Alma Mater Studiorum · Università di Bologna

Scuola di Scienze

Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

Analysis of thermophilic and mesophilic proteins through contact map networks

Relatore:
Prof. Daniel Remondini

Presentata da:
Arianna Pace

Anno Accademico 2015/2016

Abstract

In questa tesi si affronta lo studio di proteine termofile e mesofile con un approccio a network, con lo scopo di cercarne differenze strutturali che giustifichino la minore termolabilità delle proteine termofile.

La teoria dei grafi, nata come teoria matematica, ha subito negli ultimi anni, in particolare dalla fine degli anni '90 grazie al notevole sviluppo tecnologico, notevoli progressi trovando applicazioni in vari ambiti, tra cui quello della biologia, risultando un nuovo strumento per approcciare problemi complessi. Applicare la teoria dei grafi allo studio delle proteine significa modellizzarne la struttura 3D con una matrice 2D, la mappa di contatto proteica, operando una compressione dell'informazione. La perdita di dettaglio è compensata dall'ottenimento di un oggetto matematico facilmente trattabile con una chiara interpretazione fisica. In questa tesi è stato scelto di concentrarsi sulle proprietà spettrali del suo Laplaciano in quanto queste sono strettamente legate alle proprietà vibrazionali del sistema. L'ipotesi è che si possa trovare così una differenza tra proteine termofile e mesofile in quanto, secondo l'ipotesi di stati equivalenti, si suppone che una proteina termofila a temperatura ambiente sia più rigida di una mesofila e che queste abbiano una flessibilità simile solamente alle loro rispettive temperature ottimali.

Il database analizzato è stato costruito come una serie di coppie di proteine omologhe, una mesofila e una termofila. Questo permette di cercare differenze tra proteine simili, le cui differenze ci si aspetta siano dovute agli adattamenti per sopravvivere in habitat con diverse temperature. Su questo dataset sono state effettuate sia misure strutturali più tradizionali, sia è stato studiato lo spettro del Laplaciano delle loro mappe di contatto. Se i primi non hanno presentato differenze significative tra i due gruppi di proteine, un risultato interessante è stato ottenuto proprio con l'approccio a network. I primi autovalori del Laplaciano, associati quindi con basse frequenze di vibrazione, riescono a discriminare proteine termofile e mesofile, in oltre il 65% delle coppie – da confrontare con percentuali di discriminazione in letteratura recente che, utilizzando solo proprietà strutturali delle proteine, non arrivano al 60% [17].

La tesi è satata suddivisa in cinque capitoli per presentare sia un'introduzione teorica biologica e matematica, sia il lavoro svolto. ll primo capitolo inizia con una breve introduzione biologica al problema. La prima sezione descrive come siano fatte

le proteine e la loro struttura. Nella seconda si delinea la relazione tra la struttura, la funzionalità proteica e la temperatura, quindi si introducono gli estremofili che vivono in condizioni "estreme" per i canoni delle cellule umane, concludendo la sezione con la descrizione degli adattamenti riscontrati negli estremozimi, proteine che sono funzionali ad alte temperature. L'ultima parte del capitolo è invece riservata ai metodi sperimentali usati per conoscere la forma di una proteina, in particolare viene approfondita la cristallografia a raggi X, e alla conservazione e diffusione di tali informazioni nell'archivio a libero accesso Protein Data Bank.

Il secondo capitolo ha lo scopo di presentare i metodi utilizzati per l'analisi delle proteine. Per questo motivo, prima di definire le mappe di contatto e di discuterne le diverse variazioni, sono introdotti degli elementi di teoria dei grafi. Il capitolo si conclude con una presentazione del Laplaciano di un network e di come il suo spettro possa essere usato come strumento d'analisi.

L'elaborato quindi prosegue con un capitolo dedicato alla creazione del database di proteine poi usato per l'analisi. Si inizia con la descrizione della ricerca in letteratura per raccogliere coppie di proteine omologhe una proveniente da un organismo termofilo e l'altra da uno mesofilo, che fossero molto simili. Da questo studio, è stato ottenuto un elenco di 447 coppie; non tutte sono state inserite nel database finale, ma solo quelle erano conformi ai criteri di qualità descritti, scendendo così a 65 paia. Questa selezione è descritta nella seconda parte del capitolo.

Il quarto capitolo contiene il resoconto delle analisi effettuate. Per prima cosa si presentano le matrici di distanza e gli istogrammi delle distanze tra residui delle proteine nel database. Quindi vengono mostrate le mappe di contatto, le frequenze di contatto per diagonale e i valori di Contact Order e della sua variante Long Range Contact Order. Si passa poi allo spettro del Laplaciano, analizzato per cercare differenze tra le mappe dei termofili e quelle dei mesofili, e si conclude guardando alle vibrazioni dei residui, approssimando la proteina ad un network di sfere collegate con molle di uguale costante elastica. È quest'ultima analisi che mostra i risultati più promettenti per comprendere le differenze strutturali tra mesofili e termofili, mentre per le altre le differenze tra i due gruppi non sono significative.

Nell'ultimo capitolo si presentano le riflessioni conclusive. Le conclusioni riassumono i risultati ottenuti, che vengono passati in rassegna e discussi criticamente, e propongono spunti per lavori futuri a partire da quanto presentato in questa tesi.

Contents

List of Figures

List of Tables

1

Protein structure and temperature

In this chapter some background is given about how different proteins behave across different temperature ranges and why it is interesting to find proteins that keep their functionality intact at extreme temperatures. Firstly a brief description of the structure of a generic protein will be given, followed by a brief summary of the known strategies that proteins adapt in order to survive high temperatures whilst maintaining their structure intact. Most of the content from these sections is informed by the textbook *Cambell Biology*, ninth edition [33]. The final section discusses the techniques that are used to obtain a three-dimensional model of a protein and provide a description of the Protein Data Bank archive from which the structures used in this thesis were taken from.

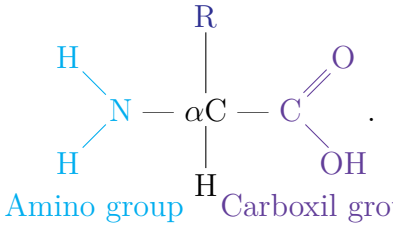
1.1 What is a protein

Proteins are essential for life as we know it; almost all the dynamic functions that take place within a living being depend on proteins. In particular, life would not be possible without *enzymes* that speed up chemical reactions. Most of the enzymes in living organisms are proteins, working as catalysts. In nature there is a huge number of various proteins; humans have tens of thousands different variants. However, every protein consists of a specific sequence of the same twenty *amino acids*. Amino acids link together with *peptide bonds*, forming a chain. After the amino acids lose an H_2O molecule to create the peptide bond, they are defined as *residue*. Proteins are made of one or more amino acid polymers, named *polypeptides*. The polypeptides are called protein when they are folded into a three-dimensional structure. The folding and coiling of the protein is generally spontaneous under normal cellular conditions. The bonds formed between different parts of the chain, that define the final shape of the protein, depend on its particular amino acid sequence. That said, there is a great

deal of research interest around how the amino acid chain transitions to a protein via a series of intermediate states: the *protein folding*. The functionality of the protein is defined by its structure which ultimately determines how it interacts with other molecules or macromolecules. But before we getting there, we must first consider in more detail the fundamental building blocks of a protein: the amino acids.

1.1.1 Amino acids

Proteins are ordinate chains of amino acids that fold into a three dimensional structure. Each of the twenty different amino acids has the same basic structure: an amino group and a carboxyl group, held together by a carbon atom, called the α carbon. The other two links of the α carbon are made with a side chain R, which is the part that differentiate one amino acid from the others, and with an hydrogen atom. The schematic representation of a amino acid monomers is shown below:



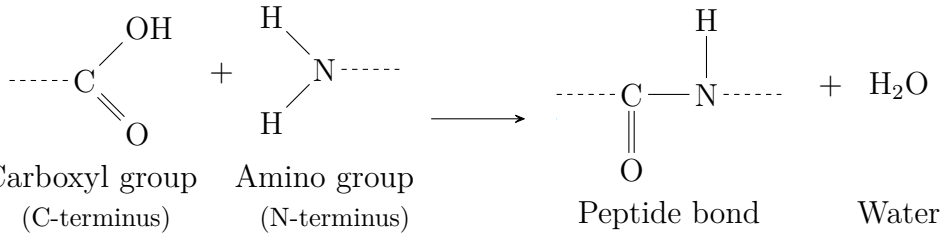
The side chain varies significantly for different amino acids. It can be as subtle as the displacement of a single hydrogen atom, as is the case of the smallest amino acid possible Glycine that is the only amino acid that does not have a β carbon, or as complex as a carbon skeleton with other functional groups attached, as in the one of Glutamine. The physical and chemical properties of the group R determines the properties and the functional roles of a particular amino acid. We can classify each amino acid depending on these properties, as shown in table 1.1.

Hydrophobic			**Hydrophilic**					
Non polar			Polar			Electrically charged		
						Basic, positively charged		
Glycine (Gly or G)	Alanine (Ala or A)	Valine (Val or V)	Serine (Ser or S)	Threonine (Thr or T)	Cysteine (Cys or C)	Lysine (Lys or K)	Arginine (arg or R)	Histidine (His or H)
						Acidic, negatively charged		
Leucine (Leu or L)	Isoleucine (Ile or I)	Methionine (Met or M)	Tyrosine (Tyr or Y)	Asparagine (Asn or N)	Glutamine (Gln or Q)	Aspartic acid (Asp or D)	Glutamic acid (Glu or E)	
Phenylalanine (Phe or F)	Trypyophan (Trip or W)	Proline (Pro or P)						

Table 1.1: Classification of all the 20 amino acids based on the properties of their side chain.

The various amino acids connect together in sequence, by means of the peptide bonds. Peptide bonds are formed as a result of dehydration synthesis, where a

covalent bond is created by removing a water molecule from the carboxyl group of one amino acid and the amino group of the other, as shown by the following schematic reaction:



These concatenations of amino acids are polymers called peptides – or polypeptides when they are formed by numerous residues. Peptides, unless they are cyclic, have two terminals: an amino end, said N-terminus, and a carboxyl one, called C-terminus.

The linear order of amino acids in the peptide, which determines the chemical properties of the molecule, is specific and unique for each one. The first protein to have its amino acid sequenced was the insulin hormone in the early 1950s, thanks to the work of Frederick Sanger [36]. He and his team worked over 12 years before determining the sequence. Prior to this breakthrough, it was only possible to measure with a certain precision the relative quantities of amino acids within a protein, without any sense of how they were arranged. Nowadays protein sequencing is an automated process and there are various methods to perform it. One process is the *Edman degradation*. In this method the N-terminus of the protein is labelled and detached form the chain without breaking other peptide bonds. By repeating this reaction, one can learn the amino acid order of the protein. An indirect way to discern the sequence of a protein, is by sequencing the gene that produces it. If the genetic material that codes for the protein is known, this is actually easier than obtaining the sequence from the protein itself.

Learning the order of the amino acids composing a polypeptide is important, but it is not sufficient to discern the properties of the final protein. That is why knowing the three-dimensional structure is crucial for the study of a protein.

1.1.2 Protein structure

The shape of a protein is the result of interactions and bonds that take place between its the components. This bonds can be seen as belonging to different structure orders. The *primary structure* is the linear chain of residues, that, as we have already seen, are held together by peptide bonds. This primary structure is quite far from the shape of the final protein. When a cell synthesizes a polypeptide, it folds itself helped by specialized proteins called *chaperones*, firstly by assuming a *secondary structure* and then compacting into a *tertiary* one. An additional *quaternary structure* is present only for certain proteins.

The secondary structure is composed by coils and foils; both of these structures are the result of hydrogen bonds between backbone oxygens and amide hydrogens. Coiled parts are called α-helix, structures that are held together by hydrogen bonding between every forth residues. A protein can have more α-helices spaced out by flat regions. The other regular secondary structure is the β-pleated sheet, or simply β-sheet. In this case two or more strands of the polypeptide, the β-strands, typically composed of 3 to 10 amino acids, are lying in parallel. This formation is kept together by a wide network of hydrogen bonds between neighbours.

The tertiary structure is superimposed to the secondary one. The main difference between the secondary and the tertiary structure is that, while the first is realized by hydrogen bonds between the backbone elements, the second is the result of interactions between the various side chains. In fact, hydrogen bonds are considered belonging to the tertiary structure when they are connecting hydrophilic side chains, but there are more types of interactions that can occur at this level. One is the hydrophobic interaction. This interaction is obtained by the spontaneous formation of clusters of hydrophobic side chains in the core of the protein, repelling the contact with water, and the stability of these clusters is increased by the formation of Van der Waals interactions between these non-polar R groups. Another kind of bonds that can occur at this level are ionic bonds between opposite charged side chains. Furthermore, also part of the secondary structure are disulphide bridges (–S–S–) that can form between two cysteines that happen to have their sulfhydryl groups (–SH) near each other, further reinforcing the stability of the protein.

In addition to these three structure levels – primary, secondary and tertiary – that are always present, there can be also a *quaternary* one. The quaternary structure is the overall functional macromolecule that is the result of an aggregation of more than one polypeptide chain. A well known example of this structure is haemoglobin, that consists of four polypeptide subunits.

Every protein has its own structure and interestingly enough there are also proteins that present functional parts that remain unfolded. Since the structure of the protein is formed by interactions between its components, it can be jeopardized when those bondings are put under stress, like in the case of high thermal agitation.

1.2 Structure and temperature

Many factors of the environment where the protein is affect its structure and, accordingly, its functionality. In particular temperature, pH and pressure have major consequences on the protein structure and on how efficiently it works. Each protein has its optimal conditions, that is when its most active shape is favoured.

With regards to the temperature, the functionality of a protein changes quite rapidly. Initially, as the temperature rises so does the activity of the protein. Then, above a certain threshold, which is considered the *optimal temperature* for that

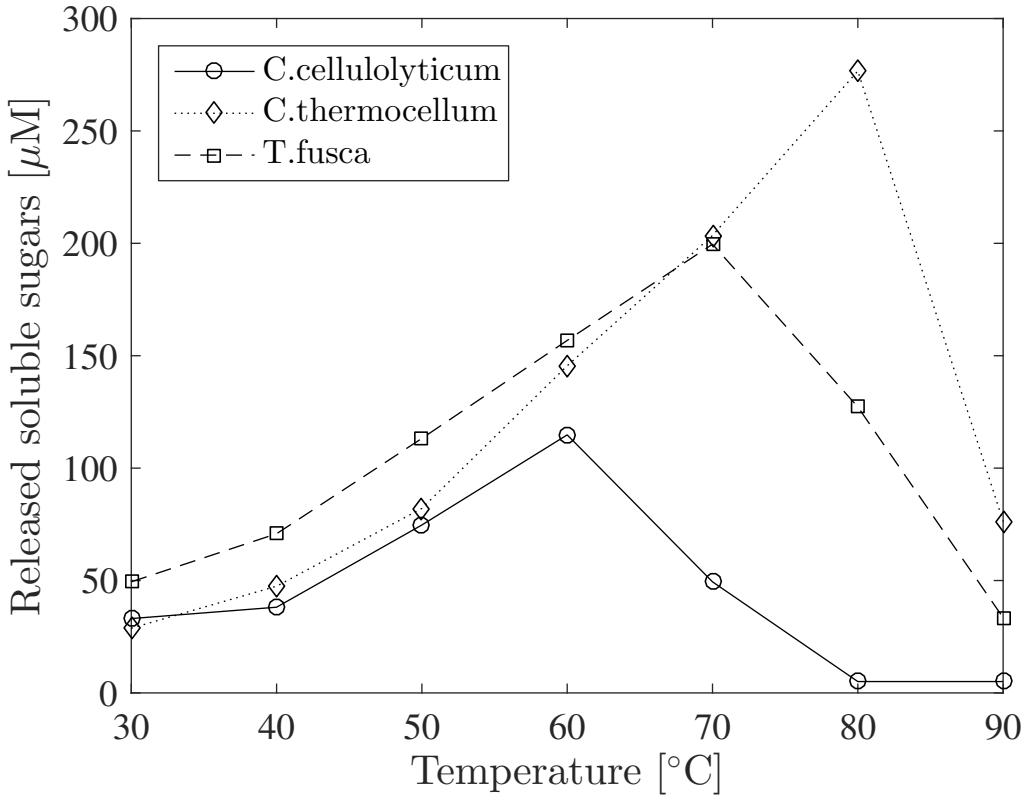


Figure 1.1: The temperature dependence in the activities of three homologous proteins, *Cel9 cellulases*. The protein *cellulase* is an enzyme that helps the glycolysis of the complex sugar molecule cellulose into monosaccharides. These enzymes come from three different bacteria which live at three different optimal temperature: the *Clostridium cellulolyticum*, the *Thermobifida fusca* and the *Clostridium thermocellum*. In the graph it can be seen at different temperature the quantity of released sugars, the product of the reaction that this protein enhances, assayed after the same amount of time and the same conditions for each one. The data plotted in this figure is taken from [27].

protein, the activity declines. An example of these curves is shown in figure 1.1, where three *homologous* enzymes, i.e. that perform the same task in three different organisms, have different optimal temperatures. The trend of those curves is due to thermal agitation. As a matter of fact, thermal agitation augment the probability of collision between the active site of the protein and the substrates, accelerating the reaction, but this is true only up to a certain temperature, the optimal one for that enzyme, at which the enzyme is the most active. After that, a further increase in temperature results in the break of bonds that stabilise the active structure of the protein, lowering its functionality and eventually causing its *denaturation*. A protein is said to denature when the disruption of its structure causes the loss of its

functionality.

