N_Δ is on the boundary of D(S) then
9:             C = ∅
10:            checked = checkedΔ ∪ N_Δ
11:            return (C, checked)
12:        end if
13:        if At least one empty tetrahedron of N_Δ not in stack then
14:            while empty Δ_n ∈ N_Δ not in stack do
15:                Insert Δ_n to stack
16:            end while
17:        else
18:            Pop Δ from stack
19:            C = C ∪ stack
20:        end if
21:    end while
22:    return (C, checked)
23: end procedure
```

endpoints. This implies not only that the tetrahedron has empty space inside, but that the empty space is enclosed inside the protein.

An empty tetrahedron is non bounded if it exists on a path of adjacent empty tetrahedra ending on an open tetrahedron (one of its edges is larger than the sum of the endpoints' radii) at the boundary of the Delaunay triangulation $\mathcal{D}(S)$. Let the graph $G = (V, E)$ have as set of vertices all empty tetrahedra in $\mathcal{D}(S)$. An edge $\Delta_1\Delta_2 \in E$ exists between two vertices, if the corresponding tetrahedra Δ_1 and Δ_2 share a face. The set of bounded tetrahedra then can be found by removing from G all the connected components

having at least one open tetrahedron at the boundary of $\mathcal{D}(S)$ (Algorithm 3).

5.5.3 Algorithm

To compute the local void of each residue, we first get the set of connected components consisting of open bounded tetrahedra. This set is obtained by iterating over the Delaunay triangulation $\mathcal{D}(S)$ of the set S (Algorithm 3). Once we obtain such a set, we can calculate for each tetrahedron the volume of the overlap with the spheres centered at its vertices. Subtracting these volumes from the volume of the tetrahedron gives the empty space or void of the tetrahedron. This void accounts for a part of the void of each residue having at least one vertex of the tetrahedron. The sum of the void of each atom of the residue accounts for the total void of the residue.

The Delaunay triangulation of S was obtained using the Delaunay class from the spatial module of the Scipy package [30] and the code to obtain the void using the three methods presented here can be found at `https://github.com/rodogi/biographs`.

5.6 Results

In order to test the three methods, we calculated the local void of each amino acid in two structurally similar proteins: Cholera Toxin B pentamer and Heat Labile Enterotoxin B pentamer (PDB files are 1EEI and 1EFI, respectively). The two proteins have five same-length chains each, and are structurally superimposable. They differ in 19 positions of the sequence, where amino acids occupying these positions are not equal. That is, the toxins differ in $19 \times 5 = 95$ positions of a total of $83 \times 5 = 515$, which gives a nice framework for comparing the different values of the void measurements.

We are interested to know how different the void of each residue will be when measured by each method. Table 5.3 shows the sensitivity of the measures to the atomic positions in the 3D structures of the two toxins, rarely two residues have a similar values independently from the measurement used. The average difference ($CtxB_5$ - $hLBT_5$) between the measures is quite low, however, with an average difference of -6.26 $Å^3$, 7.57 $Å^3$, and -4.41 $Å^3$ for the method convex hull, Delaunay, and empty tetrahedra, respectively (Table 5.1). The empty tetrahedra method has a distribution of the difference of void with a mean and median closest to zero than the other two methods, however, it is the method showing a larger standard deviation. The fact that central tendency values mean and median are close to zero implies that most void difference is redistributed across the protein as a consequence of the

Table 5.1: Central tendency measures of the difference in void between $CtxB_5$ and $hLTB_5$ across the three methods.

Method	Mean (Å^3)	Median (Å^3)	Standard Deviation (Å^3)
Convex Hull	-6.26	-1.04	53.96
Delaunay	7.57	-0.09	68.61
Empty tetrahedra	-4.41	0.33	95.13

distinct arrangement of atoms in both toxins. The sum of the differences is -645.65 Å^3, 780.41 Å^3, and -454.26 Å^3 for the method convex hull, Delaunay and empty tetrahedra, respectively, showing that convex hull and empty tetrahedra methods find more void in $hLBT_5$, opposite to Delaunay method which has a positive difference sum.

The empty tetrahedra (ET) method shows little variation between the two toxins for residues with small void (Figure 5.13a), that could explain the small values of the mean and median difference between the two toxins. The correlation coefficient is $r = 0.75$ between the voids of both molecules if we consider chain D, and $r = 0.66$ when we consider the whole protein. The probability values p (for the null-hypothesis of voids between amino acids not being correlated) are 9×10^{-20} and 7×10^{-68}, respectively. The dispersion between the two toxins is incrementing for large void values (Figure 5.13a). This can be also seen by the difference between the maximum values of void between the two toxins (Figure). Void between the toxins is balanced between 10–90 percentile ranks, but $hLTB_5$ has much greater void values in the last 10 percentile ranks, showing larger big local-voids for its positions (Figure 5.13a).

5.6.1 Large voids in $hLTB_5$

Positions 59, 62, and 63 show very large voids in $hLTB_5$ but not so in $CtxB_5$ (Table 5.2). The positions are all on the "hole" of the donut-like shape of the toxins (Figure 5.14), where a lot of empty space is found. The effect of a much larger void for the subset of positions belonging to $hLTB_5$ could be an agglomeration of small voids forming a huge cavity near the hole of the toxins. We recall that void in the empty tetrahedra (ET) method is calculated from the sum of the void of atoms in what we call an *empty tetrahedron*, which has both of its edges shorter than the sum of the atomic endpoints' radii.

We focus on the case of residue lysine on chain D at position 62, having

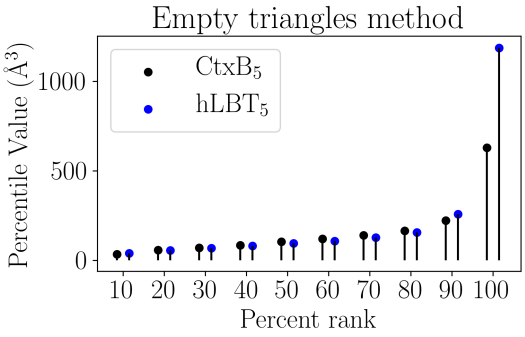
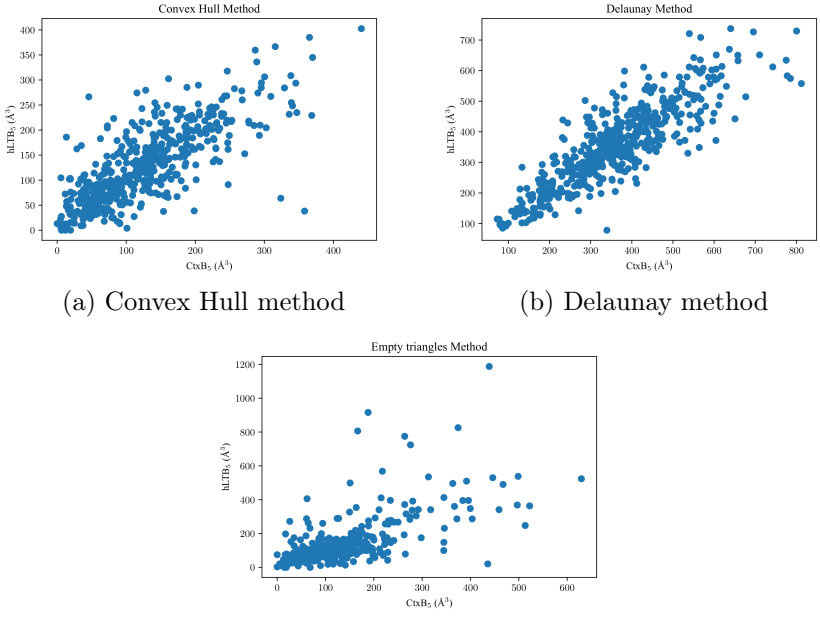


Figure 5.12: The percentile values of the void of $CtxB_5$ and $hLTB_5$ for empty tetrahedra method across all 515 positions.



(a) Convex Hull method

(b) Delaunay method

(c) Empty tetrahedra method

Figure 5.13: $CtxB_5$ vs. $hLTB_5$ void for each residue of the 515 residues in the two toxins.

Table 5.2: Voids obtained with CT method for some residues at positions 59, 62, and 63 are shown for $hLTB_5$ and $CtxB_5$.

Residue	Void (Å^3)	
	$CtxB_5$	$hLTB_5$
D62	1187.47	438.68
F63	915.8	188.12
H59	825.65	374.3
E59	806.14	166.57
E63	774.48	263.82
H63	724.01	275.83
E62	568.34	217.63
G62	534.57	313

9 atoms. In the case of $hLTB_5$ this residue has the largest local void 1187.47 Å^3, and 438.68 Å^3 for the $CtxB_5$ also being a large void (9th largest void). Even if D62 has a large void in both toxins, the void difference of 748.78 Å^3 is extremely large. Analyzing the Delaunay tessellation of both toxins targeting atoms in residue D62; we found that the number of tetrahedra adjacent to an atom of D67 is 160 and 166 for $CtxB_5$ and $hLTB_5$, respectively. So the number of the tetrahedra cannot explain the void difference as they are very similar, but the sum of the volumes can: 895.70 Å^3 for $CtxB_5$ vs. 4111.87 Å^3 for $hLTB_5$. Now, if we only select empty tetrahedra adjacent to the residue, that is, tetrahedra such that all edges are larger than the endpoints' radii, there are 21 and 26 for $CtxB_5$ and $hLTB_5$, respectively. The sum of those tetrahedra for each toxin accounts for 1181.47 Å^3, so the total void in $hLTB_5$ and 479.3 Å^3 in $CtxB_5$.

5.6.2 Delaunay Method cutoff

The cutoff used for the Delaunay method (Section 5.4) selects the tetrahedra to take into account for the void around residues. The cutoff has been set to 5 Ångströms for the results of this work after analysis of the length of edges in the Delaunay tessellation $\mathcal{D}$ of 252 proteins. A tetrahedron of $\mathcal{D}$, has the property of containing no other atomic coordinate of the protein structure inside, but this does not guarantees the avoidance of aberrant cases where tetrahedra has very large edges mostly between atomic coordinates on the surface or at two extremes of a central hole in the case of oligomeric proteins. In order to have a clearer view of the length of edges in the Delaunay tessel-

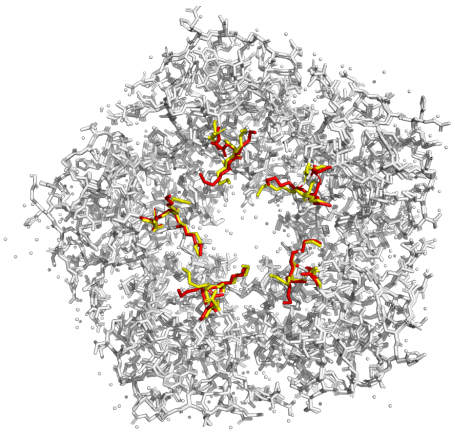


Figure 5.14: Positions 59, 62, and 63 in $CtxB_5$ and $hLTB_5$, the two toxins are aligned and colors yellow and red represent the atoms of $CtxB_5$ and $hLTB_5$, respectively.