It is clear that every protein has its own optimal temperature. Out of the window of temperatures where the structure is solid enough to maintain its active sites stable, but at the same time flexible enough to be able to interact with the substrates, the protein can not perform its task. Since cellular life depends on protein activities, major damages, that can lead to the death of the organism, arise when proteins are out of their temperature range. Most of the cells, human ones comprised, live at temperatures between 35 °C and 40 °C. Complex organisms, instead, can survive out of their preferred interval if they are able to spend energy to keep their cells within the right temperatures, as in the case of mammals and birds. Organisms that live at these temperatures are called *mesophiles* (which literally means *who loves the middle*, and it is composed by the Greek words μέσος, *middle*, and φιλία, *love*). There are, however, other organisms that thrive at other temperature ranges thanks to some extraordinary adaptations. They are part of a group called extremophiles (literally meaning *who loves extreme conditions*, from the Latin *extremus*), and they can be classified into two types:

- the *thermophiles* (*who loves the heat*, from the Greek θερμότητα, *heat*), who thrive at temperatures between 41 °C and 122 °C – if their growth temperature is higher than 85 °C they are called *hyperthermophiles* and interestingly many of them can not even reproduce themselves at temperature below 80 °C;

- the *psychrophile* or *cryophiles* (*who loves the cold*, from the Greek ψυχρός or κρύος, *cold*), who live at temperature between 20 °C and −10 °C Celsius.

In nature other types of extremophiles exist. As an example, there are organisms that can survive high salt concentrations or extreme pH values. Anyway, they fall outside the aim of this thesis and so they will not be considered here.

1.2.1 Interest and utility of extremophilic enzymes: the PCR example

Apart from the obvious academic interest and the intellectual curiosity in life forms that have inhospitable places, such as hot springs or brine pockets surrounded by sea ice, as their habitats, there are other good reasons to study them. The extremophiles indeed express enzymes that can be operative at "extreme" optimal temperatures. These enzymes are hence called *extremozymes*. They can be used for pharmaceutical, industrial and for research purposes in reactions at not standard temperatures. One of the possibly most significant example of such usage is the *Polimerase Chain Reaction* technique, commonly referred to with its acronym PCR.

PCR was developed in 1983 by Kery Mullis, although a very similar idea was firstly introduced in 1971 by the Norwegian Kjell Kleppe [22]. This method won

Mullis a Nobel Prize for Chemistry in 1993. It is thanks to this technique that scientists have been able to isolate and study the DNA of HIV and to analyse genetic materials from very scarce sources, like fossils or dried blood on crime scenes. The method itself is quite simple and it allows to replicate in a test tube a single DNA segment, obtaining billions of copies in a few hours [29]. This multiplication is made by an enzyme, the *DNA polymerase*, that copies a single strand of DNA when a *primer*, i.e. a small amount of complementary bases, is already attached to the starting point on the original strand. In order to have a single DNA strand the DNA's double helix has to be melted, which means that the hydrogen bonds connecting complementary bases have to be broken. This is obtained by heating the DNA at 94 –98 °C. Then there is the annealing step, which occurs at a lower temperature, where the primers bind with the initial part of the genetic code which needs to be copied. At this point everything is ready for the *DNA polimerase* to perform its task, doubling the initial amount of DNA. This cycle has to be repeated in order to have the desired number of DNA strands, usually something around 20-30 times, getting $\sim 2^{20-30} \simeq 10^6 - 10^9$ pieces of the same DNA.

Initially, they used the enzyme *DNA polymerase I* form *Escherichia Coli*, which works well at 30 °C but denatures at the high temperatures required to melt the DNA. This forced the researchers to add this enzyme at every cycle after denaturation and it was difficult to achieve an automated procedure. What Mullis himself said to be "*one of the most important improvements in the process*" [30] was the introduction of an enzyme extracted from another organism, the thermophilic *Thermus Acquaticus*, that has its peak of activity around 75 °C-80 °C [23]. The major advantage was to have an enzyme that could survive long incubation periods even at 95 °C, and so it only needed to be added at the beginning of the reaction, making it possible to be automated. What is more, it also improved the performance of this technique because it produced an increase in specificity, yield and sensitivity of the process [35]. Nowadays also other *DNA polymerases* are used [39]. An example is the one extracted from the termophilic organism *Pyrococcus Furiosus*, which is more accurate. In fact it has an error rate in the order of 10^{-6} and the copies contain less than 10% of the mutations caused by the *Thermus Acquaticus*'s enzyme [25].

1.2.2 The extremophiles and their evolution

Ascertained that it is worth studying the extremophiles, and in particular their proteins, it seems crucial to try to understand how they can survive in their *impossible*, from an anthropocentric view, habitats. Thermophiles have been able to adapt to high temperatures, so their enzymes are likely to have optimal temperatures at higher points than mesophilic ones. In fact, due to the little dimension of the thermophilic cells – in the scale of the micrometer – insulation from the hot environment appears impossible; therefore the cell components, including its proteins, have to be heat

resistant [40]. This is possible thanks to their firmer structure, which is less prone to disaggregate because of thermal agitation. Psychrophile, on the other hand, have evolved to survive even below 0 °C, but still the point of water freezing in cells remains a lower limit for life – with the exception of the nematode *Panagrolaimus davidi*, that can survive water freezing in its body. The littler thermal agitation in this case makes the contact between substrates and proteins less probable, causing the activity to be consistently lower then a mesophilic homologous. This is the reason why their enzymes have to be very efficient in order to produce enough product at life-compatible time rate.

So, how is it possible for extremozymes to function in those conditions? Since in this thesis we are going to compare thermophilic proteins to mesophilic ones, it will follow a discussion about the adaptations of organisms that thrives at high temperatures. There are various micro-organisms with this ability, and can be either bacteria – like photosynthetic bacteria, enterobacteria and thionic bacteria – or, as it happens for the majority of known thermophiles, archaebacteria – like *Pyrococcus* or *Thermoccocus*. According to Morozkina et al.[28], in 2010 more than 70 species, 29 genera, and 10 orders of thermophiles were known, but still the matter about the origins of these life forms is not totally resolved. On the one hand it seems that, from phylogenetic studies [1] , the thermophiles should have appeared at the time of the origin of life itself on Earth, preceding the mesophiles. On the other, some authors still prefer the hypothesis which sustain that the thermophiles descended from the mesophiles organisms, as a consequence of adaptation to high temperature [28]. It is anyway possible that both these things happened, and some thermophiles are the result of a mesophile colonizing – or recolonizing – an hot environment while others directly originated in the extreme habitat [3].

1.2.3 Extremozymes adapted to hot habitats

The problem of the extraordinary thermal stability of thermophiles has been the subject of studies since they were discovered. It was found that there are different ways they adapted themselves for living in their extreme habitats. For example, differences have been found between thermophilic archea and thermophilic bacteria in the type of membrane lipids [28].

With respect to the stability of the protein structures, a certain variability of adaptations has developed, potentially connected to the evolutionary history of

[1]The phylogenetic studies were carried on characterising some genes in the ribosomal RNA, to be precise the 16S, in prokaryotes, and the 18S, in eukaryotes, rRNA genes. The choice of this particular segment of the genetic material present in every cells, is due the fact that are essential components of the cell, and so one can find them in all self-replicating systems, and that their sequence changes slowly with time, allowing to relate very distant species [47].

the expressing organism. It is suggested that if the organism had originated in the hot environment it would prefer "structure-based" mechanisms of adaptations, presenting more compact and more hydrophobic proteins; if instead the organism had a mesophile ancestor it would use a "sequence-based" one, resulting in proteins with similar structures but with stronger interactions [3]. Both methods are made possible by many strategies, the most recurring are here listed.

Amino acid composition There are some more labile amino acids that are more likely to undergo modifications at high temperatures, and therefore their half lives are shorter in those conditions. In fact, applying temperatures beyond 100°C, the thermal stabilities of the common amino acids are (Val,Leu)> Ile> Tyr> Lys> His> Met> Thr> Ser> Trp> (Asp,Glu,Arg,Cys) [34]. As a consequence in thermophilic proteins, compared to mesophiles homologous, there are less residues that are particularly thermolabile, such as asparagine, cysteine, glutamic acid and aspartic acid [10]. What is more, some other changes in the amino acid composition may lead to a more robust protein. As an example, Van den Burg et al.[43] operating some "rigidifying" mutation – such as glycine and alanine been replaced, respectively, by alanine and glycine in delicate regions of the protein – to a relatively thermostable enzyme, managed to have a hyperstable one. Furthermore, it seems that charged residues are preferred [3], increasing the possibility of having a stronger ionic network.

Ionic networks The study of thermophilic protein structures indicates that ion pairs on the surface of subunits and domains may be important for their stability. In fact, networks of ionic interaction have a longer range than hydrophobic ones and do not depend on the alteration that water undergoes at high temperature. On hyperthermophilic proteins, more extensive ionic networks, spatially alternating of positive and negative charges, have been observed than on thermophilic or mesophilic counterparts [10] [34]. An interesting study by Vetriani et al.[44] showed how subtle changes in the amino acid sequence of a protein, made in order to reinforce its ionic network, result in major changes in its thermostability. They substituted only two bases in the hexameric glutamate dehydrogenases from *Thermococcus litoralis* and, although the single swings proved to have an adverse effect on thermostability, together they protracted of almost 4-times the half life of the protein at 104 °C.

Hydrophobic packing Notwithstanding the fact that many amino acids are more stable when they are inside a hydrophobic packing, it is still not clear what is the temperature role in the strength of hydrophobic interactions. It is not certain whether at higher temperatures they become stronger or weaker. In any case, it is a common feature of stable globular protein to have a closely packed hydrophobic core [34].

Cooperative association It has been observed that sometimes the structure of a thermophilic *oligomer*, i.e. a protein complex made of two or more subunits, can be more elaborated than its mesophilic homologous, that could even consists of just a monomer. This is due to the denaturation of monomers that follows the dissociation of the oligomer. Therefore, thanks to strong inter-molecular forces, the process of the unfolding is forestall [10] [34].

A compact structure Some thermophilic proteins present a more compact structure compared to mesophilic ones. A common alteration concerns the α-helixes to β-layers ratio [10]. Preferring for their secondary structure the more packed β-layer, extrmozymes have less cavities and a lowered area to volume ratio. Another strategy consist of preclude N– and C– termini's movements. Preventing the loose ends from fraying stabilizes the structure, limiting the chances of unraveling. This can be obtained in different ways, for example keeping the termini in hydrophobic pockets, substituting disordered loops with α-helix structures or using ion-pairing [34].

1.3 Studying the structure of a protein

It is now well clear how important the protein structure is: from the three-dimensional disposition of its residues descends a number of properties, like the very function of the protein and the temperature range in which the function is active. As a consequence, being able to understand and study the three-dimensional assemblage of the various proteins is crucial. So far, two main methods for experimentally obtaining them have been used:

- *X-ray crystallography*;
- *Nuclear Magnetic Resonance* (NMR) spectroscopy.

To these two, Bioinformatics is to be added, a newer approach that does not need a direct observation of the folded protein, but relies only on the predictions that softwares can make from the linear sequence of the amino acids. Even if the predictions can be quite accurate, the protein folding problem is still not resolved, so this last method is mainly used as a complementary approach in understanding protein structure.

All the protein structures examined in this thesis were obtained with X-ray crystallography. As a matter of fact, while the two methods usually model structures with the same backbone topology, they often produce different local features, like for example surface loops [48]. This is caused by the different samples used for these techniques. In the NMR spectroscopy case, highly purified proteins are dissolved in an aqueous solution, while for the X-ray crystallography, as the name suggests, the sample consists of a solid protein crystal. This is one of the reasons why X-ray

crystallography yields to higher resolution results, which is ultimately the cause of the choice of using only structures obtained this technique.

In the next section we are going to explain how protein X-ray crystallography works.

1.3.1 Protein X-ray crystallography

Determining the correct three-dimensional protein structure is not an easy task. The atoms in a single molecule are thousands, and finding the exact position of each of them is not simple. The first structures (of haemoglobin and a related protein) were obtained in 1959 with X-ray crystallography. This achievement was attained many years after this technique was born with the first pioneering studies on simple inorganic crystals by Max von Laue in 1912 and by Bragg father and son in the following years.

X-ray crystallography is possible tanks to the regularity of crystals and the wave proprieties of the electromagnetic radiation. An X-ray beam striking an electron makes it oscillate at the same frequency of the original beam. An oscillating electron emits a spherical wave with the same frequency as the incident one. This phenomenon is called elastic scattering. From the analysis of the scattered light wave it is possible to resolve the electron density of the crystal. This wave is "simply" the sum of all the scattered waves by all the electrons in the crystal. The periodicity of the crystal makes it possible to go from the scattered pattern to the electron density distribution.