lation, to select a suitable cutoff for the removal of bulk space, we measured the length of each edge on the Delaunay tessellation of 252 proteins, for a total of 72,108,048 tetrahedron edges. The mean value of the set of edge lengths is 3.87 Å, with a median value of 3.43 Åand standard deviation 3.49 Å. We can infer from these values that edge lengths vary little around mean but some come close to zero (min is 0.23 Å) and the maximum is 367.14 Å, which amounts for more than the diameter size of some proteins. Very large edge lengths, can produce an extremely large void in proteins when using the Delaunay triangulation (Subsection 5.6.1).

In order to make the data more manageable, we select a random sample of 100,000 edge lengths (Figure 5.15). The data shows a steep change in the distribution of edge length at around 5 Ångströms, the number of values smaller than 5 is 83,906, equal to 84% of all edge lengths for the sample. The remaining lengths are accountable for very large values of void when no cutoff is used, like in the case of the empty tetrahedra method (Section 5.5); even if the void is trapped inside the protein it represents bulk empty space.

The distribution of the length of the edges shows first a flat behavior where edge lengths vary very slowly and suddenly explode as seen in Figure 5.15. The first part of the distribution has a very stable evolution of the lengths, varying slowly for the first 80% of the edges, between 0–5 Å. The other 20% of the edges have values between 5–292 Å. It is around 5 Å that the distribution changes as seen in Figure 5.16. It is relevant to underline

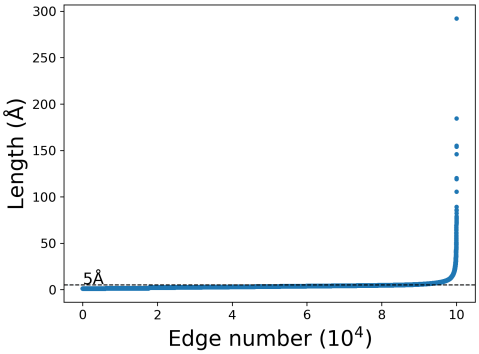
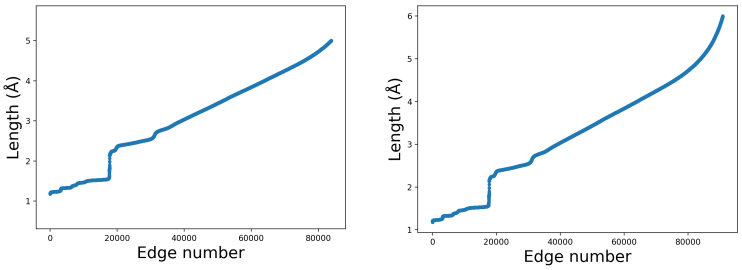


Figure 5.15: Sample of 100,000 edge lengths of tetrahedra of the Delaunay tessellation of 252 proteins.



(a) (b)

Figure 5.16: Distribution of length edges smaller than 5 Å in (a) and 6 Å in (b).

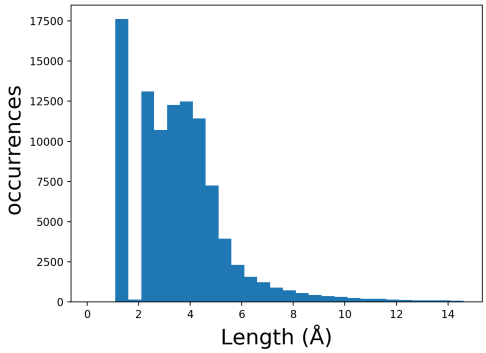


Figure 5.17: Histogram of the length of 100,000 Delaunay edges taking values from 1.2–15 Ångströms. The bins have a size of 0.5 Å and start at 0.1 Å.

the importance of this cutoff as it coincides with the chemical distance under which most interactions happen in the protein structure [21], and is thus the cutoff of 5 Å that was used to compute the void using the Delaunay method in the results presented here.

5.6.3 Gap in atomic distances

By considering the random sample of 100,000 edges, we found a gap in the lengths between 1.6 Å and 2.1 Å, the number of edges of length smaller than 2.1 Åis 17762 so 17.762% of the total, and smaller than 1.6 Å is 17613 so 17.613% of the total, leaving only around 0.1% of the edges in the interval 1.6–2.1 Å. The gap of values is more evident when considering the distribution of lengths on intervals of 0.5 Å. The 1.6–2.1 Å interval is considerable when we take into account that the minimum value of the sample is 1.2 Å, so the interval 1.2–1.6 Å has 17613 edge lengths, compared to 149 in the interval 1.6–2.1 Å (Figure 5.17). In the next interval of the same size, 2.1–2.6 Å there are 13119 edges, so the gap at 1.6–2.1 Å is surprising but congruent with a difference in the distance length between covalent and non-covalent interactions.

Figure 5.18 show the gap forming at 1.6–2.1 Å, relative to the length of edges in the Delaunay tessellation of proteins. Each figure shows a different scale of the same plot zooming in at the interval 1.6–2.1 Å. When considering the total number of edges (72,108,048), we find a very similar pattern from the random sample: 17.8% of edges have a length in the interval 1.1–1.6 Å, only 0.1% in the interval 1.6–2.1 Å, and 13% in 2.1–2.6 Å. As previously

mentioned, this suggest a clear difference between covalent and non-covalent atomic interactions, where the 17% of the distances are actually covalent interactions within the 3D structure, and non-covalent interactions start in the interval 2.1–2.6 Å.

5.6.4 Distribution of void

For each measure, we computed the void of residues belonging to 252 proteins, so the local void of 230,522 residues. The distributions vary to a great extent depending on the method used (Figure 5.19). The Delaunay method shows a distribution of void with a Gaussian-like shape, while the Convex hull method features a half-normal distribution where the void has a median of 107.9 Å^3. The empty tetrahedra method shows an exponential distribution with a median of 106 Å^3.

The Delaunay method shows a normal distribution across a broad range going from 0 to 1282 Å^3, with a mean value of 351.56 Å^3 and standard deviation 153.79 Å^3. The cutoff chosen to calculate the void values is 5 Å, meaning that only tetrahedra with edge lengths smaller than 5 Å are considered to be part of the local void of amino acids (Section 5.4). This value is close to the change in slope of the distribution of 100,00 edge lengths belonging to tetrahedra in the Delaunay tessellation of 252 proteins (Figure 5.15). Restraining the lengths of edges of tetrahedra with a 5 Å cutoff, ignores values after the steep change in the distribution of edge lengths, suggesting that external empty space is excluded from the computation. Indeed, the tetrahedra connecting atoms at very long distances within the protein should account for the external or bulk space of the protein, and the tetrahedra within 5 Å accounts for the void in the protein.

Void trapped inside the protein is measure by the empty tetrahedra method. The mean void captured by this method is smaller than the one of the Delaunay method (229 Å^3), but captures extremely large void (maximum void being 91233 Å^3). This suggests that many tetrahedra accountable for external void are taken into account.

Finally, the convex hull method shows a smaller range of voids compared to the Delaunay method (Figure 5.19). With a median of 107.9 Å^3, it has a median almost equal to the one of the empty tetrahedra method (106 Å^3). The maximum value of void captured by the convex hull method is 535 Å^3, consequence of the geometrical constraint of building void depending on the positions of atoms relative to each one of the faces of the residues (Section 5.3).

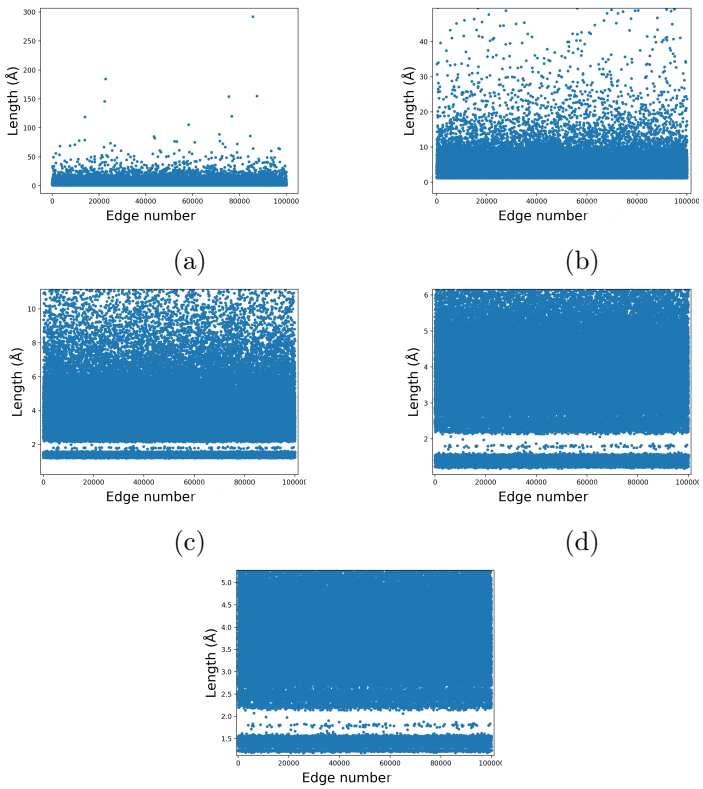


Figure 5.18: Scatter plot of the length of a random sample of 100,000 edges of tetrahedra in the Delaunay Tessellation belonging to 252 proteins. Subfigure (a) to (e) show the same scatter plot at different scales. Subfigure (b) shows a zoomed in version of Subfigure (a), and Subfigure (e) shows the closest zoom of the plot, where a gap for the interval 1.6–2.1 Å can be seen more clearly.

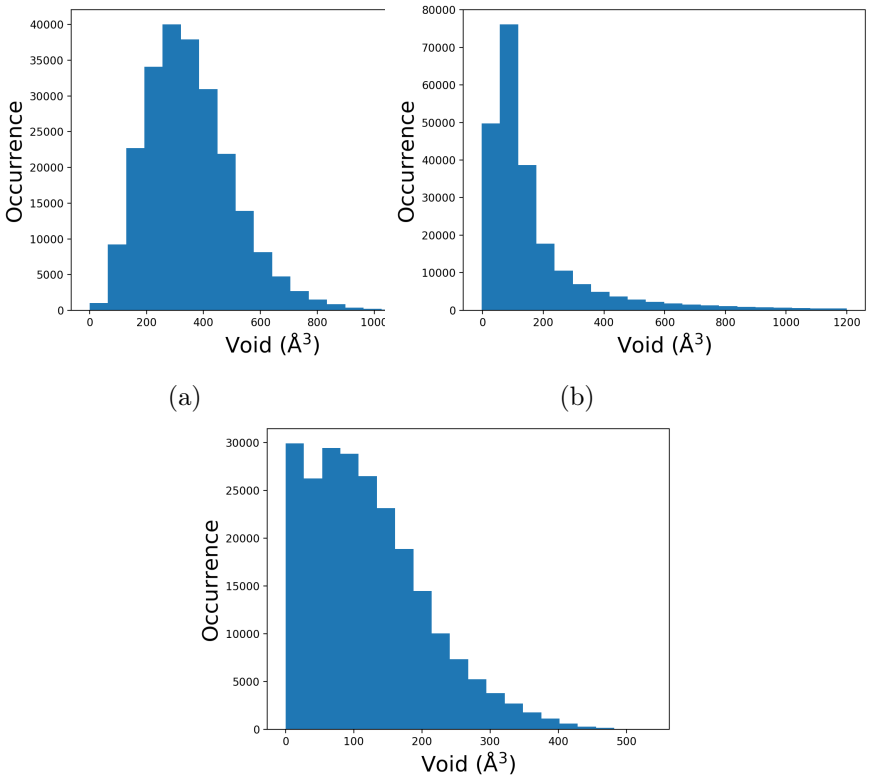


Figure 5.19: Distribution of void across 230,522 residues in 252 proteins. (a) Delaunay Method. (b) Empty tetrahedra Method. (c) Convex hull Method.

5.6.5 Void and Accessible Surface Area

The Delaunay method defines the void of an amino acid as the sum of the volumes of tetrahedra appended to one, two or three atoms of the amino acid. Tetrahedra that have their four vertices (atoms) on a single amino acid are not considered for the void, as they measure space within the same amino acid. Moreover, we filter the set of tetrahedra by removing tetrahedra with one edge larger than five Ångströms. This is done to avoid aberrant cases happening mainly between atoms on the surface of the protein.

We calculated the void of residues belonging to toxins $CtxB_5$ and $hLTB_5$ each one having 515 sequence positions from which 95 differ, and are considered as mutated positions from one another.

The correlation between the void of the 515 residues in both proteins is $r = 0.86$. The void is indeed very similar due to the very similar structures between the two proteins. The voids related to the mutated positions are less correlated ($r = 0.72$) between the two proteins. Mutated positions have on average smaller voids relative to the total number of positions: that means voids of 324 $Å^3$ and 325 $Å^3$ for mutated positions in the cholera toxin and head labile, respectively; and 367 $Å^3$ and 362 $Å^3$ in both toxins, for non-mutated positions.

In the case of accessible surface area (ASA), the void is larger for buried positions. This is due to the fact that tetrahedra having larger edges are usually in the surface. We computed the void and accessible surface area of the residues of a database of 250 proteins. The distribution of the void follows a normal law (Figure 5.20).

In order to study the relation between the distribution of void and accessible surface area, we separated the values of void depending on whether the residue had access to the surface or not (ASA$>$ 0 or ASA$=$ 0, respectively). We noticed that buried positions tend to have larger voids than surface exposed positions (Figure 5.21).

5.6.6 Conclusion

Proteins are robust objects that need to adapt under certain circumstances to improve fitness. The equilibrium between robustness and adaptability/fragility is reflected by the protein structure. Void is on average the same across positions in the protein, showing a “Goldilocks” effect of the empty space distribution, where residues have averaged values of void that are neither too small nor too big. The void of a residue is dependent on the position in the 3D structure of the residue, but up to available space inside

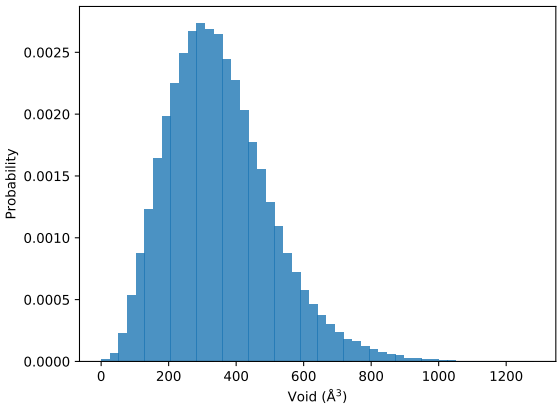


Figure 5.20: The distribution of void over 250 proteins and 230522 residues.

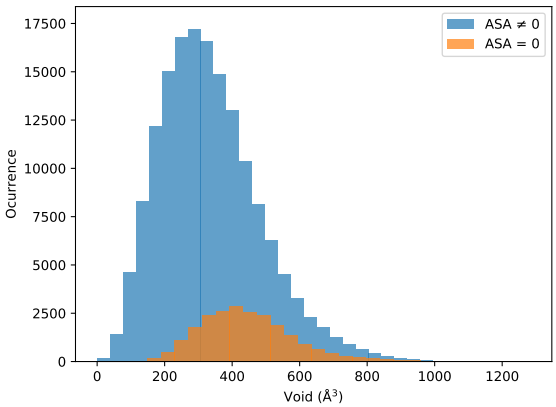


Figure 5.21: Distribution of void over buried and surface positions

the protein. That is, buried residues have more space, and therefore empty space around them than surface positions.

The distribution of degree, weight and empty space across residues being similar, suggests that the difference between positions is minimal. A different value of positions is relative to the aggregated changes on the neighbors of the position. The exceptions to the Goldilocks effect of the empty space and more generally, of the local structures of residues, are found on positions having special neighborhoods, where residues have average structures themselves but together conform a distinctive atomic arrangement.

The methods used to calculate the void that depend on the Delaunay tessellation of their atomic structures can contain aberrant tetrahedra, usually on the surface or in "holes" for oligomeric proteins. We found that edge lengths of tetrahedra varies little around its mean, which is smaller than the chemical distance of atomic interaction. However, some edge lengths can have values exceeding the total length of complete proteins in some cases, suggesting that very large voids that are computed without a cutoff for edge length measure the bulk space trapped in the protein. Methods not dependent on cutoffs eliminating aberrant cases like those described before can be used instead, but at the expense of more algorithmic complexity. However, we show that the vast majority of edge lengths falls near the chemical distance of 5 Ångströms, which suggest a fair computation of void using the chemical cutoff. Indeed, the two of our methods using cutoffs showed a strong correlation of void values for residues belonging to two structurally similar toxins, suggesting a fair estimation of local void of residues.

Further study on the use of a cutoff will need to be done by comparing void values from our measures to cutoff-independent methods. This, to shed some light on the trade off between algorithmic complexity and accuracy of the void values. Specifically, a study including a large protein data set could be used to compare voids. On the other hand, our methods allow for a computation of relative void, or local void, of amino acids, which are essential for our study of local structures.

5.6.7 Supplementary table

Table 5.3: Void values of residues belonging to $CtxB_5$ and $hLTB_5$ for chain D, using the three methods. The difference *diff* between the values of each residue in two molecules is shown, and the residue name follows the format *chain + position*.