For understanding what happens when there is a diffraction we follow the approach presented in *Principles of Protein X-Ray Crystallography*, by Jan Drenth [14]; so we start considering the scattering from a couple of electrons, the first one in the origin of our frame of reference and the other in $\vec{r}$. We consider an incident X-ray wave with wave vector $|\vec{k_0}| = 2\pi/\lambda$ and the diffracted light with wave vector $\vec{k}$, which has the same magnitude of the incident one thanks to the fact that the scattering is elastic. The amplitude A of the wave scattered from the first electron is the same of that from the second, but there is a difference in the phase. This is caused by a difference in the light path of $\vec{r} \cdot \vec{k_0}\,\lambda/(2\pi) - \vec{r} \cdot \vec{k}\,\lambda/(2\pi)$, which results in a phase difference of $\vec{r} \cdot (\vec{k} - \vec{k_0}) = \vec{r} \cdot \vec{q}$, where $|\vec{q}| = 2\sin\vartheta\,(2\pi)/\lambda$, with ϑ being the incident angle of the primary wave with an imaginary reflecting plane. This leads, thanks to the adding property of electromagnetic waves, to a total scattered wave of $A[1 + \exp(i\vec{r} \cdot \vec{q})]$. Supposing now to shift the origin of the system of $-\vec{R}$, so that the first electron is now in $\vec{R}$ and the second one is in $\vec{R} + \vec{r}$, following the previous line of reasoning, another phase displacement of $\exp(i\vec{R} \cdot \vec{q})$ has to be added to both the waves, resulting in a scattered wave with the same amplitude:

$$A \exp(i\vec{R} \cdot \vec{q})[1 + e^{i\vec{r} \cdot \vec{q}}]. \tag{1.1}$$

Let us now consider an atom and its electron cloud with $\varrho(\vec{r})$ density. We have now the origin of the reference frame in the nucleus, since a displacement would only result in a phase shift common to all electrons as just seen. The amount of scattered light depends on the number of electrons and their position in the cloud, so the atomic scattering factor f is defined as:

$$f(\vec{q}) = \int_V \varrho(\vec{r}) e^{i\vec{r}\cdot\vec{q}} d\vec{r}. \tag{1.2}$$

The atomic scattering factor can be looked up in tables, where f is expressed as a function of the module of $\vec{q}$, as the electron cloud of an atom is assumed spherically symmetric and so the direction of $\vec{q}$ is irrelevant. What is more, f, which is actually the Fourier transform of the electron density map, is always real thanks to this symmetry of the cloud. From the single atoms we now move on considering a unit cell. The scattering of the unit cell is nothing but the sum of the scattering from the single atoms that compose it. Since the nucleus of the j–th atom is not centred in the origin of the system, but it is now in $\vec{r}_j$, a phase angle of $\vec{r}\cdot\vec{q}$ must be added to the atomic scattering factor f_j. The structure factor $F(\vec{q})$ is then defined as:

$$F(\vec{q}) = \sum_{j=1}^{n} f_j e^{i\vec{r}_j\cdot\vec{q}}. \tag{1.3}$$

The last step consist in considering the whole crystal. Suppose that the crystal has n_1 cells in direction $\vec{a}$, n_2 in direction $\vec{b}$ and n_3 in direction $\vec{c}$, being $\vec{a}$, $\vec{b}$ and $\vec{c}$ the translation vectors of the crystal. Then the total scattering factor is:

$$K(\vec{q}) = F(\vec{q}) \times \sum_{s=0}^{n_1} e^{is\vec{a}\cdot\vec{q}} \times \sum_{u=0}^{n_2} e^{iu\vec{b}\cdot\vec{q}} \times \sum_{v=0}^{n_3} e^{iv\vec{c}\cdot\vec{q}}. \tag{1.4}$$

But since the number of cells in every direction is a very high one, $\sum_{s=0}^{n_1} \exp(is\vec{a}\cdot\vec{q})$ would be almost equal to zero, unless $\vec{a}\cdot\vec{q} = 2\pi h$, with h integer, and so on for the other translation vectors. Hence follow the Laue conditions:

$$\begin{cases} \vec{a}\cdot\vec{q} = 2\pi h, & h \in \mathbb{Z} \\ \vec{b}\cdot\vec{q} = 2\pi l, & l \in \mathbb{Z} \\ \vec{c}\cdot\vec{q} = 2\pi k, & k \in \mathbb{Z}. \end{cases} \tag{1.5}$$

From these equation it appears clear that it is crucial to have a good periodicity in the crystal in order to be able to analyse the data. Unfortunately obtaining a suitable single protein crystal is the least understood step in the whole process, so it is mainly a trial-and-error procedure that leads to the precipitation of the protein form its solution. The purity of the protein is surely an important factor. In fact, compound other then the protein itself should be absent and also all the molecules in the protein should present the same surface properties. The crystallization of protein is obtained in four steps:

1. The protein's purity has to be determined (e.g. with a mass spectrometry). If it is not satisfactory, further purification will be necessary.

2. The protein is dissolved in a solvent, which is usually a water-buffer solution. Membrane proteins that are insoluble in such solvent require the addition of a detergent.

3. The solution is then brought to supersaturation. During this step little aggregates are formed. Those aggregations will become the nuclei for the growth of the crystal. Precipitation of the protein is achieved in more then one way, like changing the pH or the temperature, or, one of the most common method, increasing the concentration of the protein by adding a salt (*salt-out*) or polyethyleneglycol (PEG) to the solution.

4. After the formation of the nuclei, the actual crystal growth begins. More molecules or other small nuclei get attached to the starting nuclei, forming the crystal. This step is quite critical. One of the reason is that, if the supersaturation is too high, too many nuclei are formed resulting in many small crystals. So it is important to have a lower level of supersaturation then in the previous step. What is more the optimal growth process is a very slow one to achieve a high order degree, so it would be better to not change the temperature as a way to modify the supersaturation.

Once a pure enough and big enough (usually $0.3\,\text{mm} \times 0.3\,\text{mm} \times 0.3\,\text{mm}$, approximately $15\,\mu\text{g}$) crystal is obtained, it is exposed to an X-ray diffraction trial. To do so, the crystal has to be mounted on an appropriate support. Usually two roads can be taken: putting it in a capillary test tube at or near room temperature or suspend it in a small loop in a stream of liquid nitrogen at a temperature range of $100\,\text{K}$ to $120\,\text{K}$. In the first case the crystal is pushed carefully in an air gap between two layers of mother liquor, i.e. the solvent used to precipitate the proteins. It is very important to not change the environment of the crystal, because the spherical or egg-shaped macromolecules are loosely packed in the crystal, surrounded by the solvent that fills the gaps between them, so any loss of the mother liquor destabilizes the crystal. The second option has the advantage of the low temperature in maintaining the crystal structure. In fact, the very exposure to the X-ray radiation damages the crystal and the diffraction pattern dies after a few hours at room temperature. Nevertheless, the cooling of the crystal has to be treated very carefully. It has to occur suddenly, hence the names of the technique *flash freezing* or *shock cooling*, because the mother liquor in the crystal must freeze to a vitreous substance and not crystallize, otherwise ice crystals would damage the protein crystal structure.

The diffraction pattern is then acquired. Different X-ray sources and detectors can be used for this goal. Very briefly, as this could lead us far form the aim of

describing protein crystallography, it is possible to obtain an X-ray beam from three different sources:

Sealed or rotating anode tubes Both of these tubes have at their base the same functioning idea: a cathode emits an electron beam, the electrons accelerate towards the anode and collide with it at high speed. Most of the energy of the impact is converted to heat, that has to be removed by cooling the anode, usually with water, in the case of the sealed tube, or by rotating it, as the name suggests this is the case of a rotating anode tube, in order to change the place of impact and give it time to cool down before being hit again. However, a part of the energy is emitted as X-rays thanks to the interaction between the electrons and the anode material. The spectrum of the X-ray emitted presents a smooth continuous region, called the *"Bremsstrahlung" radiation*, that is radiation emitted due to the deceleration of the charged particles, electrons in this case, and sharp peaks called *characteristic radiation*, emitted by electrons of outer shells lowering their energy level in order to take the place of K-shell electrons ejected after a collision with a fast moving electron coming from the cathode.

Particle accelerators Synchrotron radiation is obtained from a particle accelerator, a big and expensive facility. Charged particles, as electrons or positrons, are accelerated and injected into the storage ring where they circulate. Every time the particle beam changes direction, having its path bended by a magnet, it accelerates and some radiation is emitted. This lost in energy is compensated for with radiofrequency input at every cycle, and when the first synchrotrons were built it was considered an annoying wast of energy, being their primary objective to obtain high energy particles' collisions. Anyhow nowadays synchrotron radiation is very interesting for various applications and experiments since the quality of the X-ray beam is higher than that of a ordinary X-ray tube. The intensity of the synchrotron light beam obtained is up to two orders higher, the beam has a low divergence, a monochromator can select any suitable wavelength in the spectral range and the light is polarized.

To collect the diffracted light, single photon counters were used since the early years of X-ray crystallography, but they, although very precise, have the big disadvantages that it takes several weeks to obtain a complete data set for one protein. Photographic film was a good alternative, with a resolution superior than modern day detectors, but the process of developing the film is quite time consuming as well and it has a limited dynamic range, making necessary to use three consecutive films when the full X-ray intensities' range was needed. A first alternative is using a image plate. Image plates have a higher sensibility and dynamic range than photographic films. Image plates have to be read after the exposure to the X-rays, since they retain energy, proportional to the number of photons that hit that area, that can be released on

illumination with light. More convenient are area detectors that can process the signal immediately after the detection. They can be gas-filled ionization chambers or semiconductor detectors.

Once the diffraction patterns, for various crystal orientations, are collected, it is time to do some data analysis and convert it into an electron density map. The data set usually consist of hundreds of single two-dimensional images that have to be merged, that is identifying the same peaks in different images, and scaled in order to have the same intensity scale. The data is a representation in the reciprocal space - the reciprocal lattice is the Fourier transform of the crystal lattice written as a Bravais lattice - of the crystal lattice. It is easy to see that the total scattering factor, which is proportional to the the structure factor from equation 1.3 with the Laue conditions 1.5, is actually the Fourier transform of the electron density writing it as an integration on the all cell volume:

$$\begin{aligned} F(\vec{q}) &= \int_{cell} \varrho(\vec{r}) e^{i\vec{r}\cdot\vec{q}} dV \\ &= V \int_{x=0}^{1}\int_{y=0}^{1}\int_{z=0}^{1} \varrho(x,y,z) e^{i\left[\vec{a}x+\vec{b}y+\vec{c}z\right]\cdot\vec{q}} dxdydz \\ &= V \int_{x=0}^{1}\int_{y=0}^{1}\int_{z=0}^{1} \varrho(x,y,z) e^{i2\pi[hx+ly+kz]} dxdydz, \qquad \text{for } h,l,k \in \mathbb{Z} \\ &= F(h,l,k) = |F(h,l,k)|\, e^{i\alpha(h,l,k)}, \qquad \text{for } h,l,k \in \mathbb{Z}. \end{aligned}$$

From this results that the electron density map can be obtained with the inverse transform, i.e. the Fourier transform of the structure factor:

$$\varrho(x,y,z) = \frac{1}{V} \sum_h \sum_l \sum_k |F(h,l,k)|\, e^{i\alpha(h,l,k)} e^{-i2\pi[hx+ly+kz]}. \tag{1.6}$$

The integration has been replaced by a summation because of the Laue conditions. The reason why $F(h,l,k)$ has been explicity separated into its module and its phase it is going to be clear soon, and it has to do with what we detect from a crystallographic experiment. In fact, the data collected from a crystallographic experiment is basically the intensity of the scattered waves, that can be written as:

$$I(h,l,k) = (AS)^2 \left|F(h,l,k)\right|^2, \tag{1.7}$$

where S is a proportionality constant. This now makes clear that if $|F(h,l,k)|$ can be deduced from the experiment, the phase $\exp[i\alpha(h,l,k)]$ can not, leading to what is called the *phase problem*. Fortunately several methods have been developed to have an initial guess for the phase and then iteratively perfect it while maximizing the correlation between the diffraction data and the model obtained.

There are some parameter that one can calculate from the obtained model and the diffraction pattern that can quantify the quality of the model itself. The most important ones are here listed and described, citing as typical values the ones indicated in the relevant sections of the online portal PDB-101, at the website rcsb.org [4].

Resolution The *resolution* of an electron density map depends on various things and is a first parameter that shows the quality of the data collected. As a matter of fact, resolution tells how much detail is present in the diffraction pattern and, as a consequence, in the final model. This depend on the experimental equipment used, on the conditions of the experiment, e.g. the temperature, and on the purity and the order of the crystal. A structure that presents all its fine details, showing all its atoms on the electron density map, is said to be at *high-resolution*, having little resolution values around 1 Å. On the contrary, for *low-resolution* structures, that have resolution values of 3 Å of higher, only the basic skeleton of the protein can be seen and the atom position is inferred. The majority of the structures have a resolution in between those two peaks.

Temperature factors or B-factors If the resolution is a characteristic of the whole model, more detail on the accurate positioning of the single atoms can be obtained with the *temperature factors*, also called *B-factors*. These values give an insight on the displacement of an atom from its mean position. The more the atom moves from its average, i.e. the more flexible it is in the protein crystal, the higher the associated B-factor. The typical values for the temperature factors are around 15-30 $Å^2$, with possible peaks way larger than 30 $Å^2$ for very flexible regions. Such parts of a protein can be shown on a 3D chart representing each atom in its coordinates and adding a red color to the ones with the highest B-factors and a blue color to the ones with the lowest, B-factors in between represented by an appropriate blend that allows to see how close are to one of the extremes.

R-value and R-free value The *R-value* carries another important information about the model obtained. It is a measure of how well the diffraction pattern simulated from an hypothetical crystal made with the calculated protein model compares to the experimental diffraction pattern. This provides a quantitative information about the quality of the reconstruction carried out from the empirical data. On the one hand if the two patterns overlap perfectly the R-value is 0, while on the other if the atoms of the proteins are positioned randomly the R-value is close to 0.63. These of course are extremes, whereas typical values of real cases are about 0.2. Sometimes the R-value can be utilized as a feedback for the reconstruction algorithm in order to refine the atomic model. Therefore the R-value of the final model can present a bias, for this reason one can look at the *R-free value*, an R-value obtained

on a 10% of the experimental data that is not used to optimise the model. A good model should have its R-value similar to the R-free value, but in any case the R-free is usually higher, with typical values around 0.25.

Once the model is refined and the obtained electron density map is satisfactory, it is usually deposited with all the additional information needed in a crystallographic database, such as the Protein Data Bank.

1.3.2 The Protein Data Bank and its files

The Protein Data Bank (PDB) [4] archive was first announced in 1971 [31]. Before the PDB was established, punched cards, one for each atom, were the only method to exchange protein structures' coordinates; this reciprocity was active between only a few research laboratories. The establishment of the PDB, a joint operation by the Brookhaven National Laboratory and Cambridge Crystallographic Data Centre, made the data exchange possible for anyone. In 1999, the management moved from Brookhaven to the consortium called Research Collaboratory of Structural Bioinformatics (RCSB).

At present the PDB site, URL www.rcsb.org, provides the user with 116085 biological macromolecular structures and a number of tools for search the database, visualize entries and analyse them. At the moment, in the database there are more 100000 protein structures, nearly 3000 nucleic acid ones and more than 5000 protein and nucleic acid complexes. Proteins structures were mostly obtained with X-ray crystallography, only less than about 1/10 of them were obtain with NMR spectroscopy.

Each entry can be downloaded as a text file that presents the ".pdb" extension. The file contains various information about the protein and its structure, more then just its atoms' coordinates. Therefore it is divided into sections. The following is a list of the sections you can find in a PDB file, even if not every one has all of them.

- Title; it presents the main descriptive records.
- Remark; where various comments are inserted.
- Primary structure; the peptide or nucleotide linear sequence is presented.
- Heterogen; in this section non-standard groups, i.e. groups that are not part of the polymer as described in the primary structure section or are unknown/non-standard amino/nucleic acids, are described.
- Secondary structure; here the helix and β-sheet are described.
- Connectivity annotation; the existence and location of disulfide bonds or other linkages are stated in this section.

- Miscellaneous features; it may describe proprieties in the macromolecule.
- Crystallographic and coordinate transformation; it describes the crystallographic cell, the geometry of the crystallography, and the coordinate transformation operators.
- Coordinate; it collects the atomic coordinate data.
- Bookkeeping; final information.