	Delaunay Method			Convex Hull Method			Empty tetrahedra Method		
Residue	$CtxB_5$	$hLTB_5$	*diff*	$CtxB_5$	$hLTB_5$	*diff*	$CtxB_5$	$hLTB_5$	*diff*
D1	177.41	177.50	-0.09	37.63	21.18	16.45	344.64	99.71	244.93
D2	382.24	598.54	-216.30	142.77	74.41	68.36	67.09	51.91	15.18
D3	319.07	286.84	32.23	47.98	52.61	-4.63	107.83	158.99	-51.16
D4	352.28	241.59	110.70	171.39	112.06	59.32	185.90	187.15	-1.25
D5	450.99	418.52	32.47	93.68	167.15	-73.46	110.72	141.90	-31.18
D6	299.58	213.76	85.83	61.56	42.49	19.07	19.43	29.09	-9.66
D7	335.97	358.35	-22.38	85.45	62.39	23.06	5.20	20.66	-15.46
D8	491.40	475.60	15.81	106.92	89.93	16.99	173.07	136.77	36.31
D9	327.82	311.72	16.10	41.45	41.24	0.21	17.73	14.79	2.95
D10	133.56	166.34	-32.78	6.26	0.00	6.27	13.04	46.23	-33.19
D11	399.69	383.07	16.62	37.67	65.19	-27.52	107.85	173.40	-65.55
D12	611.42	487.70	123.72	186.27	209.70	-23.42	344.93	412.82	-67.89
D13	205.53	212.69	-7.16	11.61	72.34	-60.73	629.39	523.81	105.58
D14	324.19	356.05	-31.86	50.97	41.95	9.03	34.85	127.63	-92.78
D15	395.81	339.60	56.21	58.09	91.79	-33.70	55.15	80.87	-25.71
D16	459.87	373.94	85.94	84.51	93.30	-8.79	34.64	55.92	-21.28
D17	434.57	408.99	25.58	143.41	97.94	45.46	27.53	46.84	-19.31
D18	423.20	488.55	-65.35	153.79	145.57	8.23	105.73	79.61	26.12
D19	213.50	129.11	84.39	51.95	77.95	-26.00	46.54	88.14	-41.61
D20	480.38	433.45	46.94	81.82	223.12	-141.30	133.39	104.08	29.31
D21	270.55	141.87	128.68	34.50	8.92	25.58	34.31	50.98	-16.67
D22	373.29	337.31	35.97	54.20	49.73	4.47	107.16	75.75	31.41
D23	255.93	259.60	-3.67	185.10	128.03	57.07	157.84	34.39	123.45
D24	392.01	412.36	-20.34	101.43	104.77	-3.34	153.41	98.84	54.56
D25	327.90	370.51	-42.62	198.43	38.62	159.81	200.23	161.49	38.75
D26	387.96	335.72	52.23	110.87	92.37	18.50	91.52	199.45	-107.93
D27	637.13	669.71	-32.58	246.27	317.68	-71.41	118.16	131.23	-13.08
D28	397.50	430.89	-33.39	178.35	145.93	32.42	135.56	151.04	-15.48
D29	472.61	425.59	47.03	128.18	279.45	-151.28	99.70	81.35	18.35
D30	297.10	293.86	3.24	104.51	109.95	-5.44	109.89	120.61	-10.72

	Delaunay Method			Convex Hull Method			Empty tetrahedra Method		
Residue	$CtxB_5$	$hLTB_5$	*diff*	$CtxB_5$	$hLTB_5$	*diff*	$CtxB_5$	$hLTB_5$	*diff*
D31	442.01	338.29	103.72	204.28	252.03	-47.76	65.53	95.76	-30.23
D32	351.92	296.78	55.13	40.10	41.25	-1.15	38.21	44.11	-5.90
D33	199.88	182.22	17.65	11.71	11.96	-0.24	65.44	64.77	0.67
D34	176.84	170.43	6.41	133.40	76.07	57.33	399.59	347.90	51.69
D35	477.72	500.92	-23.20	302.71	204.56	98.15	188.89	245.04	-56.14
D36	596.59	435.10	161.49	159.19	249.37	-90.17	65.64	80.70	-15.06
D37	409.45	397.47	11.99	214.64	201.27	13.38	150.56	157.55	-6.99
D38	308.79	412.35	-103.56	62.61	182.85	-120.24	55.45	93.04	-37.59
D39	361.35	470.01	-108.66	244.78	213.92	30.87	147.49	143.58	3.91
D40	401.40	429.07	-27.68	194.51	166.78	27.72	105.60	108.17	-2.57
D41	382.30	380.42	1.88	69.26	57.52	11.73	107.29	63.32	43.97
D42	777.98	582.98	195.00	148.38	194.65	-46.27	95.61	95.27	0.34
D43	117.33	127.25	-9.92	42.61	74.85	-32.23	191.50	37.84	153.66
D44	185.42	154.21	31.22	24.77	25.12	-0.35	102.28	52.15	50.13
D45	80.28	94.78	-14.50	8.54	4.85	3.69	115.84	18.82	97.02
D46	298.83	332.35	-33.52	18.82	103.59	-84.77	53.04	124.19	-71.14
D47	348.17	392.93	-44.77	112.15	92.53	19.62	96.29	83.79	12.51
D48	775.32	633.93	141.39	238.10	233.08	5.02	104.69	101.14	3.55
D49	573.49	607.83	-34.34	165.39	199.85	-34.46	86.24	56.31	29.92
D50	243.71	302.80	-59.09	151.05	160.49	-9.44	81.66	61.58	20.08
D51	364.16	443.59	-79.44	198.88	221.56	-22.68	171.67	186.55	-14.88
D52	339.99	330.60	9.40	80.84	93.13	-12.29	116.87	139.31	-22.44
D53	144.68	215.24	-70.56	135.30	106.01	29.29	287.70	304.54	-16.84
D54	179.93	158.84	21.09	5.80	5.75	0.06	0.00	74.99	-75.00
D55	72.65	114.77	-42.12	19.05	46.70	-27.65	371.86	286.58	85.27
D56	339.17	343.93	-4.76	109.24	69.59	39.65	446.02	530.00	-83.98
D57	521.27	547.93	-26.67	154.29	103.39	50.90	181.91	153.86	28.05
D58	375.00	274.29	100.71	38.37	75.61	-37.24	230.03	255.71	-25.68
D59	115.59	122.46	-6.87	24.00	73.23	-49.23	215.04	411.05	-196.01
D60	206.93	226.02	-19.08	73.30	59.87	13.43	152.28	131.68	20.61
D61	607.58	573.28	34.30	170.83	145.85	24.98	95.64	73.27	22.37
D62	359.78	204.80	154.98	98.94	145.65	-46.71	438.68	1187.47	-748.79
D63	198.00	231.63	-33.63	189.87	196.51	-6.64	149.76	327.07	-177.31
D64	252.96	281.70	-28.74	42.52	89.34	-46.83	92.38	99.85	-7.47
D65	439.86	535.05	-95.19	150.31	215.24	-64.93	69.40	82.16	-12.76
D66	395.50	456.44	-60.94	336.02	231.46	104.55	92.79	96.10	-3.31
D67	506.75	507.28	-0.53	300.34	306.21	-5.86	135.58	118.43	17.15

	Delaunay Method			Convex Hull Method			Empty tetrahedra Method		
Residue	$CtxB_5$	$hLTB_5$	*diff*	$CtxB_5$	$hLTB_5$	*diff*	$CtxB_5$	$hLTB_5$	*diff*
D68	557.29	601.76	-44.47	241.16	232.33	8.82	66.63	57.33	9.30
D69	446.22	527.59	-81.37	174.41	242.26	-67.85	140.65	137.37	3.28
D70	432.03	348.28	83.75	135.30	147.90	-12.60	44.54	30.01	14.53
D71	398.23	384.37	13.86	77.30	114.28	-36.98	81.69	67.58	14.10
D72	443.06	445.44	-2.38	195.02	229.17	-34.15	144.76	149.90	-5.14
D73	492.66	482.72	9.94	369.87	344.93	24.95	111.93	137.37	-25.43
D74	255.52	271.95	-16.43	136.30	151.94	-15.64	71.62	57.35	14.28
D75	212.76	321.25	-108.49	40.01	110.69	-70.68	121.02	123.23	-2.21
D76	554.02	607.20	-53.18	365.33	385.02	-19.69	196.95	112.27	84.69
D77	269.89	283.16	-13.28	231.09	214.85	16.24	146.51	97.11	49.40
D78	242.99	231.81	11.18	40.34	49.99	-9.65	68.19	0.00	68.20
D79	250.32	298.26	-47.95	154.66	153.84	0.82	202.11	57.90	144.21
D80	181.13	267.82	-86.69	33.02	34.07	-1.05	131.77	84.95	46.83
D81	329.94	246.41	83.53	115.37	165.82	-50.45	148.06	83.10	64.96
D82	435.14	476.84	-41.70	99.30	185.87	-86.57	110.84	131.54	-20.70
D83	398.19	361.48	36.70	164.95	101.62	63.33	152.97	201.91	-48.94
D84	568.25	506.40	61.85	247.63	91.06	156.57	48.59	100.01	-51.42
D85	429.09	435.13	-6.04	178.15	238.82	-60.67	99.78	98.55	1.23
D86	334.45	347.93	-13.47	123.75	79.16	44.59	105.47	71.66	33.81
D87	599.78	460.57	139.21	124.69	142.88	-18.19	65.27	50.95	14.32
D88	579.44	483.17	96.28	195.24	222.91	-27.67	291.87	341.47	-49.60
D89	434.68	282.80	151.87	73.42	70.77	2.65	5.42	10.02	-4.59
D90	217.17	300.53	-83.36	86.12	73.92	12.20	144.76	55.08	89.69
D91	317.22	298.48	18.74	162.47	115.92	46.55	157.34	118.27	39.07
D92	134.49	147.93	-13.43	57.94	67.31	-9.37	39.38	46.90	-7.52
D93	437.19	483.70	-46.50	19.81	30.76	-10.96	46.91	38.14	8.77
D94	529.13	371.95	157.19	134.95	109.52	25.43	66.01	105.81	-39.81
D95	450.49	431.81	18.68	93.36	86.16	7.19	19.80	26.83	-7.03
D96	389.32	332.86	56.47	226.49	183.15	43.34	149.81	104.97	44.84
D97	321.85	311.69	10.16	13.80	57.24	-43.44	107.37	95.00	12.37
D98	279.02	269.86	9.17	50.85	51.71	-0.86	96.86	62.81	34.05
D99	307.76	377.08	-69.32	187.52	182.15	5.36	176.72	108.90	67.82
D100	345.56	318.74	26.82	83.77	91.66	-7.89	32.75	82.49	-49.73
D101	367.01	389.59	-22.58	154.10	104.67	49.43	159.53	156.41	3.12
D102	176.99	173.25	3.74	45.82	266.37	-220.56	34.87	175.18	-140.32
D103	175.69	243.47	-67.78	51.56	82.30	-30.74	119.74	102.22	17.52

Chapter 6

Overview

The work produced in this dissertation, is at the frontier between data science, mathematics and bioinformatics. It was based on the modeling of the protein structure as a network, where nodes are atoms or amino acids that are connected through a link if they are close enough in the structure. Proteins are molecules whose function is mainly characterized by their structure. Having an optimal structure is thus a necessary condition for the well functioning of a protein. A variation in the structure can yield negative or positive consequences consequences to the function. In nature, proteins are molecules that vary through mutations in their genetic sequence, sometimes causing deleterious effects to the function of a protein. The analysis of the effects of mutations on the structure is therefore the first step to understanding those effects on the protein function.

Proteins are robust biological objects that deal with a paradigm: they should be robust to tolerate mutations and adaptable to new functions in order to evolve. In reality, mutations seldom have an impact in the protein function, as proteins are indeed robust. Some mutations, however, can be deleterious of the wild type function, and in some cases give conditions for adaptability of a new function. The structure of proteins is consequential to the effect mutations have on the function, because the sequence of the protein, where mutations take place, contains the information of the final function and structure of the protein. Indeed, the structure can be thought as the tool for the protein to conduct its function. Here, we are interested in the study of protein structure *per se*, and we studied it based on its atomic coordinates. First, by developing a framework for studying structural change (when mutations happen). Second, using the same framework, we compared the effects of mutations on structure and function basing on a dataset of functional change produced experimentally [42]. We also compared the structural effects to the level of buriedness of the mutated position, to look at the relation between local and global structural change. At the same time, we conducted a survey on local structures of amino acids for a large protein database, studying the connectivity of residues under a network approach. Third, we implemented three methods to study the void or empty space around amino acids, this in order to quantify individual void "trapped" inside the protein and as a measure of adjacent discrete geometrical objects to the residue in question. Concerning these three parts, in the following we describe our main conclusions and possible extensions of our work.