The begin of every line in the file is a six letter word that defines the record stored in that line. Every file starts with the title section. The first record is the "`HEADER`" one, that contains the classification of the molecule in the file, the deposition date and the IDcode, an unique identifier of that entry within the PDB. Other information about that set of coordinates, of the file and of the molecule follows, e.g. if the entry is obsolete and has been substituted with new entries or the method used to resolve that structure. The lines with the record name "`ATOM  `", in the coordinate section, present the atomic (x, y, z) coordinates in Angstroms for every standard amino acid (or nucleotide). This lines store the most important information about the structure of the protein.

For a complete information about the structure of a pdb file, it is possible to refer to the available on-line documentation (http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html).

With the data in the PDB archive you can perform a number of inquiries on the structure of the protein. This is also simplified thanks to the Python library BioPython [20]. It downloads the PDB file matching the provided PDB ID and from this it retrieves, with simple Python commands, information without the user having to search for it "manually" in all its lines.

In this thesis the three dimensional structures of the thermophilic and mesophilic proteins in the database, which will be described in chapter 3, have been reduced to 2D matrices, the Protein Contact Maps, as detailed in chapter 2. These are then analysed with a network approach and a series of measurements on them have been executed. The result of this analysis are presented in chapter 4.

2

Methods of analysis

In this chapter the methods used to analyse the proteins are presented. In particular, after some background about graph theory, it is discussed what a Protein Contact Map is and how it is possible to retrieve information about the dynamic of the protein from that network. The elements of graph theory are mainly taken from the book *Graph Theory*, by Reinhard Diestel [12].

2.1 Elements of graph theory

A protein contact map is a useful tool and it allows to look at the structure of a protein as a network. For this reason, we first briefly introduce the definition of network (or graph) and its mathematical description.

Graphs, as mathematical objects, made their first appearance in 1736 when Euler used them to solve the seven bridges of Könisberg problem [9]. The river Pregel was crossing the city, forming two big islands, which were connected to each other and the two sides of the city by seven bridges. The problem consisted in planning a walk through the city that would use each bridge once and only once. Euler imagined the problem as a graph, where the mainlands and the island were vertices and the bridges were links to and forth those vertices, whose schematic representation can be seen in picture 2.1. It was the first time a problem was formalised in terms of nodes and links connecting the nodes, which are the two main components of every network, as illustrated in the next section.

2.1.1 Definition of graph

Euler for solving his problem used two different groups of elements, the links and the nodes, and indeed a graph $G = (V, E)$ is defined as a couple of sets, one

containing the *vertices*, or nodes or points, and the other the *edges*, or links or lines, that connect them:

$$V = \{v_1, v_2, ...\} \tag{2.1}$$

$$E \subseteq [V]^2 = \left\{e_{(1,2)} = (v_1, v_2),\, e_{(1,3)} = (v_1, v_3), ...\right\} \tag{2.2}$$

As a convention, the set of nodes of a graph is referred to as $V(G)$ and the set of links as $E(G)$, whatever the actual names of the subsets, and the set of the edges that have one end in the vertex v_i is denoted by $E(v_i)$. A graph is said to be a *multigraph* if multiple links between the same two vertices are allowed, as in the case of the Könisberg problem. A link from and to the same node, i.e. of the type $e_{(i,i)} = (v_i, v_i)$, is called a *loop*. The *order* $|G|$, or equivalently $|V|$, of a graph is defined as the number of vertices in it; even though it is possible to have graphs with infinite number of nodes in this discussion only *finite* graphs will be considered, since proteins have a finite number of elements.

From this broad definition one can devise many different kinds of graphs. One classification separates undirect from direct graphs. A graph is said to be *undirect* when the order of the vertices in the edges is immaterial, namely $e_{(i,j)} = (v_i, v_j) = (v_j, v_i) = e_{(j,i)}$, so if the vertex v_i is connected to v_j then v_j is connected with v_i. Conversely, a graph is called *direct* when the edges have a directionality and the connection between v_i and v_j does not imply a link between v_j and v_i; "one-way streets" between nodes are allowed in this case.

Another dichotomy can then be seen between weighted and unweighted networks. *Weighted* networks $G = (V, E, W)$ have an additional information: a set of "weights" or "costs" $W = \left\{w_{(i,j)} \neq 0 \text{ if } (v_i, v_j) \in E\right\}$, one for each edge. *Unweighted* ones have no such set or can be seen as a special case of weighted graphs where all the $w_{(i,j)} = 1$ if $(v_i, v_j) \in E$ and $w_{(i,j)} = 0$ otherwise.

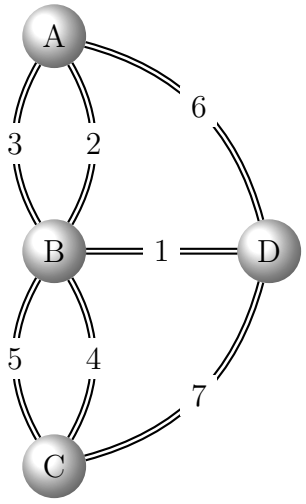


Figure 2.1: The schematic representation of the Könisberg's bridges problem.

Representing a graph

Graphs can be represented in many ways. They can be drawn as points (the vertices) connected with lines (the edges). An example of this representation can be seen in picture 2.1, where the nodes are labelled with capital letters and the links with numbers. Alternatively, using a more mathematical approach, a graph can be seen as a matrix $\mathbf{A}$ called the *adjacency matrix*, that is unique for each graph, given an ordering choice of the

nodes. The general element $a_{i,j}$ of the adjacency matrix is defined as:

$$a_{i,j} = \begin{cases} w_{(i,j)} & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

From the definition it is clear that only graphs with loops have nonzero diagonal elements and that undirect graphs result in a symmetric matrix.

As an example, using the alphabetical order shown in picture 2.1, the same graph of the bridges of Könisberg could be alternatively represented as a matrix:

$$A = \begin{bmatrix} 0 & 2 & 0 & 1 \\ 2 & 0 & 2 & 1 \\ 0 & 2 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}.$$

Since this graph is a multigraph, the element $a_{i,j}$ of the matrix is weighted with the number of bridges linking the nodes v_i and v_j.

2.1.2 Main proprieties of a graph and its components

After having introduced what a graph is and how to represent it, it is important to list some of the proprieties of a graph and its components. Those proprieties characterise each graph and are often used to analyse them.

The *neighbourhood* of a vertex v_i is the set $N(v) = \{v_j \in V(G) | (v_i, v_j) \in E(G)\}$. A node v_j is said to be a *neighbour* of or *adjacent* to v_i, with $i \neq j$, if it is in the neighbourhood of v_i. Conversely, v_i and v_j are said to be *independent* if there is no edge between them.

The *degree* $|E(v_i)| = d(v_i) = k_i$ of a node v_i is defined as:

$$d(v_i) = \sum_{j=1}^{N} a_{i,j},$$

that is the number of its neighbours if the graph is unweighted, or the sum of the weights associated with the edges that have an end in v_i. A vertex whose degree is null is called *isolated*. A graph whose nodes have all the same degree k is called *regular* or *k-regular*. From the definition of the node degree, a number of proprieties of a graph follows, like the *minimum degree* of a graph G $\delta(G) = \min\{k_i | v_i \in V(G)\}$, or the *maximum degree* $\Delta(G) = \max\{k_i | v_i \in V(G)\}$ and the *average degree* of G $d(G) = \frac{1}{|G|} \sum_i k_i$.

Now the definition of path will be introduced, which has always been central in network analysis since its beginning. In the case of a network of streets or transport

routes the idea of describing paths comes quite natural. Mathematically speaking, a *path* is a non-empty graph P, whose sets of vertices and edges are of the kind of:

$$V(P) = \{v_0, v_1, ..., v_k\} \tag{2.4}$$
$$E(P) = \{(v_0, v_1), (v_1, v_2), ..., (v_{k-1}, v_k)\}, \tag{2.5}$$

where the v_i are all different nodes. Often paths are noted simply as a sequence of vertices, implying a link between consecutive ones, like for example $P = v_0\, v_1\, ...\, v_k$. The nodes v_0 and v_k at the extremity of the path are said the *endvertices* or *ends* of P, while the others $v_1, ..., v_{k-1}$ are called the *inner* vertices of P. Two paths are said to be *independent* if they do not contain any inner vertices of the other. The *length* of a path P is the number of edges in $E(P)$. A path P of length k can be written as P^k. In the case of a weighted graph, the length or *cost* of a path P is the sum of the weights of the edges in $E(P)$. A *cycle* is a path $P = v_0\, ...\, v_{k-1}$ with at least two nodes at witch is added a link between the last vertex and the first, so a cycle can be written as $C = v_0\, v_1\, ...\, v_{k-1}\, v_0$.

A next important propriety of a graph, tightly bind to the definition of path, is its connectivity. A graph G is said to be *connected* if you can move from any vertex to any other through a path in G. When it is impossible to link any two nodes with a path in G, the graph is called *disconnected*. A connected graph that does not contain any cycle is called a *tree*. A graph whose components are trees is called a *forest*.

2.1.3 The New Science of Networks

To conclude this overview of graph theory, it is worth underlining that the field has seen major developments since its beginnings in the 18th century, particularly in the last decades. As a matter of fact, during the late 90s, thanks to technology breakthroughs and growing popularity of computers, it was possible to apply it to the study of bigger and bigger databases of network structures. This new possibility made it convenient to think (or re-think where possible) of complex systems as networks. As a matter of fact, with this approach considerable progress has been made by mathematicians, biologists and social science researches, who, even if coming from different backgrounds, all introduced new ideas and added new results to the graph theory. The blossoming of the graph theory in different branches of science has been called the "new science of networks" [46].

As just stated, one area where this network approach has given many interesting results is biology. Biological networks can be of various kinds: can be logical, as in the case of protein-protein or gene-protein interactions, or they can represent a physical network, as in the case of the nervous system or of a protein [9]. A more detailed discussion about how to analyse proteins as networks, more specifically as Protein Contact Maps, will be carried out in the next section.

2.2 Protein Contact Map

As described in section 1.3, the structure of a protein can be obtained with experimental methods and illustrated as a set of coordinates for each atom of the residues in the protein. This provides a complete information, at the given resolution, about the structure of the protein in the particular conditions in which it has been crystallized. The idea behind a *Protein Contact Map* (PCM) is to collapse the three-dimensional structure into a two-dimensional matrix, transforming the spatial information of where each residue is into a network of contacts. This simplification allows to describe the protein structure with less numerical values while retaining useful information about its relevant properties. How to devise a PCM is described in the following sections.

2.2.1 Definition of Protein Contact Map

So what is exactly a Protein Contact Map? A PCM is an undirect graph with no loops, defined as the adjacency matrix of a network where the vertices are the residues in the protein and the edges are the contacts between them, therefore its generic element $a_{i,j}$ is defined as:

$$a_{i,j} = \begin{cases} 1 & \text{if the } i\text{-th residue and the } j\text{-th one are in contact} \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

Since there is a natural order of the residues in the protein, given by the primary structure, the adjacency matrix of this network is well defined. The PCM of a protein represents, as the name suggest, the physical network of contacts between the protein's components, allowing the possibility to have a clear physical interpretation for most of the network properties that a PCM exhibits.

This definition, anyway, is quite broad and it allows different matrices to be created for the same protein, depending on the definition of contact and on what spatial point is chosen to represent the residue. Some of the most common possibilities for these various cases are discussed below.

2.2.2 Defining the distance matrix

In order to obtain a PCM, the first thing is to construct the *distance matrix*, whose generic element is:

$$d_{i,j} = d(i,j), \tag{2.7}$$

where $d(i,j)$ is the distance between the i-th and the j-th residue. Usually the distance is the Euclidean distance:

$$d(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{2.8}$$

between $\vec{r}_i = (x_i, y_i, z_i)$ and $\vec{r}_j = (x_j, y_j, z_j)$, which are respectively the position of the i-th and the j-th residue. Accordingly, the first decision to make is in what point $\vec{r}$ to collapse the whole residue, that is composed by many atoms – a brief description of how is made a residue can be found in section 1.1.1 at page 2 – and therefore there is no straightforward way to identify only three spatial coordinates to describe its position. There are instead many possibilities, from the most intuitive to the most sophisticated, and one has to make a choice. Some of the most common ones are presented here below.

α-carbons and β-carbons Usually the residue is represented by the coordinates of a carbon in the amino acid. The most standard choice is to use the coordinates of the α-carbons. Even if the carbon-alfa is a backbone atom which is present in every residues and therefore it seems an obvious choice, the protein structure appears to be better approximated by the β-carbons [15]. This means that the positions of the residues are placed on the coordinates of the carbon-beta for every residue with the exception of Glycine that does not have a β-carbons, in this case the α-carbon's coordinates are used.

Center of mass Another possibility is to utilize the coordinates of the center of mass of the residue. In this way the position of all the atoms in the residue is taken into account, giving a weighted average as the location of the residue. In fact, the center of mass is defined as:

$$\vec{M} = \frac{\sum_i m_i \cdot \vec{r_i}}{\sum_i m_i}, \tag{2.9}$$

where the sum is done over all the heavy atoms, i.e. all the atoms except for the hydrogen, of the residue, m_i is their mass and $\vec{r_i}$ their position.

All atoms Still another idea could be not to choose a fix coordinate for the residue, but, when computing the distance from another residue, to utilise the positions of the two closest atoms. In this way one calculates the shortest distance between any two atoms of the residues. A common way to refine this idea is to take into account the Van der Waals radius of each atom. The *Van der Waals radius* of an atom is also referred to as the atomic volume; in fact it represents the minimum distance from the nucleus at which another atom can come close to it, like the radius of an imaginary sphere that could be seen as the volume of the atom. The distance between two residues in this case would be calculated as the distance between the closest atoms minus the Van der Waals radii of those atmos. In this way the "dimensions" of the atoms are taken into consideration and the distance can be considered more precise. What is more, another variation is to consider the contribution from all atoms except the backbone ones, since they have been seen to provide a less specific information compared with the side chain atoms [5].

The decision of what to use as the residues' positions can change the aspect of the final contact map, since it influences the distances between residues. As it will be described in the next section, the distances between residues are crucial to define whether two residues are in contact or not and hence to define the contact map of the protein.

2.2.3 Defining the contacts

One of the easiest way to tell whether two residues i and j are in contact is to consider a spatial interval $\mathcal{I}$, said *cut-off*, and if the distance $d(i, j) \in \mathcal{I}$ then there is a link.

The choice of the cut-off interval $\mathcal{I}$ is clearly decisive to define the interactions included in the Contact Map and it should be pondered also on the chosen way to calculate the distance between residues. In literature one can find many different choices of this interval, the most common one is to have an upper limit for distances computed between α-carbons or β-carbons at around 8 Å, but other common options are 6 Å or 12 Å. For the all atom contacts, an upper threshold of 4.5 Å is usually set [21].

It is also possible to set a lower limit for $\mathcal{I}$ higher than 0 Å, excluding from the contact matrix the links between some very close residues. This option is described in the next paragraph.

Removing the protein backbone

Some authors choose to have not only an upper threshold, but also a lower one. This is positioned at the average length of the peptide bond [11], that is ~ 4 Å for the Cα. The idea behind this choice is to eliminate the trivial contacts that are present due to the primary structure of the protein. This is very useful when the study is focusing on contacts that are more sensitive on environmental stimuli.

As a matter of fact, removing the backbone links is a strategy that is used even after having defined the cut-off with only the upper limit. Usually in this case some diagonals are erased from the Contact Map. The typical number of diagonals whose content is all put to 0, so the existing contacts are ignored, is between 3 and 5 [21][24]. A more refined possibility, instead of deciding a fixed number of diagonals with no contacts, is to exploit the network proprieties of the Contact Map, in particular the one of connection; the contacts are deleted from as many diagonals as one less than number that would make the graph disconnect. In this way the Contact Map's propriety of being connected is preserved and the number of diagonals to cancel depends on the protein considered and not on an *a priori* choice.

2.3 The Laplacian matrix and its spectrum

A number of interesting analysis can be carried out from the contact map of a protein, but in this work a particular focus was given to the Laplacian spectra of the networks. This choice has been made because the Laplacian spectrum is closely related to the concept of vibrations.

2.3.1 Defining the Laplacian of a graph

The Laplacian, also known as the *Kirchoff matrix*, of a graph G is a positive semi-definite matrix defined as:

$$L(G) = D(G) - A(G), \tag{2.10}$$

where $D(G) = diag(d(v)),\ v \in V(G)$ is the *degree* matrix, i.e. a matrix whose elements are non null only on the main diagonal where a generic $d_{i\,i}$ is the degree of the i-th vertex, and $A(G)$ is the adjacency matrix, where $a_{i,j} = 1$ if v_i is adjacent to v_j and $a_{i,j} = 0$ otherwise. From this definition it is clear that the elements of L are of the kind:

$$l_{i\,j} = \begin{cases} deg(i), & \text{if } i == j, \\ -1, & \text{if } i \neq j \text{ and } i \text{ is connected to } j, \\ 0, & \text{otherwise.} \end{cases} \tag{2.11}$$

The Laplacian spectrum

The set of the eigenvalues of $L(G)$ is the Laplacian spectrum of G. They are n, as the number of nodes in G, and all non-negative, so they are usually arranged from the smallest to the highest:

$$0 = \lambda_0 \leq \lambda_1 \leq ... \leq \lambda_{n-1} \leq \Lambda.$$

The first eigenvalue λ_1 is the trivial solution and it is equal to 0. The λ_{n-1}, the highest eigenvalue, is limited by $\Lambda = 2 \cdot \max_i (deg(i))$.

The second eigenvalue of the Laplacian, λ_2, is called the *Fiedler value*, or *Fiedler eigenvalue*, and its related eigenvector $\vec{v}_2$ is known as *Fiedler vector*. The Fiedler value, also called *algebraic connectivity*, gives an indication of how well connected a graph is and it is 0 if the graph is not connected; in fact the number of zeros in the spectrum is the number of connected components in the graph. What is more, the Fiedler vector is often use to obtaining partitions of the graph, as described in the next section. In the next paragraph instead it will be defined a normalised Laplacian.

The normalised Laplacian

As just seen, the Laplacian eigenvalues have an upper bound that depends on the maximum degree of the nodes of the graph. It is possible to solve this by defining the *normalized Laplacian* $\mathcal{L}$ as:

$$\mathcal{L} = \begin{cases} 1, & \text{if } i == j, \\ -\frac{1}{\sqrt{deg(i) \cdot deg(j)}}, & \text{if } i \neq j \text{ and } i \text{ is connected to } j, \\ 0, & \text{otherwise} \end{cases} \tag{2.12}$$

from which it follows that $\mathcal{L} = D^{-1/2} \cdot L \cdot D^{-1/2} = I - D^{-1/2} \cdot A \cdot D^{-1/2}$.

$\mathcal{L}$ have its eigenvalues in the interval [0,2] [8], eliminating the dependence with the nodes' degrees. This is particularly useful when comparing eigenvalues of networks with different orders [6]. The spectrum of the normalised Laplacian results as a version of the Laplacian spectrum shifted by k^{-1} if the graph is k-regular.

An interesting propriety of the normalised Laplacian is that its spectrum is stricly connected with the proprieties of random walks on the network. As a matter of fact, the more closely its eigenvalues are packed around 1, the faster a random walk converges to the stationary distribution.

2.3.2 Spectral Clustering

Another use of the spectrum of a graph Laplacian is to separate the graph into partitions. This is obtained with the Fiedler vector, that will be indicated as $\vec{v}_1$, with a process known as *Spectral Clustering*. The components of the Fiedler vector are one for each node in the network, allowing to divide them into two groups: one made of vertices whose Fiedler vector component is positive and another one with the nodes that have an associated component negative. This process can be iterated on the two subnetworks until the components of the new Fiedler vector are all with the same sign [11].

The nodes can also be classified as "strongly" or "weakly" part of a cluster, depending on how far its component of $\vec{v}_1$ is from 0 [16]. Nodes that have values equal to 0, or very close, can be considered as bridges between two clusters, that when removed further isolate the two partitions.

This is not the only kind of information that the Laplacian spectrum can provide; it is also connected with the vibrational proprieties of the network as described in the next section.

2.3.3 Vibrations and the Laplacian spectrum

The Laplacian spectrum of a graph is not only related with its connectivity, but also with its vibrational proprieties. This can be clearly seen when we imagine the

network as a set of balls, representing the vertices, connected with springs, the links, that all have the same elastic constant k, immerse into a thermal bath of temperature T. The probability distribution of $\vec{x}$, whose i-th component is the displacements x_i of the i-th node from its equilibrium position, can be express with a Boltzmann distribution:

$$P(\vec{x}) = \frac{e^{-\beta V(\vec{x})}}{Z} = \frac{\exp(-\beta k/2\ \vec{x}^T \mathbf{L} \vec{x})}{\int d\vec{x} \exp(-\beta k/2\ \vec{x}^T \mathbf{L} \vec{x})}, \tag{2.13}$$

where $V(\vec{x})$ is the system's potential vibrational energy, Z is the partition function of the network, $\beta = (k_B T)^{-1}$ is the inverse temperature and $\mathbf{L}$ is the Laplacian matrix of the graph.

From expression (2.13), it is possible to compute $\Delta x_i = \sqrt{\int x_i^2 P(\vec{x})}$, the mean displacement of the i-th node. Estrada and Hatano in *A vibrational approach to node centrality and vulnerability in complex networks* [16] made this calculation obtaining the result:

$$\Delta x_i = \sqrt{\frac{1}{\beta k} \sum_{j=2}^{n} \frac{u_{ij}^2}{\lambda_j}} = \sqrt{\frac{l_{ii}^+}{\beta k}}, \tag{2.14}$$

where u_{ij} is an element of $\mathbf{U}$, a matrix whose columns are the Laplacian eigenvectors, and l_{ii} of $\mathbf{L}^+$, the Moore–Penrose pseudo-inverse of the graph Laplacian. The authors proposed to use this value as a measure of node vulnerability, since the higher Δx_i the higher the i-th node is affected by the surrounding environment.

3

The creation of the database

To analyse the structure of the thermophilic and mesophilic proteins with a contact map approach, the construction of an appropriate database was necessary. The aim was to have a database containing couples of proteins, from a thermophilic and mesophilic organism respectively, that share a common ancestor and are still similar to each other. In the next sections it is explained how the database was created, from the early choice of which couples to take into account to how it was refined with respect to the desired characteristics, including a discussion about the quality of the obtained database.

3.1 Creation of the database

To choose the couples to include in the database, searches in literature were conducted. The aim was to find analyses of homologous proteins, one thermophilic and the other mesophilic, in order to add the couples used to the list of possible entries of the database. Requiring homology and not only analogy, i.e. requiring that the proteins have developed from a common ancestor and not only that they carry out the same function (but each with its independent evolution history), is key for this analysis. As a matter of fact, the final objective is to possibly find structural differences between the two protein families, with a special focus on differences due to thermal adaptation. Analysing two homologous proteins therefore augments the likelihood that the discrepancy in their structures could be caused by the different temperatures at which the organisms evolved, and not by other reasons or by random neutral mutations. Accordingly, it was also required that the homologous couples has similar enough sequences and structures, to take into account proteins that have a common ancestor in their recent history, maximizing the chances that their differences are only due to thermal adaptation. A discussion on what "similar enough"

means will be carried out in section 3.2.2, measuring the difference in length of the primary structures and the coefficients Identity, MaxSub and TM-score of the two proteins. This section, instead, will focus on the results of the literature review looking for homologous couples.

3.1.1 A database by Glyakina et al.

The paper "*Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms*" by Glyakina et al. [19] presents a wide database of already well aligned thermophilic and mesophilic proteins. In figure 3.1 it is shown an example of a couple of aligned proteins taken from their list. In orange there is the C-phycocyanin protein from the thermophilic organism *Thermosynechococcus vulcanus*, while in cyan the same protein form the mesophilic *Synechococcus elongatus*. This pair of protein is really similar and the good alignment can be appreciated also from the figure. The database contains more than 300 couples of very akin protein chains that have been selected as follows.

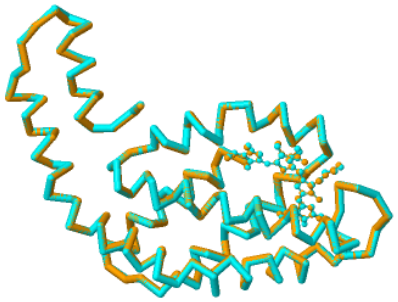


Figure 3.1: Example of two aligned proteins in Glyakina database: two C-phycocyanin proteins, with PDB ID 1KTP for the thermophilic and 1JBO for the mesophilic.

Firstly they clustered all the proteins in the PDB archive available in September 2005 into homologous groups, paying attention to the organism they belonged to in order to discern whether they were thermophilic or mesophilic. Then, they structurally aligned all the proteins in each cluster and calculated the *MaxSub*, whose definition follows in equation 3.1, value between all the thermophilic and mesophilic ones. *MaxSub* value was introduced in [38] as a method for evaluating structure prediction models and it is defined as:

$$MaxSub = \frac{\sum_i \left(1 + (d_i/d_0)^2\right)^{-1}}{\min(N_1, N_2)}, \tag{3.1}$$

where the sum is on every i-th residue, d_i is the distance between the i-th residues in the two compared proteins after alignment, d_0 the distance threshold parameter that was set to 3.5 Å, in accordance with the authors that proposed the *MaxSub* score, and N_1, N_2 are the numbers of residues of the two proteins. At this point, from each cluster they chose the couple with the highest *MaxSub*. Finally, they eliminated from the database all the pairs that presented a *MaxSub* value below 70%.

With this procedure they obtained a database containing 396 alignments. Out of these, 74 aligned intervals referred to the same chains of the same proteins. With this correction, the database in the end contains 322 pairs of highly similar proteins chain.

Some years have passed from September 2005 when the PDB was scanned for this article and many of the original files are now obsolete. That is the reason why some proteins don't have the same PDB IDs as the ones stated in the supplementary materials of this work. This can also be the case with some proteins selected by Taylor and Vaisman [42], whose database is described in the next section.

3.1.2 A database by Taylor and Vaisman

Another database of couples of homologous thermophilic or hyperthermophilic and mesophilic proteins was compiled by Taylor and Vaisman, on which they performed their analyses in the paper "*Discrimination and Classification of Thermophilic and Mesophilic Proteins*" [42]. Their aim was to assemble two databases with pairs of proteins with a similar structure, associating to the mesophilic protein a thermophilic or a hyperthermophilic one.

From the entire PDB, they selected all the proteins with no missing coordinates of the alpha-carbons or alternate atoms. What is more, they decided to use only proteins with crystallographic resolution equal or lower than 2.2 Å and *R-factor* smaller than 0.23 – both this parameters have been already introduced at the end of section 1.3.1. Then, using the name of the organism that synthesized each protein, the set was divided into three subsets: one of mesophilic, one of thermophilic and the last one of hyperthermophilic proteins. Then the sets of thermophilic and hyperthermophilic were reduced to non-reundant sets, making sure that no protein had a sequence identity greater than 30% with any other in the set. At this point the proteins from the extremophiles organisms were aligned with their mesophilic counterparts, looking for proteins with the same *Enzyme Commission number* (*EC number*) [1] or function annotation. The couples formed in this way were then kept if they had a *Root Mean Square Deviation* (RMSD) smaller than 4 Å on at least 80% of the aligned structures. For the majority of the proteins it was impossible to find even one mesophilic match, but for some there were many of them (in this cases maximum five mesophilic proteins were kept in the database).

Putting together the two databases they created, since for this work we do not differentiate between thermophilic or hyperthermophilic proteins, a total of 126 pairs resulted.

3.1.3 Papers analysing a few couples

[1] The EC number is a system of classification for enzymes that takes into account the chemical reactions that they are involved in, so it does not specify the enzyme itself, but the reaction that it catalyses. For this reason different enzymes that catalyse the same reactions have the same EC number, even if those enzymes had a parallel evolution and not a common ancestor.

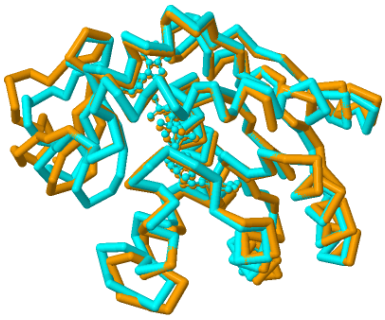


Figure 3.2: Example of two aligned proteins: two homologous Adenylate Kinase whose PDB IDs are 1ZIP for the thermophilic protein and 1P3J for the mesophilic one.

In addition to the available databases, a few other proteins were added to our list. In the literature various papers citing some carefully picked couples of homologous thermophilic and mesophilic proteins can be found. They analyse a limited number of couples (some even just one). Not all the couples that were found were inserted in the beginning list; some were neglected because of a preliminary measure of how similar the thermophilic structure was to the mesophilic. This comparison was done because in the majority of the papers the similarity among proteins was not taken into account. The analysis, knowing the Protein Data Bank ID of the promising proteins couples from the papers, was carried out on http://www.rcsb.org [4] with the Comparison Tool, choosing jFATCAT as the Structure Alignment method. Only couples of proteins with a high value of *Identity* and *Similarity* were selected for the database, the cutoffs chosen to be Identity $> 55.0\%$ and Similarity $> 60.0\%$. This is the reason why, form a first review of the literature, only sixteen pairs that conformed with the good alignment requested were found. In table A.2, which can be found in appendix A, all the pairs are listed with the reference to the article where they were found, their PDB ID, their values of *Identity* and *Similarity*, their length and the resolution of the density map in the PDB file.

In picture 3.2 it is shown an example of an alignment of two chosen proteins. The pair presented is a couple of Adenylate Kinase. In orange the protein of a thermophilic organism, *Bacillus stearothermophilus*, can be seen, while in cyan the protein from a mesophilic one *Bacillus subtilis*. This couple of proteins was analysed in [2], where the authors reported a criticism about the common use of pairing thermophilic and mesophilic proteins from organisms that do not share a recent common progenitor. By doing that, proteins that can be different because of different evolution strategies, that may or may not be related to temperature adaptation, could be compared causing misleading results. Therefore, the two Adenylate Kinases that were studied in the article belong to closely related organism. This reprimand has confirmed the necessity of having a good database in order to be able to draw conclusions from the data. For this database a phylogenetic analysis of the organisms of each protein couple is not carried out, but it is imposed that the proteins in the selected pairs should be very similar. Thanks to this high similarity we can assume a close common ancestor and the little changes are more likely to be caused by the different temperature of their habitat rather that other evolutionary adaptations.

3.1.4 The final list

Putting together all the couples cited in the papers mentioned above, a final list with the PDB IDs of all the couples of proteins was created. This list consists of 894 proteins, that is 447 pairs: 322 from Glyakina, plus 126 from Taylor, plus 16 from various other publications, minus 17 that where identically repeated in the different subsets. In appendix A, table A.1 reports the PDB IDs of all the couples.

From the PDB IDs in this list it is possible to retrieve the correspondent PDB files. What is a PDB file and what it contains has been already described in section 1.3.2. From the PDB files information was collected to asses the quality of the database and refine it removing the pairs that do not have the wanted characteristics. In the next section this final selection is going to be discussed.