6.1 Framework

The framework for studying structural change is based on a network approach, where amino acids at the interface of two chains are called hot spots and compose the nodes of the network. The connectivity between nodes is based on spatial proximity, where two nodes are connected if they are at distance less than five Ångströms from each other. Assuming that all proteins bear mutations by similar mechanisms, a case of study is a good model of investigation. We constructed the hot spot network of the Cholera toxin B pentamer (PDB file 1EEI), obtaining the hotspot amino acids. We mutated *in silico* each hotspot amino acid by asparagine and constructed the network of the resulting PDB file. One novel finding is that structural changes follow a cascade mechanism where the local reorganization of the atoms at the site of the mutation disturbs the chemical neighbors of the mutated hotspot which in turn disturb their chemical neighbors, as in a domino effect. We found that motifs in the secondary structure of the protein (α-helix and β-sheet) were found to have a large perturbation propagation in general. As a matter of fact, a large part of the mutations propagated in the protein structure beyond the chemical distance of the mutated position.

We clustered nodes into communities of secondary structural motifs to qualify the topological variations produced by mutations. We encountered two different scenarios: mutations perturbing connections between different communities of structural motifs, and those whose perturbations happened within the communities. This led to the hypothesis that a mutation could yield an alternative network, which although topologically different from the original, would be considered structurally equivalent, which fits the updated definition of protein function as an ensemble of conformations [45].

At the end of the first part of the work, we tested the alternative structure hypothesis on a cancer related mutation which disconnected different communities of nodes in the original network. The idea was to find another mutation which could neutralize these negative effects on the structure, by having an *ad hoc* alternative network. We found a mutation that brought close enough these communities to act as a correction mechanism when the protein was simultaneously mutated by the cancerous mutation. Finally, we understood the need to consider the complete set of amino acids in the network, given that structural change happening at a distance cutoff could be consequence of structural change happening exclusively at non-hotspot positions.

6.2 Local structure

6.2.1 Local structure of amino acids

Each node in the amino acid network (AAN) represents an amino acid in the protein. Nodes are labeled by amino acid position in the sequence. Labels of a same position in two amino acid networks (e.g. corresponding to wild type and mutated proteins), are identical. An edge connects two nodes in the network if its incident amino acids are interacting in the structure. An interaction between two amino acids occurs when two atoms, one in each amino acid, are in turn interacting. The interaction distance of atoms is defined by a distance cutoff. This cutoff, is usually equal to five Ångströms in our work, representing chemical interaction distance between atoms. The use of multiple cutoffs in Chapter 4 however, permitted us to investigate the structural surroundings of amino acids within and beyond chemical distance. In the network, each edge has a weight (a real number) corresponding to the number of atomic interactions shared between the two incident amino acids of the edge.

We used the ANN to set the local structure of each amino acid, which is the subset of atoms of the protein structure around the amino acid, under a given distance cutoff, belonging to other amino acids. The degree of a node represents its number of amino acid interactions. Weight and degree are properties that depend on the local structure of amino acids. The weight of a node in the amino acid network (the sum of the weights of edges incident to that node), is the number of atomic interactions of the node.

The local structure of an amino acid depends on its surrounding, but does it depend on the amino acid itself too? We addressed this question with the statistical analysis of the local structures of residues in a database of 252 proteins.

A variety of degrees and weights can be found across most amino acid types. Some buried positions are not packed to their full potential, showing evidence of empty space or void around positions. Altogether, packing is correlated to accessible surface area, but positions at the core of the protein can have the same level of packing as positions at the surface.

The average weight of an edge of nodes, the ratio weight to degree, is called the neighborhood watch, and is almost constant across amino acid types. This implies a restriction for the packing of amino acids that should be further investigated in future work. We concluded that amino acid types had large range of interchangeability in the protein structure and the pairwise geometrical, chemical and atomic interaction properties of a position appear

a new reasonable tool to investigate the folding mechanisms of a protein.

Local structure not depending on amino acid type, and only partially correlated to the accessible surface area, points towards the hypothesis that protein robustness is a consequence of the interchangeability of local structures across amino acids. The movement of atoms in the 3D protein structure as a consequence of a protein mutation, is not relevant on its own to study of functional change (due to mutations) ending in fragility (loss of protein function), or adaptability (change of protein function). New qualitative measures must be implemented for the analysis of structural change.

6.2.2 Local structure of functional positions

Experimental and evolutionary studies have pointed out independently to the evidence that some positions in the sequence drive the functional change of proteins. For the case study of the PSD-95 protein, it was shown experimentally that the protein contains a set of twenty positions that when mutated, the protein loses its native function substantially more often than when mutations occur elsewhere. Moreover, the effects are independent from what amino acid is replacing which. Functionally sensitive positions, as we call them, tend to make the protein lose its native function when mutated by any other amino acid.

The question we raise is to know whether these positions are structurally distinguishable from others, in terms of local structure, or, is the structural surrounding or local structure of functionally sensitive positions distinct from that of the rest of positions? If so, it starts differing at what distance?

We showed that neither functionally sensitive nor buried positions are distinguishable structurally from the rest when the local structure is defined on a very short distance (less than four Ångströms). However, some differences appear between types of position (functionally sensitive/non-functionally sensitive, buried/surface positions) when using a cutoff larger than four Ångströms.

We used three measures of structural change based on our previous work on hotspot networks. The first counts the number of atomic interactions perturbed by the mutation. The second is the number of perturbed amino acids, that is, amino acids that have at least one perturbed atom. And the third is the largest distance in the 3D structure between the position of the mutation and a perturbed amino acid.

When atomic interaction, is set to be four Ångströms or less, no difference is made between positions, for any measure. The use of larger cutoffs than chemical distance, aims at investigating the structural relations between two

parts of the protein that are not necessarily chemically linked. This is important to investigate, as it allows for a more complete inventory of structural differences around different positions.

Moreover, mutations at functionally sensitive positions, perturbed more amino acids in general. We found that there is a correlation of up to seventy percent between functional change and number of perturbed amino acids by the mutation. On the other hand, there is no significant link between functional change and number of perturbed interactions. Amino acid perturbation plays thus a more important role in functional change than interaction perturbation does, i.e. it is not the number of interactions the mutation can perturb that counts in terms of functional change, but the number of amino acids. Indeed, whether one atomic pair between two amino acids, or all atomic pairs are perturbed by the mutation is not relevant to the functional change.

Indeed, we found that the number of perturbed atomic or amino acid interactions are not characteristic to functionally sensitive positions, as they affect similarly the rest of the functionally robust positions. On the other hand, a larger number of perturbed amino acids was found to be correlated to functionally sensitive positions. We suggested that this result could underlie the existence of several communication paths between amino acids as a method of error-correction. Further work needs to be done to test this hypothesis, however.

The correlation between number of perturbed nodes and functional change, implies a relation between the local structure of functionally sensitive positions and the structural change of the protein when mutated. The change in the atomic configuration of the position consequence of replacing one amino acid by another, can be approached by a high-throughput statistical analysis where a mutation at the position would be simulated by a random change in the atomic configuration of the protein. Indeed, this would allow for an arbitrary number of structural change values relying on only on the local structural surrounding of the position being mutated.

6.3 Local void

The final part of my dissertation was the implementation of three different algorithms to quantify the empty space around groups of atoms. This was done by using the Delaunay tessellation of the set of 3D atomic coordinates of the structure as well as the convex hull of residues' atomic coordinates. In addition, one of the algorithms considers the atoms to be spheres of radii

equal to their van der Waals radii, in order to take into account the atomic sizes.

The methods used to calculate the void that depend on the Delaunay tessellation of their atomic structures can contain aberrant tetrahedra, usually on the surface, or in "holes" for the case of oligomeric proteins. We found that edge lengths of tetrahedra varies little around their mean, which is smaller than the chemical distance of atomic interaction. However, edge lengths can have values exceeding the total length of complete proteins in some cases, suggesting that very large voids that are computed without a cutoff for edge length, measure the bulk space trapped in the protein. Methods not dependent on cutoffs, but eliminating cases like these can be used instead, at the expense of more algorithmic complexity. Furthermore, we showed that the vast majority of edge lengths fall near the chemical distance of 5 Ångströms, which suggest a fair computation of the void within the protein is done using the chemical cutoff. Indeed, two of our methods using cutoffs, showed a strong correlation of void values for residues belonging to two structurally similar toxins, suggesting a fair estimation of local void around residues.

Void using the Delaunay tessellation and a cutoff, is distributed normally across positions in the protein, showing a "Goldilocks" effect of the empty space around residues, where residues have averaged values of void that are neither too small nor too big from the mean value. The void of a residue is dependent on the position in the 3D structure of the residue, but up to available space inside the protein, that is, buried residues have more space, and therefore more empty pace around them compared to surface positions.

6.4 Future work

In the future, we propose to approximate the local structure of amino acids by a sphere centered at the centroid of each amino acid, and radius equal to five Ångströms, to measure the distribution of the atoms of the local structure by partitioning the sphere in several subsets of 3D points, and counting the number of atoms per subset. This would indicate the spatial distribution of local structures, perhaps differentiating between two local structures otherwise similar in terms of weight and degree, neighborhood watch or buried and surface position.

Further study on tessellation methods using a cutoff, will need to be done by comparing void values from our measures with cutoff-independent methods, to shed some light on the trade off between algorithmic complexity and accuracy of the void values. Specifically, a study including a large protein

data set could be used to compare voids. Finally, our methods allow for a computation of relative void, or local void of amino acids, favorable for the study of local structures.

Chapter 7

Introduction (Français)

Le travail produit dans cette thèse se trouve à la frontière entre la science des données, les mathématiques et la bio-informatique. Il a était basé sur la modélisation de la structure des protéines comme des réseaux, où les noeuds sont des atomes ou des acides aminés qui sont reliés par un lien s'ils sont assez proches dans la structure. Les protéines sont des molécules dont la fonction est principalement caractérisée par leur structure. Avoir une structure optimale est donc une condition nécessaire au bon fonctionnement d'une protéine. Une variation dans la structure peut avoir des conséquences négatives ou positives sur la fonction. Dans la nature, les protéines sont des molécules qui varient à travers des mutations dans leur séquence génétique, causant parfois des effets délétères à la fonction d'une protéine. L'analyse des effets des mutations sur la structure est donc la première étape pour comprendre ces effets sur la fonction des protéines.