3.2 Quality of the database and its final refinements

The assessment of the quality of the list of proteins collected and the decision of what pair to include in the final databased used for the analysis is a crucial step. The aim of this evaluation was to confidently obtain a database that could allow to produce good protein contact maps and to compare pairs with very high functional and structural similarities. To obtain this, two aspects were evaluated: the quality of the single PDB files, i.e. of the atomic model reconstructed from the X-ray crystallography and how good the two proteins in each pair are matching. After the check of the atomic model quality, the proteins that do not comply with the wanted standards are removed.

3.2.1 Quality of the atomic models

Checking the quality of the atomic models of the proteins in the database means looking, file by file, at the quality parameters of the crystallography reconstruction: the resolution, the R-factor and the number of atoms whose position was not found when modelling the diffraction patterns into the proteins' structures. However, before checking those parameters, models obtained with a technique different from the X-rays crystallography were discarded.

Method: X-rays crystallography or NMR spectroscopy

As described in section 1.3, there are two main methods in use to obtain an experimental 3D model of the atomic structure of a protein: the X-rays crystallography and NMR spectroscopy. Due to the higher resolution obtained with the first technique, it was chosen to use only structures obtained with the crystallography

method, eliminating the others. In the list of PDB files 8 protein models were reconstructed with NMR spectroscopy, causing 7 couples – one had both models obtained with NMR – to be removed from the database. This reduces the database to 440 pairs of proteins.

Resolution

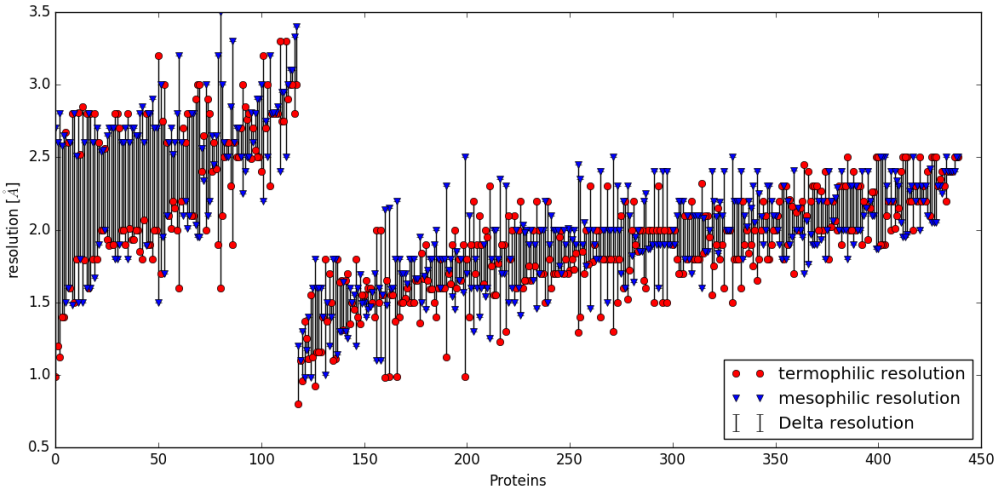


Figure 3.3: The resolution of all the couples in the database. The bars connect each thermophilic protein's resolution to its mesophilic counterpart. The first 133 pairs are the ones that have at least one protein with resolution over the chosen threshold of 2.5 Å, while the others are the better ones.

Resolution is a measure of the quality of the crystal. If all the proteins in the crystal are perfectly aligned, then they will scatter the light with the same angle and the diffraction pattern will discriminate well the fine details of the protein and present a resolution value close to 1 Å. Otherwise, if the proteins are even slightly differently arranged in the crystal, the diffraction pattern will not carry as much information. In models with resolutions around 3 Å only the approximate contours of the protein will be distinguishable.

In graph 3.3 one can see the resolution of each file in the database, arranged by couples. The highest resolution in the database is 0.8 Å, while the lowest is 3.5 Å, denoting a consistent variation between the proteins with the finest resolution and the one with the smallest. The average on the whole database is 2.04 Å with standard deviation of 0.46 Å.

In order to discard low quality protein structures, 2.5 Å has been chosen as upper threshold for the value of resolution, relaxing a little bit the chosen value of 2.2 Å by Taylor and Vaisman [42] for their database. The reason for this change is that in

this case there will be an additional quality check about the missing residues: he proteins will be discarded in a second time, in all the cases where a resolution of less than 2.5 Å was not enough to delineate the position of all the residues. The more severe threshold of 2.2 Å is anyway imposed on the structures that do not declare their R-factor, as stated in the next section.

With this quality check, out of the 880 files, 133 protein models, of which 57 thermophilic and 76 mesophilic, present a resolution that did not comply with the chosen threshold. This causes 118 pairs, those that have at least one protein with a resolution over 2.5 Å, to be neglected for the analysis.

R-factor

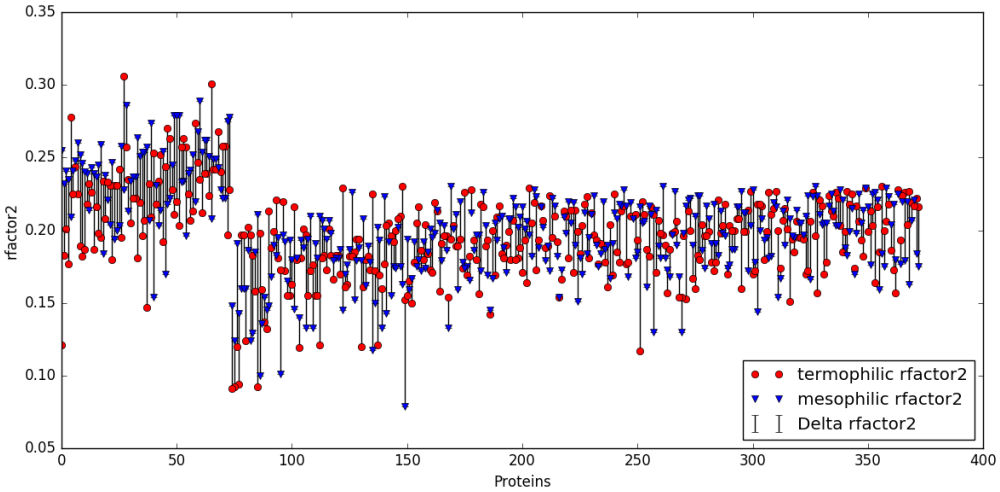


Figure 3.4: R-factor of each protein in the database. The bars indicate how much the R-factor of the two proteins in every couple is distant from each other. The first 85 pairs are the ones that have at least one protein with R-factor over the limiting threshold of 0.23, followed by the good ones. 67 couples are missing because their files do not contain information about the R-factor of their model.

The R-factor is a quality parameter of the reconstructed model compared with the experimental diffraction pattern. If the simulated diffraction pattern from the reconstructed model is exactly the same as the experimental one, the model presents a R-factor of 0, while if the model is made of just randomly positioned atoms, the R-factor is then in the neighbourhood of 0.63.

In figure 3.4 the mean value of the R-factor for each couple in the database is shown. The bars are an indication of the difference between the R-factors of the two proteins in the various pairs. Not all files reported this information in the header, in that case the free R-factor is used. Since the free R-value is higher than the

R-factor for the reason described in section 1.3.1, the values were therefore *corrected* by subtracting 0.04, which is the average difference between the R-factor and the free R-factor over the 745 files that reported both values. Nonetheless, in 67 files there were neither the R-factor nor the free R-factor; these files were kept only if the resolution value was below 2.2 Å, a more restrictive threshold than for the others. Between the ones that have that information, the lowest value for the R-factor is 0.079, while the highest is 0.306; the mean value is 0.198 with a standard deviation of 0.032.

Using in this case the threshold proposed by Taylor and Vaisman [42], only the models with a R-factor lower than 0.23 were kept in the database used in the analysis. Because of this imposition, 85 protein models out of 880 had to be discarded.

Missing residues and atoms

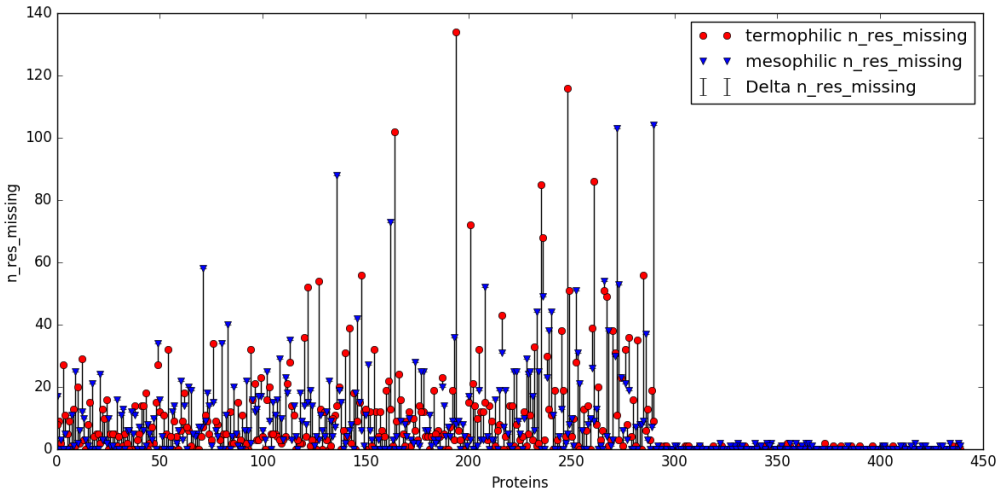


Figure 3.5: For each couple it is shown the number of missing residues for the termophilic and the homologous mesophilic protein. The first 304 couples have more than 1 missing residue; 89 of the others have 0 missing residues.

To produce a protein contact map on which meaningful measures can be performed, all the positions of the residues have to be known as precise as possible. In fact, if in the PDB file some residues or relevant atoms are missing, the contact matrix that is obtained has some missing rows and columns. That is to say that if the coordinates of the residue are missing it will not be possible to obtain any information on that node of the network. Therefore it was checked whether there are missing residues or atoms in the structures in the database.

From the image 3.5 it is possible to appreciate how many proteins have some missing residues. In the database, 160 thermophilic and 172 mesophilic proteins have

no missing residues, but they form only 89 pairs with no missing residues. If instead at most one missing residue per protein chain is accepted, the number of satisfactory couples reaches 135.

It was then checked the amount of missing Cα and Cβ atoms in each protein structure. In the 89 models with no missing residues, there are zero proteins with missing Cα and 2 proteins with 1 missing Cβ. These two proteins were not removed from the database because the coordinates of the missing Cβ can be well approximated by the correspondent Cα's.

The database after the removal of the unwanted pairs

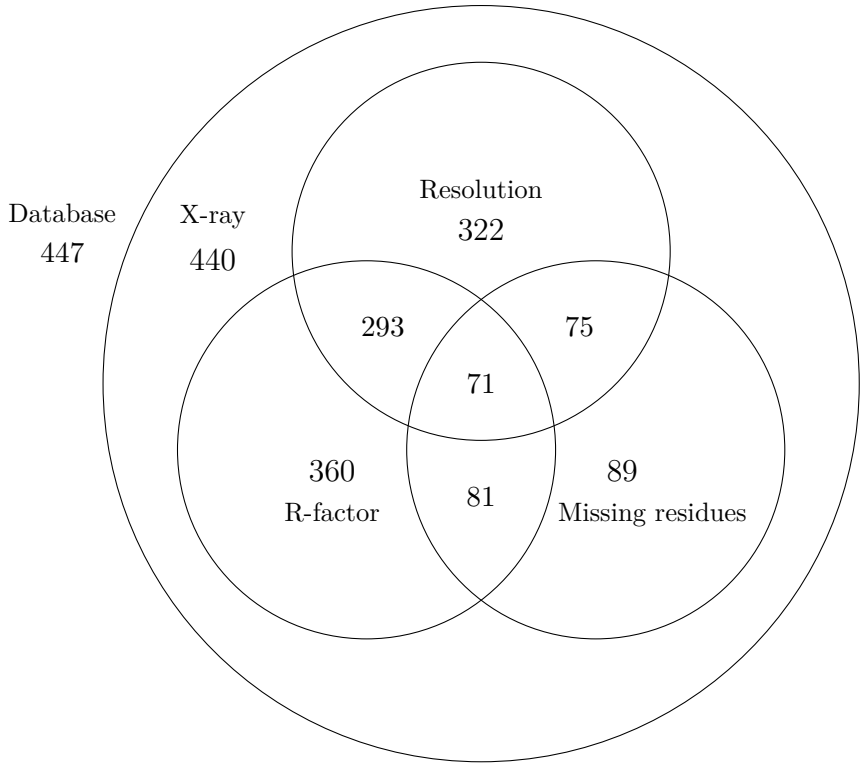


Figure 3.6: The number of pairs that remained in the database after the quality check on the PDB files.

According to the results of the quality check on the PDB files, some couples were removed from the database of 447 couples of proteins. As it is shown in picture 3.6, the couples whose models were obtained with X-ray crystallography are 440, the couples that have a satisfying resolution are 322, R-factor 360 and number of missing residues 89. Having each couple to comply to all the quality measures made on the database, only 71 pairs remain in the database. A more detailed description of these proteins can be found in appendix B at the table B.1.

In the next section a discussion about how comparable the two proteins in each couple follows.

3.2.2 Quality of the pairs of proteins

The final aim of the analysis is to discriminate which one in every couple of very similar homologous protein chains is a thermophilic and which one is a mesophilic. For this reason various parameters have been taken into account to quantify the similarity between the two proteins in each of the couples that remained in the database after the quality check on the file. In particular the length difference and the coefficients *Identity*, *MaxSub* and *TM-score* were assessed, as described in the next sections.

Length of the proteins

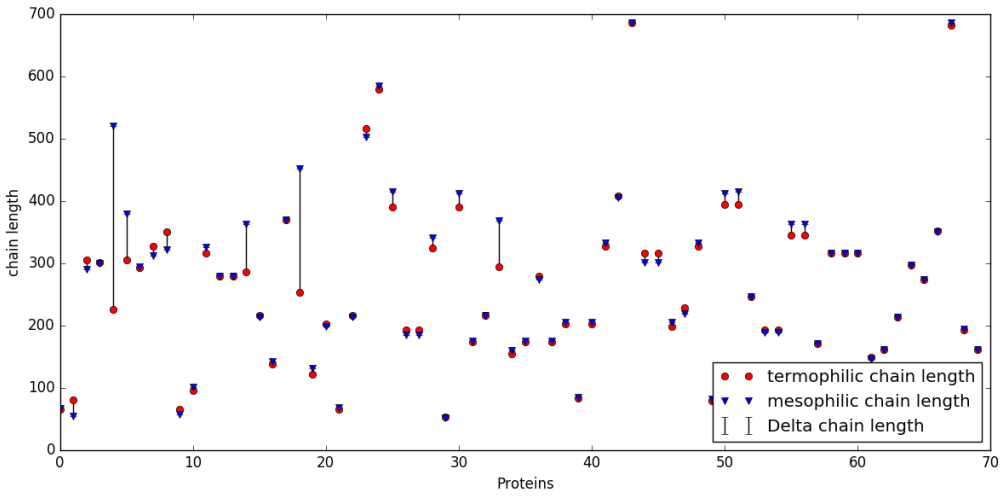


Figure 3.7: The length of each protein chain in the database. The bars highlight the difference between the thermophilic protein and its mesophilic homologous.

The easiest thing to check in order to assess how similar the proteins are in each couple is to look at their length. Moreover, the difference in length of the protein chains is wanted to be as little as possible because it is going to cause a difference in the dimensions of the contact maps. Contact maps of different dimensions are more difficult to compare; in particular there are some measures, like the Laplacian spectrum – the Laplacian has been introduced in section 2.3 – that result in vectors with length the number of nodes in the network, that could easily be set side by side only with other same length vectors. Therefore it is clear that having proteins

with the same number of residues would be a great advantage when analysing the differences between the thermophilic and the mesophilic ones.