Les protéines sont des objets biologiques robustes qui répondent à un paradigme: elles doivent être robustes pour tolérer des mutations et adaptables à de nouvelles fonctions pour évoluer. En réalité, les mutations ont rarement un impact sur la fonction des protéines car les protéines sont en effet robustes. Certaines mutations peuvent cependant être délétères à la fonction de la protéine dite *wild type*, et dans certains cas, fournir des conditions pour l'adaptabilité d'une nouvelle fonction. La structure des protéines est la cause de l'effet des mutations sur la fonction, parce que la séquence de la protéine, où les mutations ont lieu, contient l'information de la fonction finale et structure de la protéine. En effet, la structure peut être considérée comme l'outil de la protéine pour mener à bien sa fonction. Ici, nous sommes intéressés par l'étude de la structure des protéines *en soi*, et nous l'avons étudié en fonction de ses coordonnées atomiques. D'abord, en développant un cadre pour l'étude du changement structurel (quand des mutations se produisent). Deuxièmement, en utilisant le même cadre, nous avons comparé les effets des mutations sur la structure et la fonction en fonction d'un ensemble de changements produits expérimentalement [42]. Nous avons également comparé les effets structurels au niveau de l'enterrement de la position mutée, examiné la relation entre le changement structurel local et globale. Parallèlement, nous avons mené une enquête sur les structures locales d'acides aminés sur une large base de données de protéines, en étudiant la connectivité des résidus dans le cadre d'une approche en réseau. Troisièmement, nous avons mis en oeuvre trois méthodes pour étudier le vide ou l'espace vide autour des acides aminés, afin de quantifier vide individuel 'piégé' à l'intérieur de la protéine et comme mesure de la géométrie discrète d'objets adjacents au résidu en question. En ce qui concerne ces trois parties, nous décrivons ci-dessous nos principales conclusions et les extensions possibles de notre travail.

7.1 Le cadre

Le cadre d'étude des changements structurels repose sur une approche en réseau, où les acides aminés l'interface de deux chaînes sont appelées hot spots et composent les noeuds du réseau. La connectivité entre les noeuds est basée sur la proximité spatiale, où deux noeuds sont connectés s'ils sont à distance moins de cinq Ångströms l'un de l'autre. En supposant que toutes les protéines portent des mutations par mécanismes similaires, un cas d'étude est un bon modèle d'enquête. Nous avons construit le réseau hotspot du pentamère B de la toxine du choléra (fichier PDB 1EEI), obtenant les acides aminés hotspot. Nous avons muté *in silico* chaque acide aminé hotspot par l'asparagine et construit le réseau résultant du fichier PDB. Une découverte nouvelle est que les changements structurels suivent un mécanisme en cascade où la réorganisation locale des atomes sur le site de la mutation perturbe les voisins chimiques du hotspot muté qui, à son tour, perturbe leurs voisins chimiques, comme dans un effet domino. Nous avons trouvé que des motifs dans la structure secondaire de la protéine (α-helix et β-sheet) ont été trouvés de faire une grande propagation de la perturbation en général. En fait, une grande partie des mutations propage dans la structure de la protéine, au-delà de la distance chimique de la position mutée.

Nous avons regroupé les noeuds en communautés de motifs structuraux secondaires pour qualifier les variations topologiques produites par des mutations. Nous avons rencontré deux scénarios différents: des mutations perturbant les connexions entre différentes communautés de motifs structurels, et ceux dont les perturbations se sont produites dans le communautés. Cela a conduit à l'hypothèse qu'une mutation pourrait produire un réseau alternatif, qui bien que topologiquement différent de l'original, serait considéré comme structurellement équivalent, ce qui correspond à la définition actuelle de la fonction des protéines comme un ensemble de conformations [45].

A la fin de la première partie du travail, nous avons testé l'hypothèse de structure alternative sur une mutation reliée à un cancer qui déconnectait différentes communautés de noeuds dans le réseau d'origine. L'idée était de trouver une autre mutation qui pourrait neutraliser ces effets négatifs sur la structure, par moyen d'un réseau alternatif *ad hoc*. Nous avons trouvé une mutation qui a assez rapproché ces communautés pour agir comme un mécanisme de correction lorsque la protéine a été simultanément muté par la mutation cancéreuse. Enfin, nous avons compris la nécessité de considérer l'ensemble complet des acides aminés dans le réseau, étant donné que le changement structurel se produisant à une distance cutoff pourrait être la conséquence de changement se produisant exclusivement à des positions non-

hotspot.

7.2 Structure locale

7.2.1 La structure locale d'acides aminés

Chaque noeud dans le réseau d'acides aminés (AAN) représente un acide aminé dans la protéine. Les noeuds sont étiquetés par position d'acide aminé dans la séquence. Les étiquettes d'une même position dans deux réseaux d'acides aminés différents (par exemple correspondant à des protéines de type wild type et mutées), sont identiques. Un lien connecte deux noeuds dans le réseau si ses acides aminés incidents interagissent dans la structure. Une interaction entre deux acides aminés se produit lorsque deux atomes, un dans chaque acide aminé, interagissent à leur tour. La distance d'interaction des atomes est défini par un cutoff de distance. Ce cutoff est généralement égale à cinq Ångströms dans notre travail, représentant la distance d'interaction chimique entre les atomes. Utilisant multiples cutoffs dans le chapitre 4, cependant, nous a permis d'étudier l'environnement structurel des acides aminés dans et au-delà de la distance chimique. Dans le réseau, chaque lien a un poids (un nombre réel) correspondant au nombre d'interactions atomiques partagées entre les deux acides aminés incidents au lien.

Nous avons utilisé le AAN pour définir la structure locale de chaque acide aminé, qui est le sous-ensemble des atomes de la structure appartenant à d'autres acides aminés de la protéine, autour de l'acide aminé, et sous un cutoff de distance donnée. Le degré d'un noeud représente son nombre d'interactions en termes d'acides aminés. Le poids et le degré sont des propriétés qui dépendent de la structure locale des acides aminés. Le poids d'un noeud dans le réseau d'acides aminés (la somme des poids de liens incidents à ce noeud), est le nombre d'interactions atomiques du noeud.

La structure locale d'un acide aminé dépend de son environnement, mais dépend-elle de l'acide aminé lui-même aussi? Nous avons abordé cette question avec l'analyse statistique des structures locales de résidus dans une base de données de 252 protéines.

Une variété de degrés et de poids peuvent être trouvés dans la plupart des types d'acides aminés. Quelques positions enfouies ne sont pas entourées à leur plein potentiel, montrant une preuve d'espace vide autour des positions. Au total, le *packing* est corrélé à l'aire de la surface accessible, mais les positions au coeur de la protéine peuvent avoir le même niveau de *packing* que les positions à la surface.

Le poids moyen d'un lien, le rapport poids à degré, est appelé le *neighborhood watch*, et est presque constant à travers tous les types d'acides aminés. Cela implique une restriction du *packing* des acides aminés qui devraient être étudiés d'avantage dans les travaux futurs. Nous avons conclu que les types d'acides aminés avait une large gamme d'interchangeabilité dans la structure de la protéine et la paire géométrique, propriétés d'interaction chimique et atomique d'une position apparaissent comme un nouvel outil pour étudier les mécanismes du repliement d'une protéine.

La structure locale ne dépend pas du type d'acide aminé, et seulement partiellement corrélée à l'aire de surface accessible, pointe vers l'hypothèse que la robustesse des protéines est une conséquence de interchangeabilité des structures locales à travers les acides aminés. Le mouvement des atomes dans la structure 3D de la protéine en conséquence d'une mutation protéique, n'est pas pertinente en elle-même pour étudier le changement (en raison de mutations) se terminant par la fragilité (perte de la fonction de la protéine), ou l'adaptabilité (changement de la fonction de la protéine). De nouvelles mesures qualitatives doivent être mises en oeuvre pour l'analyse du changement structurel.

7.2.2 Structure locale des positions fonctionnelles

Des études expérimentales et évolutives ont montré indépendamment de la preuve que certaines positions dans la séquence sont clés pour le changement fonctionnel des protéines. Pour le cas d'étude de la protéine PSD-95, il a été démontré expérimentalement que la protéine contient un ensemble de vingt positions que lorsque mutées, la protéine perd sa fonction native beaucoup plus souvent que lorsque des mutations se produisent ailleurs dans la séquence. De plus, les effets sont indépendants du type d'acide aminé mutant. Les positions fonctionnellement sensibles, comme nous les appelons ici, tendent à faire perdre à la protéine sa fonction native muté par tout autre acide aminé.

La question que nous soulevons est de savoir si ces positions se distinguent structurellement d'autres, en termes de structure locale, ou, est la structure locale de des positions fonctionnellement sensibles distinctes de celles du reste des positions? Si oui, cette structure locale commence à différer à quelle distance?

Nous avons montré que ni les positions fonctionnellement sensibles ni les positions enterrées ne se distinguent structurellement du reste lorsque la structure locale est définie sur une très courte distance (moins de quatre Ångströms). Cependant, certaines différences apparaissent entre les types

de position (fonctionnellement sensibles / non-fonctionnellement sensibles, enterrées / positions de surface) lors de l'utilisation d'un cutoff supérieure à quatre Ångströms.

Nous avons utilisé trois mesures de changement structurel basées sur nos travaux précédents sur les réseaux de hotspots. La première compte le nombre d'interactions atomiques perturbées par une mutation. La seconde est le nombre d'acides aminés perturbés, c'est-à-dire d'acides aminés ayant au moins un atome perturbé. Et la troisième est la plus grande distance dans la structure 3D entre la position de la mutation et un autre acide aminé perturbé.

Lorsque l'interaction atomique est définie sur quatre Ångströms ou moins, aucune différence n'est faite entre les positions, pour toute mesure structurale. L'utilisation de plus grands cutoffs que la distance chimique, vise à enquêter les relations structurelles entre deux parties de la protéine qui ne sont pas nécessairement chimiquement liés. Ceci est important à étudier, car cela permet un inventaire plus complet des différences structurelles autour de différentes positions.

De plus, des mutations aux positions fonctionnellement sensibles ont perturbé plus d'acides aminés en général. Nous avons trouvé que il y a une corrélation allant jusqu'à soixante-dix pour cent entre le changement fonctionnel et le nombre d'acides aminés perturbés par la mutation. D'autre part, il n'y a pas de lien significatif entre changement fonctionnel et nombre d'interactions perturbées. La perturbation des acides aminés joue donc un rôle plus important dans le changement fonctionnel que la perturbation des interactions atomiques ou d'acide aminés, c'est-à-dire que ce n'est pas le nombre d'interactions que la mutation peut perturber ce qui compte en termes de changement fonctionnel, mais le nombre d'acides aminés perturbés.