In picture 3.7 the lenght of each protein chain is shown. It is clear that three couples have very different lengths, in the worst case the thermophilic protein has 295 residues less than the mesophilic and the other couples have a difference of 198, 77, 75 and 74 residues. Those 5 pairs were removed from the database, resulting in a total of 65 couples of proteins in the final database. After the removal of the too different pairs, the highest difference between a thermophile protein and its mesophile homologue is 29 and the average is 6.43 residues, with a standard deviation of 7.51.

Percentage Identity (PID)

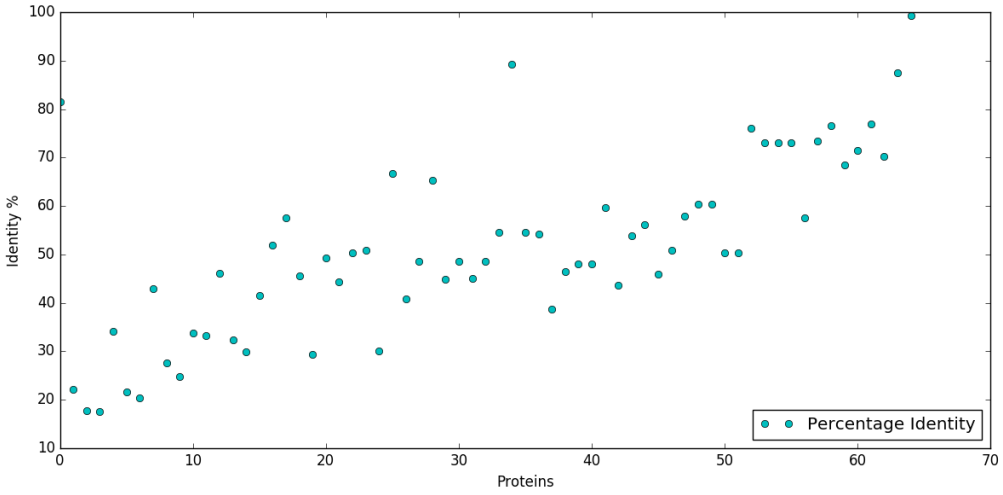


Figure 3.8: Percentage Identity of each pair of proteins in the database.

Another parameter that immediately gives an idea of how similar the proteins are, is the *Percentage Identity* (*PID*), i.e. the percentage of identical residues in the chain of interest. PID is commonly used to define families of proteins, and when its value is over 30% the two proteins are already considered closely related from an evolutionary point of view [13]. The Identity is calculated as:

$$PID = \frac{\text{number of identical aligned residues}}{\text{total number of residues}} \cdot 100, \tag{3.2}$$

The formula seams quite easy, however there is no agreement on which alignment algorithm to use and, what is more, the lengths of the two compared proteins usually are not identical. This creates a number of different definitions of the denominator that can affect the value of PID remarkably [26]. In a paper by Raghava and Barton [32] this issue is extensively examined and a variation up to 22% between

PID calculated with different algorithms is reported. Thus, a more specific definition of the PID used is necessary. The PID calculated on our database is defined as:

$$PID = \frac{\text{number of identical residues}}{\text{number of aligned residues}} \cdot 100, \tag{3.3}$$

where the alignment method is *TM-align* [50], which is an algorithm that maximises the *TM-score*, defined in the equation (3.4) in the next section, between the two proteins.

In picture 3.8 the values of the PID of all the proteins' pairs in the database are shown. From the graph it is easy to see that 7 couples have a PID below 30%. This means that those proteins might not have a close common ancestor; nonetheless they were kept in the database to evaluate their *MaxSub* and *TM-score* values, as described in the next section.

MaxSub and *Template Modelling score*

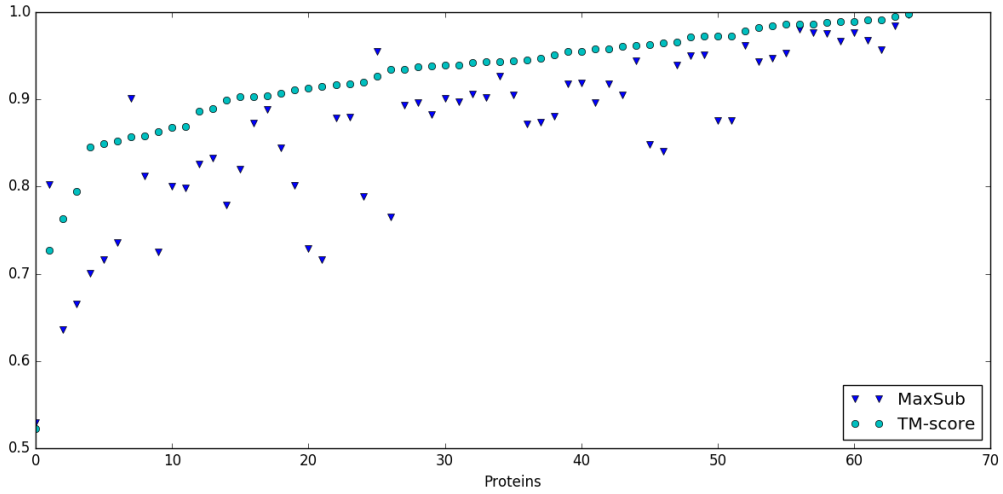


Figure 3.9: The graph shows the coefficients MaxSub and TM-score of every couple of proteins in the database.

Other two parameters that have been calculated on the database and provide meaningful information about the similarity of the two proteins in each couple are the *MaxSub* and the *Template Modelling score*(TM-score). If the Identity takes into account the primary structure of the two proteins to compare, this new coefficients look at their 3D structures and provide a quantitative value of how close they are to each other.

The MaxSub has already been defined in (3.1) earlier in this chapter as:

$$MaxSub = \frac{\sum_i \left(1 + (d_i/d_0)^2\right)^{-1}}{\min(N_1, N_2)}.$$

The values of the MaxSub always lies between 0 and 1, with higher values indicating better superimposing structures. Not taking into account the problem of how to align the two structure, the MaxSub coefficient immediately presents a problem, instrinsic to its definition: its value is dependent on the length of the two aligned proteins. To overcome this limitation, the *TM-score* is itroduced [49]. With this new parameter the score for two randomly aligned proteins is no longer reliant on the number of residues in the proteins, but results for every length about 0.17 [49]. To achieve this result, there is no need to significantly modify the MaxSub; as a matter of fact the only difference between the MaxSub and the TM-score is in the definition of d_0:

$$TM\text{-}score = \frac{\sum_i \left(1 + (d_i/d_0)^2\right)^{-1}}{\min(N_1, N_2)}, \tag{3.4}$$

where d_0, while in the case of MaxSub is a fixed value, for the TM-score is defined as a function of the length N of the targeted protein:

$$d_0(N) = 1.24\sqrt[3]{N - 15} - 1.8. \tag{3.5}$$

The values of d_0, for either the MaxSub and the TM-score, versus the protein lengths are shown in graph 3.10. They match only for proteins with 93 residues. For shorter proteins the TM-score's d_0 is lower then the MaxSub's one, requiring therefore to have closer residues to have the same TM-score and MaxSub, while for longer proteins d_0 is higher, relaxing the condition on how near the residues should be to have the same MaxSub and TM-score.

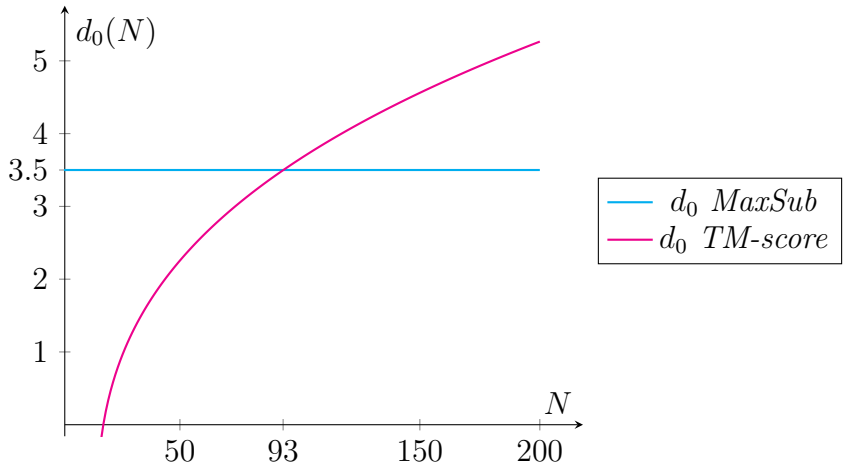


Figure 3.10: The value of d_0 for the *MaxSub* and *TM-score* at different values of N, the length of the target protein.

For our database the results of the *MaxSub* and *TM-score* are shown in figure 3.9.

In this graph is possible to appreciate how all the couples have a TM-score higher than 0.7, except for one that is near to 0.53.

4

The analysis

In this chapter the analyses performed on the database are presented. The analyses were carried out with a custom made code written in Python, the protein files were accessed with the aid of the Biopython module PDB [20].

The first aim of the examinations made was to characterise the dataset and the proteins in it. The second and more ambitious was to discriminate between the thermophilic and mesophilic proteins. The comments on to what extend this second objective has been reached will be presented in the last chapter of this thesis with the conclusions of this work, while in the next sections the results of the analysis are reported.

The fist section of this chapter presents a study of the distances between residues in the proteins, especially focusing on their distributions in the two subsets of thermophiles and mesophiles proteins. In the second one, the Protein Contact Maps creation and the characterization of the resulting contacts is discussed. This is followed by a section on the Laplacian of the PCM and its spectrum is analysed.

4.1 Distance matrix and distance histograms

Constructing a distance matrix is the first step towards the creation of a PCM, but the distance matrix is not only a tool to obtain the Contact Map. The distance matrix itself carries a lot of information, useful to give a first insight on the dataset. In fact, the distribution of the inter-residues distances, on top of been the basis from which to determine whether there is a contact or not between the residues, characterise the structure of the proteins. Therefore, histograms of the distance distributions are presented, after a description of the obtained distance matrices.

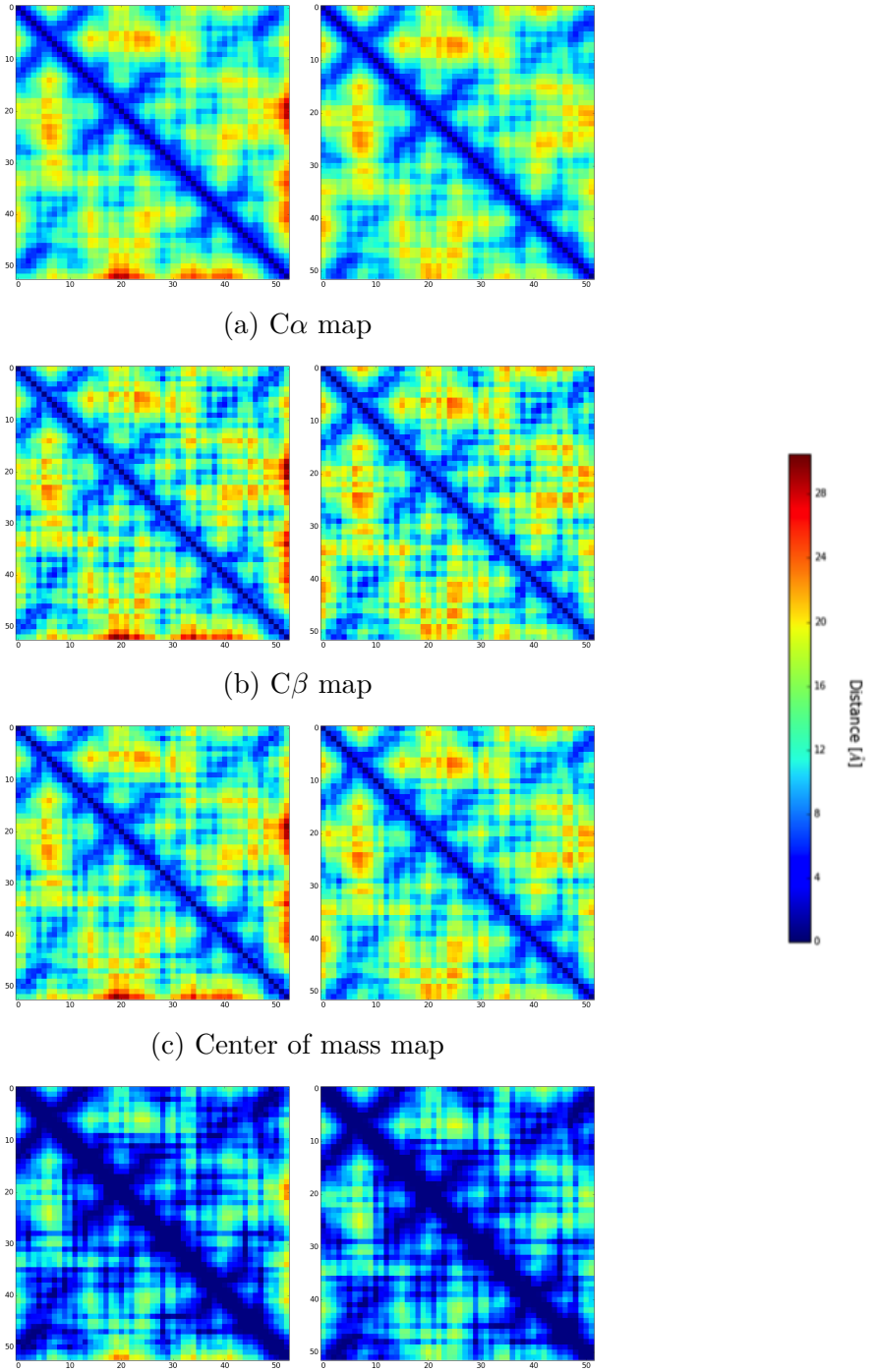
4.1.1 Distance matrices

The distance matrix of a protein is a matrix D, whose general element $d_{i,j}$ is the distance between the i-th and the j-th residue of the protein. It follows that the matrix is a $N \times N$ square matrix, with N the number of residues in the protein. Even if this matrix has such an easy definition, the definition of the distances between residues is not as obvious and there is more than one possible solution. This is the reason why for every protein in the database four corresponding distance matrices have been created. Both single atoms and whole residue methods to define the position of the residues, and therefore the distances, have been used. A first map has been created using the position of the Cα of the residues, a backbone atom that is good as a first approximation of the backbone structure of the protein. For the second one, the position of the residues has been collapsed on the Cβ, which should provide a better estimate of the whole protein structure. In the third matrix the center of mass of the residue has been used as the position of the residue, allowing in this way to consider differently residues with a very long side chain from those with a little one. For the last distance map kind obtained for this work, the distance between two residues is the shortest distances between any two atoms of the two residues from which the Van der Waals radii of the two atoms has been subtracted (had the result been negative it would have been put to 0 Å). The Van der Waals radii used for realise this distance matrix of the closest atoms, taken from [1], can be seen in table 4.1. All these methods for obtaining a distance matrix have already been described in more detail in section 2.2.2.

Atom	Van der Waals radius [Å]	Atom	Van der Waals radius [Å]
H	1.20	O	1.50
N	1.66	C	1.77
S	1.89		

Table 4.1: The Van der Waals radii used for the creation of the closest atom distance matrix, taken form [1].