En effet, nous avons trouvé que le nombre d'interactions atomiques ou d'acides aminés perturbées ne sont pas caractéristiques à des positions fonctionnellement sensibles, car elles affectent de la même manière le reste des positions fonctionnellement robustes. D'autre part, un plus grand nombre d'acides aminés perturbés a été trouvé d'être corrélé à des positions fonctionnellement sensibles. Nous avons suggéré que ce résultat pourrait soustendre l'existence de plusieurs voies de communication entre les acides aminés comme méthode de code correcteur. D'autres expérimentations doivent être réalisées pour tester cette hypothèse, cependant.

La corrélation entre le nombre de noeuds perturbés et le changement fonctionnel implique une relation entre la structure locale des positions fonctionnellement sensibles et le changement structurel de la protéine lorsqu'elle est mutée. Le changement de la configuration atomique dû au remplace-

ment de la position d'un acide aminé par un autre, peut être abordé par une analyse statistique de où une mutation à une position donnée, serait simulée par un changement aléatoire dans la configuration atomique de la protéine. En effet, cela permettrait un nombre arbitraire de changement de valeurs structurales s'appuyant seulement sur l'environnement structurel local de la position en cours de mutation.

7.3 Vide local

La dernière partie de ma dissertation était la mise en oeuvre de trois algorithmes différents pour quantifier l'espace vide autour des groupes d'atomes. Cela a été fait en utilisant la tessellation de Delaunay de l'ensemble des coordonnées atomiques 3D de la structure, ainsi que l'enveloppe convexe des coordonnées atomiques des résidus. En outre, l'un des algorithmes considère les atomes comme des sphères de rayons égaux à leurs rayons de van der Waals, afin de prendre en compte les tailles atomiques.

Les méthodes utilisées pour calculer le vide dépendent de la tessellation de Delaunay basée sur les structures atomiques des protéines, peuvent contenir des tétraèdres aberrants, généralement à la surface, ou dans des "trous" dans le cas des oligomères. Nous avons trouvé que les longueurs des arêtes des tétraèdres varient peu autour de leur moyenne, qui est plus petite que la distance chimique d'interaction atomique. Cependant, les longueurs d'arêtes peuvent avoir des valeurs dépassant la longueur totale des protéines complètes dans certains cas, ce qui suggère que de très grands vides sont calculés sans l'utilisation d'un cutoff sur la longueur des arêtes. Les méthodes non-dépendantes des cutoffs, mais qui éliminent les cas aberrants, peuvent être utilisés à la place, au détriment d'une plus grande complexité algorithmique. En outre, nous avons démontré que la grande majorité des longueurs d'arêtes tourne près de la distance chimique de 5 Ångströms, ce qui suggère qu'un calcul juste du vide est fait dans la protéine en utilisant ce seuil chimique. En effet, deux de nos méthodes utilisant des cutoffs ont montré une forte corrélation des valeurs du vide pour les résidus appartenant à deux toxines structurellement similaires, ce qui suggère une juste estimation du vide local autour des résidus.

Le vide utilisant la tessellation de Delaunay et un cutoff est distribué normalement à travers les positions dans la protéine, montrant un effet "Goldilocks" de l'espace vide autour des résidus, où ils ont des valeurs moyennes de vide qui ne sont ni trop petites ni trop grande par rapport à la valeur moyenne. Le vide d'un résidu dépend de sa position dans la structure 3D

de la protéine, plus précisément de son espace disponible à l'intérieur de la protéine. C'est-à-dire que les résidus enfouis ont plus d'espace, et donc plus d'espace vide autour d'eux par rapport aux positions de surface.

7.4 Travaux futurs

Dans le futur, nous proposons d'estimer la structure locale des acides aminés par une sphère centrée sur le centroïde de chaque acide aminé, et le rayon égal à cinq Ångströms, pour mesurer la distribution des atomes de la structure locale en partitionnant la sphère en plusieurs sous-ensembles de points 3D, et en comptant le nombre d'atomes par sous-ensemble. Cela indiquerait la distribution spatiale des structures locales, distinguant probablement entre deux structures locales autrement similaires en poids et en degré, *neighborhood watch* ou enterrée/position de surface.

D'autres études sur les méthodes de tessellation utilisant un cutoff, devront être effectuées en comparant les valeurs du vide de nos mesures avec des méthodes indépendantes de cutoffs, pour éclairer le compromis entre la complexité algorithmique et la précision des valeurs du vide. Plus précisément, une étude incluant un grand ensemble de données pourrait être utilisée pour comparer des vides. Enfin, nos méthodes permettent un calcul du vide relatif, ou vide local des acides aminés, favorable à l'étude des structures locales.

Bibliography

[1] Mounia Achoch, Rodrigo Dorantes-Gilardi, Chris Wymant, Giovanni Feverati, Kave Salamatian, Laurent Vuillon, and Claire Lesieur. Protein structural robustness to mutations: an in silico investigation. *Physical Chemistry Chemical Physics*, 18(20):13770–13780, 2016.

[2] Mounia Achoch, Giovanni Feverati, Laurent Vuillon, Kavé Salamatian, and Claire Lesieur. Protein subunit association: Not a social network. In *Theoretical Approaches to Bioinformation systems*, 2013.

[3] Gil Amitai, Rinkoo Devi Gupta, and Dan S Tawfik. Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP journal*, 1(1):67, 2007.

[4] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[5] Orly Bachar, Daniel Fischer, Ruth Nussinov, and Haim Wolfson. A computer vision based technique for 3-d sequence-independent structural comparison of proteins. *Protein Engineering, Design and Selection*, 6(3):279–287, 1993.

[6] Albert-László Barabasi. Network biology. *The Febs Journal*, 272:433, 2005.

[7] Bogdan Barz, David J Wales, and Birgit Strodel. A kinetic approach to the sequence–aggregation relationship in disease-related protein assembly. *The Journal of Physical Chemistry B*, 118(4):1003–1011, 2014.

[8] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Gossip algorithms: Design, analysis and applications. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 3, pages 1653–1664. IEEE, 2005.

[9] KV Brinda and Saraswathi Vishveshwara. A network representation of protein structures: implications for protein stability. *Biophysical journal*, 89(6):4159–4170, 2005.

[10] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[11] Marc JS De Wolf, Guido AF Van Dessel, Albert R Lagrou, Herwig JJ Hilderson, and Wilfried SH Dierick. ph-induced transitions in cholera toxin conformation: a fluorescence study. *Biochemistry*, 26(13):3799–3806, 1987.

[12] Eynat Dellus-Gur, Mikael Elias, Emilia Caselli, Fabio Prati, Merijn LM Salverda, J Arjan GM de Visser, James S Fraser, and Dan S Tawfik. Negative epistasis and evolvability in tem-1 β-lactamase—the thin line between an enzyme's conformational freedom and disorder. *Journal of molecular biology*, 427(14):2396–2409, 2015.

[13] Özlem Demir, Roberta Baronio, Faezeh Salehi, Christopher D Wassman, Linda Hall, G Wesley Hatfield, Richard Chamberlin, Peter Kaiser, Richard H Lathrop, and Rommie E Amaro. Ensemble-based computational approach discriminates functional activity of p53 cancer and rescue mutants. *PLoS computational biology*, 7(10):e1002238, 2011.

[14] Luisa Di Paola and Alessandro Giuliani. Protein contact network topology: a natural language for allostery. *Current opinion in structural biology*, 31:43–48, 2015.

[15] Declan A Doyle, Alice Lee, John Lewis, Eunjoon Kim, Morgan Sheng, and Roderick MacKinnon. Crystal structures of a complexed and peptide-free membrane protein–binding domain: molecular basis of peptide recognition by pdz. *Cell*, 85(7):1067–1076, 1996.

[16] A Elisabeth Eriksson, Walter A Baase, Xue-Jun Zhang, Dirk W Heinz, MPBE Blaber, Enoch P Baldwin, and Brian W Matthews. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*, 255(5041):178–183, 1992.

[17] Erkang Fan, Ethan A Merritt, Zhongsheng Zhang, Jason C Pickens, Claudia Roach, Misol Ahn, and Wim GJ Hol. Exploration of the

gm1 receptor-binding site of heat-labile enterotoxin and cholera toxin by phenyl-ring-containing galactose derivatives. *Acta Crystallographica Section D: Biological Crystallography*, 57(2):201–212, 2001.

[18] Victoria A Feher, Jacob D Durrant, Adam T Van Wart, and Rommie E Amaro. Computational approaches to mapping allosteric pathways. *Current opinion in structural biology*, 25:98–103, 2014.

[19] Giovanni Feverati, Mounia Achoch, Laurent Vuillon, and Claire Lesieur. Intermolecular β-strand networks avoid hub residues and favor low interconnectedness: A potential protection mechanism against chain dissociation upon mutation. *PloS one*, 9(4):e94745, 2014.

[20] Daniel Flatow, Sumudu P Leelananda, Aris Skliros, Andrzej Kloczkowski, and Robert L Jernigan. Volumes and surface areas: geometries and scaling relationships between coarse-grained and atomic structures. *Current pharmaceutical design*, 20(8):1208–1222, 2014.

[21] Greta Gronau, Sreevidhya T Krishnaji, Michelle E Kinahan, Tristan Giesa, Joyce Y Wong, David L Kaplan, and Markus J Buehler. A review of combined experimental and computational procedures for assessing biopolymer structure–process–property relationships. *Biomaterials*, 33(33):8240–8255, 2012.

[22] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2):369–387, 2002.

[23] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[24] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

[25] Yuichiro Higashimoto, Yuya Asanomi, Satoru Takakusagi, Marc S Lewis, Kohei Uosaki, Stewart R Durell, Carl W Anderson, Ettore Appella, and Kazuyasu Sakaguchi. Unfolding, aggregation, and amyloid formation by the tetramerization domain from mutant p53 associated with lung cancer. *Biochemistry*, 45(6):1608–1619, 2006.

[26] Timothy R Hirst, J Moss, B Iglewski, M Vaughan, and AT Tu. Biogenesis of cholera toxin and related oligomeric enterotoxins. *Bacterial toxins and virulence factors in disease. Marcel Dekker, Inc., New York, NY*, pages 123–184, 1995.

[27] HR Horton, LA Moran, KG Scrimgeour, MD Perry, and JD Rawn. Lipids and membranes. *Principles of Biochemistry. Pearson Prentice Hall, Upper Saddle River, New Jersey*, pages 253–292, 2006.

[28] Dmitry N Ivankov, Alexei V Finkelstein, and Fyodor A Kondrashov. A structural perspective of compensatory evolution. *Current opinion in structural biology*, 26:104–112, 2014.

[29] Tao Jia, Yang-Yu Liu, Endre Csóka, Márton Pósfai, Jean-Jacques Slotine, and Albert-László Barabási. Emergence of bimodality in controlling complex networks. *arXiv preprint arXiv:1505.06476*, 2015.

[30] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.

[31] Ulrich K Laemmli. Cleavage of structural proteins during the assembly of the head of bacteriophage t4. *nature*, 227(5259):680–685, 1970.

[32] David M Leitner. Frequency-resolved communication maps for proteins and other nanoscale materials. *The Journal of chemical physics*, 130(19):05B606, 2009.

[33] David M Leitner, Sebastian Buchenberg, Paul Brettel, and Gerhard Stock. Vibrational energy flow in the villin headpiece subdomain: Master equation simulations. *The Journal of chemical physics*, 142(7):02B608_1, 2015.

[34] Claire Lesieur. The assembly of protein oligomers—old stories and new perspectives with graph theory. In *Oligomerization of Chemical and Biological Compounds*. InTech, 2014.

[35] Claire Lesieur, Matthew J Cliff, Rachel Carter, Roger FL James, Anthony R Clarke, and Timothy R Hirst. A kinetic model of intermediate formation during assembly of cholera toxin b-subunit pentamers. *Journal of Biological Chemistry*, 277(19):16697–16704, 2002.

[36] Jie Liang and Ken A Dill. Are proteins well-packed? *Biophysical journal*, 81(2):751–766, 2001.

[37] Kaj Ulrik Linderstrøm-Lang. *Lane medical lectures: proteins and enzymes*, volume 6. Stanford University Press, 1952.

[38] Tong Liu, Steven T Whitten, and Vincent J Hilser. Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble. *Proceedings of the National Academy of Sciences*, 104(11):4347–4352, 2007.

[39] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *nature*, 473(7346):167, 2011.

[40] Sébastien Loriot, Frédéric Cazals, and Julie Bernauer. Esbtl: efficient pdb parser and data structure for the structural and geometric analysis of biological macromolecules. *Bioinformatics*, 26(8):1127–1128, 2010.

[41] Dubravka Matković-Calogović, Arianna Loregian, Maria Rosa D'Acunto, Roberto Battistutta, Alessandro Tossi, Giorgio Palù, and Giuseppe Zanotti. Crystal structure of the b subunit of escherichia coli heat-labile enterotoxin carrying peptides with anti-herpes simplex virus type 1 activity. *Journal of Biological Chemistry*, 274(13):8764–8769, 1999.

[42] Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138, 2012.

[43] Eric A Ortlund, Jamie T Bridgham, Matthew R Redinbo, and Joseph W Thornton. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, 317(5844):1544–1548, 2007.

[44] Venkata N Padmanabhan, Helen J Wang, and Philip A Chou. Resilient peer-to-peer streaming. In *Network Protocols, 2003. Proceedings. 11th IEEE International Conference on*, pages 16–27. IEEE, 2003.

[45] Gustavo Parisi, Diego Javier Zea, Alexander Miguel Monzon, and Cristina Marino-Buslje. Conformational diversity and the emergence of sequence signatures during evolution. *Current opinion in structural biology*, 32:58–65, 2015.

[46] Linus Pauling, Robert B Corey, and Herman R Branson. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205–211, 1951.

[47] Joshua L Payne and Andreas Wagner. The robustness and evolvability of transcription factor binding sites. *Science*, 343(6173):875–877, 2014.

[48] Art FY Poon and Lin Chao. Functional origins of fitness effect-sizes of compensatory mutations in the dna bacteriophage ϕx174. *Evolution*, 60(10):2032–2043, 2006.

[49] Owen JL Rackham, Martin Madera, Craig T Armstrong, Thomas L Vincent, Derek N Woolfson, and Julian Gough. The evolution and structure prediction of coiled coils across all genomes. *Journal of molecular biology*, 403(3):480–493, 2010.

[50] D Reichmann, O Rahat, S Albeck, R Meged, O Dym, and G Schreiber. The modular architecture of protein–protein binding interfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 102(1):57–62, 2005.

[51] Frederic M Richards. Areas, volumes, packing, and protein structure. *Annual review of biophysics and bioengineering*, 6(1):151–176, 1977.

[52] Mary M Rorick and Günter P Wagner. Protein structural modularity and robustness are associated with evolvability. *Genome biology and evolution*, 3:456–475, 2011.

[53] Kristian Rother, Peter Werner Hildebrand, Andrean Goede, Bjoern Gruening, and Robert Preissner. Voronoia: analyzing packing in protein structures. *Nucleic acids research*, 37(suppl_1):D393–D395, 2008.

[54] Kristian Rother, Robert Preissner, Andrean Goede, and Cornelius Frömmel. Inhomogeneous molecular density: reference packing densities and distribution of cavities within proteins. *Bioinformatics*, 19(16):2112–2121, 2003.

[55] Lloyd W Ruddock, Jeremy JF Coen, Caroline Cheesman, Robert B Freedman, and Timothy R Hirst. Assembly of the b subunit pentamer of escherichia coli heat-labile enterotoxin kinetics and molecular basis of rate-limiting steps in vitro. *Journal of Biological Chemistry*, 271(32):19118–19123, 1996.

[56] Lloyd W Ruddock, Stephen P Ruston, Sharon M Kelly, Nicholas C Price, Robert B Freedman, and Timothy R Hirst. Kinetics of acid-mediated disassembly of the b subunit pentamer of escherichia coli heat-labile enterotoxin molecular basis of ph stability. *Journal of Biological Chemistry*, 270(50):29953–29958, 1995.

[57] Merijn LM Salverda, Eynat Dellus, Florien A Gorter, Alfons JM Debets, John Van Der Oost, Rolf F Hoekstra, Dan S Tawfik, and J Arjan GM de Visser. Initial mutations direct alternative pathways of protein evolution. *PLoS genetics*, 7(3):e1001321, 2011.

[58] Uttamkumar Samanta, Ranjit P Bahadur, and Pinak Chakrabarti. Quantifying the accessible surface area of protein residues in their local environment. *Protein engineering*, 15(8):659–667, 2002.

[59] Carmelinda Savino, Adriana E Miele, Federica Draghi, Kenneth A Johnson, Giuliano Sciara, Maurizio Brunori, and Beatrice Vallone. Pattern of cavities in globins: the case of human hemoglobin. *Biopolymers*, 91(12):1097–1107, 2009.

[60] Charles D Schwieters, John J Kuszewski, Nico Tjandra, and G Marius Clore. The xplor-nih nmr molecular structure determination package. *Journal of magnetic resonance*, 160(1):65–73, 2003.

[61] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The foldx web server: an online force field. *Nucleic acids research*, 33(suppl_2):W382–W388, 2005.

[62] Maxim Shatsky, Ruth Nussinov, and Haim J Wolfson. Algorithms for multiple protein structure alignment and structure-derived multiple sequence alignment. In *Protein Structure Prediction*, pages 125–146. Springer, 2008.

[63] Peter Shirley, Michael Ashikhmin, and Steve Marschner. *Fundamentals of computer graphics*. CRC Press, 2015.

[64] A Shrake and JA Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of molecular biology*, 79(2):351–371, 1973.

[65] Penelope E Stein, Amechand Boodhoo, Gregory J Tyrrell, James L Brunton, and Randy J Read. Crystal structure of the cell-binding b oligomer of verotoxin-1 from e. coli. *Nature*, 355(6362):748–750, 1992.

[66] Severin Strobl, Arno Formella, and Thorsten Pöschel. Exact calculation of the overlap volume of spheres and mesh elements. *Journal of Computational Physics*, 311:158–172, 2016.

[67] Joanna F Swain and Lila M Gierasch. The changing landscape of protein allostery. *Current opinion in structural biology*, 16(1):102–108, 2006.

[68] Ágnes Tóth-Petróczy and Dan S Tawfik. The robustness and innovability of protein folds. *Current opinion in structural biology*, 26:131–138, 2014.

[69] Brandon H Toyama and Martin W Hetzer. Protein homeostasis: live long, won't prosper. *Nature Reviews Molecular Cell Biology*, 14(1):55–61, 2013.

[70] Chung-Jung Tsai, Antonio Del Sol, and Ruth Nussinov. Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. *Molecular Biosystems*, 5(3):207–216, 2009.

[71] Vicente Tur, Almer M van der Sloot, Carlos R Reis, Eva Szegezdi, Robbert H Cool, Afshin Samali, Luis Serrano, and Wim J Quax. Dr4-selective tumor necrosis factor-related apoptosis-inducing ligand (trail) variants obtained by structure-based design. *Journal of Biological Chemistry*, 283(29):20560–20568, 2008.

[72] Louis-Nicolas Vauquelin and Pierre J Robiquet. The discovery of a new plant principle in asparagus sativus. *Ann. Chim.(Paris)*, 57(2):1, 1806.

[73] Beatrix Vécsey-Semjén, Claire Lesieur, Roland Möllby, and F Gisou van der Goot. Conformational changes due to membrane binding and channel formation by staphylococcal α-toxin. *Journal of Biological Chemistry*, 272(9):5709–5717, 1997.

[74] Laurent Vuillon and Claire Lesieur. From local to global changes in proteins: a network view. *Current opinion in structural biology*, 31:1–8, 2015.

[75] Andreas Wagner. Neutralism and selectionism: a network-based reconciliation. *Nature Reviews Genetics*, 9(12):965, 2008.

[76] Andreas Wagner. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society of London B: Biological Sciences*, 275(1630):91–100, 2008.

[77] Andreas Wagner. Mutational robustness accelerates the origin of novel rna phenotypes through phenotypic plasticity. *Biophysical journal*, 106(4):955–965, 2014.

[78] Nils Woetzel, Mert Karakaş, Rene Staritzbichler, Ralf Müller, Brian E Weiner, and Jens Meiler. Bcl:: Score—knowledge based energy po-

tentials for ranking protein models represented by idealized secondary structure elements. *PloS one*, 7(11):e49242, 2012.

[79] Adam Zlotnick, Zhenning Tan, and Lisa Selzer. One protein, at least three structures, and many functions. *Structure*, 21(1):6–8, 2013.

[80] Jihad Zrimi, Alicia Ng Ling, Ernawati Giri-Rachman Arifin, Giovanni Feverati, and Claire Lesieur. Cholera toxin b subunits assemble into pentamers-proposition of a fly-casting mechanism. *PLoS One*, 5(12):e15347, 2010.