As an example of the obtained matrices, the maps of a couple of rubredoxin proteins (electron transport), the thermophilic one from Pyrococcus furiosus on the left and the mesophilic one from Desulfovibrio vulgaris on the right, can be seen in figure 4.1. The first two are obtained using the position of the Cα or the Cβ of each residue as the position of the entire residue, the next one is obtained using the centre of mass of the residue and the last one using the closest atoms. From those images one can see the differences between the different kind of distance maps. The alpha carbon type presents a more blurred profile than the beta carbon one, that has more sharp differences between neighbouring cells. The β-carbon maps and the



(a) Cα map

(b) Cβ map

(c) Center of mass map

(d) Closest atom map

Figure 4.1: Distance matrices for a couple of rubredoxin, the thermophilic from Pytococcus furiosus (PDB ID: 1caa.A) on the *left* and the mesophilic from Desulfovibrio vulgaris (PDB ID: 8rxn.A) on the *right*.

center of mass ones are quite similar, with only minor differences. The closest atom maps are very different from all the others due to the way they are obtained: not only are the distances calculated between the closest atoms of the residues, but also the Van der Waals radii of those atoms are subtracted from the resulting amount. It derives that many cells that present a contact have values close to 0 Å; obviously the threshold at which to consider the existence of a link between residues has to be lower than for the other matrices.

From figure 4.1 it is also possible to get a first idea of the structural differences between thermophilic and mesophilic proteins. They are indeed very subtle, not only for this example but for all the proteins in the database. Having chosen couples of very similar proteins, it was to be expected to find very similar distance matrices. To have a better understanding of the degree of similarity between the matrices, histograms showing the distances distribution have been drawn, as described in the next section.

4.1.2 Distance histograms

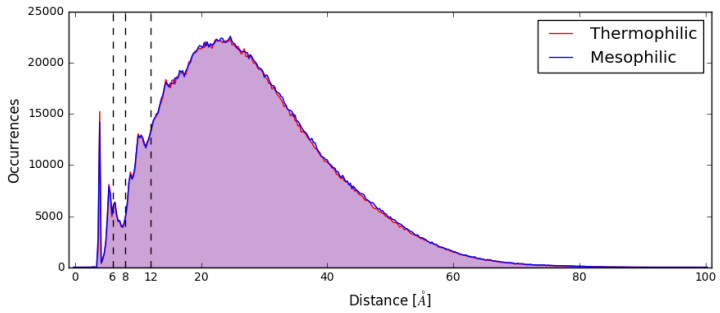
For every distance matrix a relative distance histogram has been obtained. The histograms have been obtained by binning the values of the top triangle of every distance matrix in 410 bins between [0 Å,100 Å]. The [0 Å,100 Å] interval has been chosen to represent all the distances, whose max values are 96.50 Å for the alpha carbon, 96.40 Å for the beta carbon, 96.50 Å for the center of mass and 91.01 Å for the closest atom. The number of bins has instead been chosen with the Freedman-Diaconis rule[18], that, given a sample distribution x, define the size of each bin as:

$$\text{bin size} = 2\frac{Q3(x) - Q1(x)}{\sqrt[3]{N}}, \tag{4.1}$$

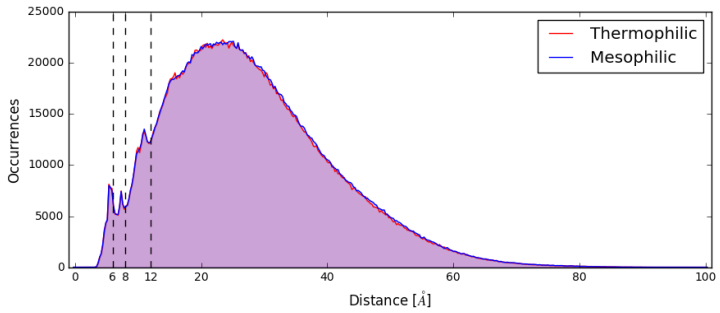
where $Q1(x)$ and $Q3(x)$ are respectively the lower and upper quartiles of the sample x and N is the number of elements in the sample. The smallest bin size for the distance distributions using this rule was 0.244 Å, a value that lead to requiring 410 bins. It was chosen the smallest value, that was obtained for the thermophilic distribution of distances between the closest atoms, because in order to compare the various graphs all had to use the same parameters and the most strict was chosen; in any case as a matter of fact, choosing any of the others would have not changed much, since the biggest bin size was of 0.246 Å for the mesophilic distribution of Cβ distances that would have lead to 406 bins.

The resulting histograms can be seen in figure 4.2, where one can compare the different shapes for the different methods used.

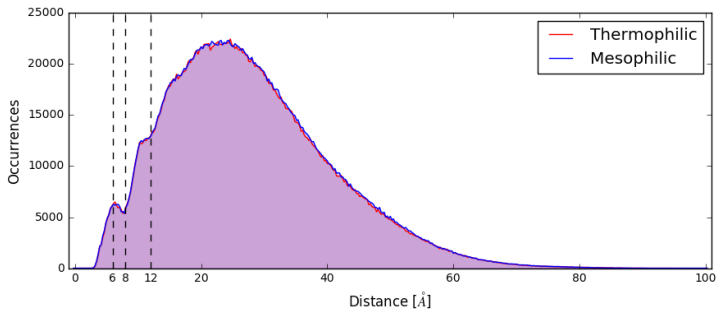
The histogram of the distances between Cαs, figure 4.2a, presents a sharp peak in the bin centered in 3.902 Å. This peak is easy to explain: it is the interval where the distance between carbons-α in residues that are linked by a peptide bond, i.e.



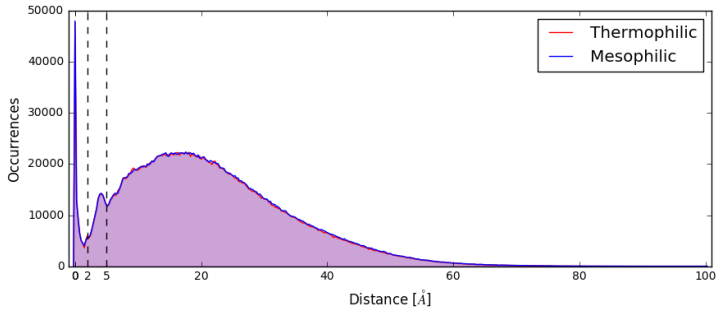
(a) Cα



(b) Cβ



(c) Center of mass



(d) Closest atoms

Figure 4.2: Distance histograms of the dataset for the various kind of distance matrices.

the adjacent residues in the primary structure, falls, being about 3.8 Å. The Cα in an alpha helices or beta sheets structures and the disulfide bonds contribute to the second and third pick, around 6 Å; the distance between bonding Cα in an α helix is about 6 Å, the distance in a β strand is about 5 Å.

The Cβ histogram, figure 4.2b, shows a quite a similar profile to the Cα one, with the main difference of not having the first peak. This is caused by the fact that the distances between the Cβ of adjacent residues are bigger than for the Cα, assuming values as big as around 6 Å, with a distribution whose mode is around 5.5 Å. The center of mass histogram, figure 4.2c, has the same broadly shape of the Cβ histogram, while the closest atoms one is pretty different from the previous ones.

The closest atoms histogram, figure 4.2d, has a very high peak (47914 occurrences, more than 3 times the number in the Cβ peak) in the bin centred in 0 Å. In this peak and in its vicinity, as a matter of fact, one can expect to find the peptide bonds but also the other bonds between residues in the protein.

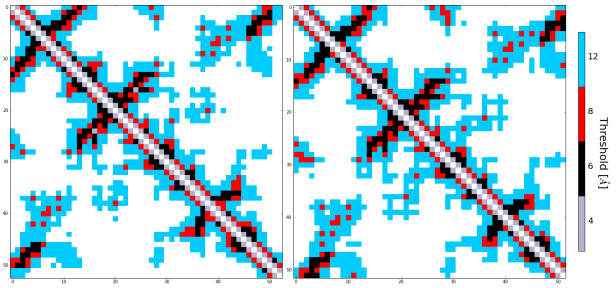
4.2 Protein Contact Maps

The PCM is a matrix A whose generic element $a_{i,j}$ is 1 if the i-th and j-th residues are in contact, 0 otherwise. To set whether there is a contact or not between any two residues the distance between them should fall in a chosen interval $\mathcal{I}$. This choice heavily modify the aspect of the resulting contact map, as it can be seen in figure 4.3, that will be commented in the next paragraph.

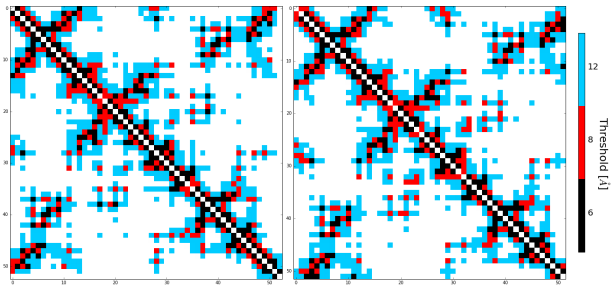
Choosing the thresholds

From the distance matrices the Protein Contact Maps were obtained for all the proteins in the database applying an upper threshold. In figure 4.3 it is possible to see the contact maps with different choices of threshold.

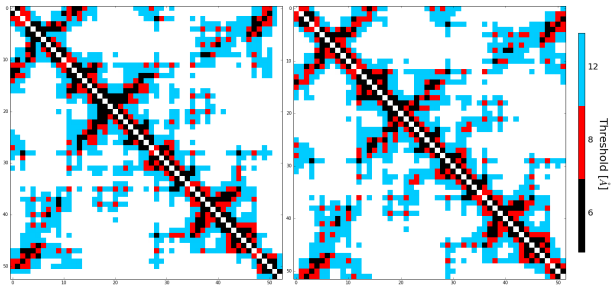
The first couple of graphs, figure 4.3a, are the Cα contact maps. For this kind of map 4 different thresholds are plotted in the charts: 4,6,8 and 12 Å. A threshold as low as 4 Å allows to see the extent of residues whose Cαs are very close to each other, about the distance that occur when the two residues are linked by a peptide bond, that contribute to the first peak in the distance distribution shown in figure 4.2a. Except for a contact in the thermophile map, all the others are concentrated on the first diagonal, confirming that the high majority of those very close Cαs is caused by the primary structure of the protein. Therefore, imposing a lower threshold to eliminate the trivial contacts of the backbone could be reasonable, but it would be a problem to lose those rare links between residues that are very close outside the first diagonal and, what is more, eliminating the first diagonal could in some cases disconnect the graph, a propriety that would be important to preserve for the following analysis on the PCM. This is why the lower limit of the interval $\mathcal{I}$ has been



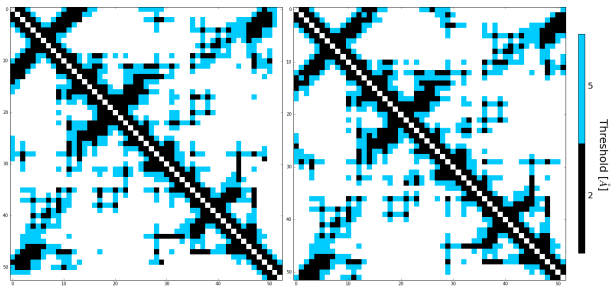
(a) Cα



(b) Cβ



(c) Center of mass



(d) Closest atoms

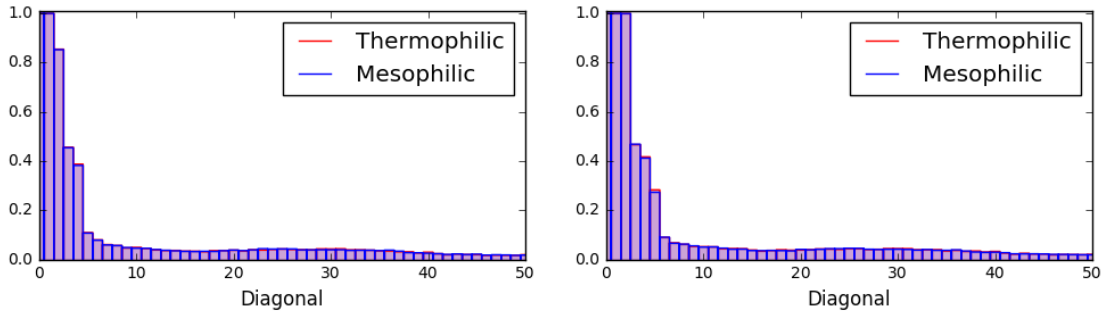
Figure 4.3: Protein contact maps for a couple of rubredoxin, the thermophilic from Pytococcus furiosus (PDB ID: 1caa.A) on the *left* and the mesophilic from Desulfovibrio vulgaris (PDB ID: 8rxn.A) on the *right*. The colors represent different threshold choices, as specified in their color bar.

set to 0 Å, and the redundant information of the links in the backbone of the protein will be eliminated successively, keeping into account the connectivity proprieties of the network. Similarly, for the other kind of maps the lower threshold has been set to zero, also because setting a lower threshold in those cases has less physical meaning, since there is not a fixed distance between, for example, the Cβs of two residues connected with a peptide bond.

For the upper threshold it is possible to see that, for the Cα, Cβ and closest atom maps, liming it at 6 Å would mean leaving out many contacts, while setting it at 12 Å accepting a lot of noise. For this reason it was chosen to use the midway and follow what is one of the most common choices in the literature, selecting as contact all the distances that are in the interval $\mathcal{I} = [0\,\text{Å}, 8\,\text{Å}]$. For the closest atoms map, see figure 4.3d, the interval chosen was $\mathcal{I} = [0\,\text{Å}, 2\,\text{Å}]$, since it is possible to see that using as upper threshold 5 Å, would signify accepting more or less the contacts with the 12 Å threshold in the other maps.

With this choices of intervals, the total number of contacts for the Cα is 175152 for the thermophiles and 175358 for the mesophiles, for the Cβ is 179216 and 179014, for the center of mass 177692 and 178508 and for the closest atom the number of contact is 198350 and 198108, slightly higher than for the other kind of maps. Interestingly, the thermophilic and mesophilic values are very similar and neither of the two types of protein have consistently a higher number of contacts than the other.

4.2.1 Contact Frequency and Contact Order



(a) Cβ (b) Closest atoms

Figure 4.4: Frequency of having a contact vs. the distance in the primary structure of the two residues, i.e. the diagonal number of the cell in the PCM, for the first 50 positions.

The total number of contacts in the obtained PCM is very similar between thermophiles and mesophiles, so a more in depth analysis was carried out to check

whether there is a reliable difference in the position distance on the primary chain between the residues that share a link. The contacts between residues next to each others in the linear chain, i.e. the i-th residue and the $i+1$, are shown in the map on the first diagonal of the matrix, i.e. the diagonal whose elements are $a_{i,i+1}$; the residues that have one residue between each others are on the second diagonal, $a_{i,i+2}$, and so on. Therefore, the contact frequency was calculated for every diagonal d as:

$$\text{contact probability}(d) = \frac{\text{total number of contacts in the } d \text{ diagonal}}{\text{total number of cells in the } d \text{ diagonal}}. \tag{4.2}$$

As it can be seen in figure 4.4, where the contact frequencies of the first 50 diagonals of the Cβ and closest atoms maps are plotted, there is no significant difference between the thermophilic and mesophilic proteins. It is anyway interesting to notice that the closest atoms kind of map has not only always a contact on the first diagonal, but also on the second one. What is more, also the frequency of contact on the 5^{th} diagonal is pretty different, being for the closest atoms almost three times the one of the Cβ, with slightly more contacts for the thermophilic proteins. It is therefore mainly in those two diagonals that this kind of map presents significantly more contacts than the others instead of having a uniform rise in every diagonal.

To refine the characterisation of the distributions of the contacts in the PCMs, another parameter was measured on the database: the Contact Order (CO). The Contact Order is a topology measurement, defined as:

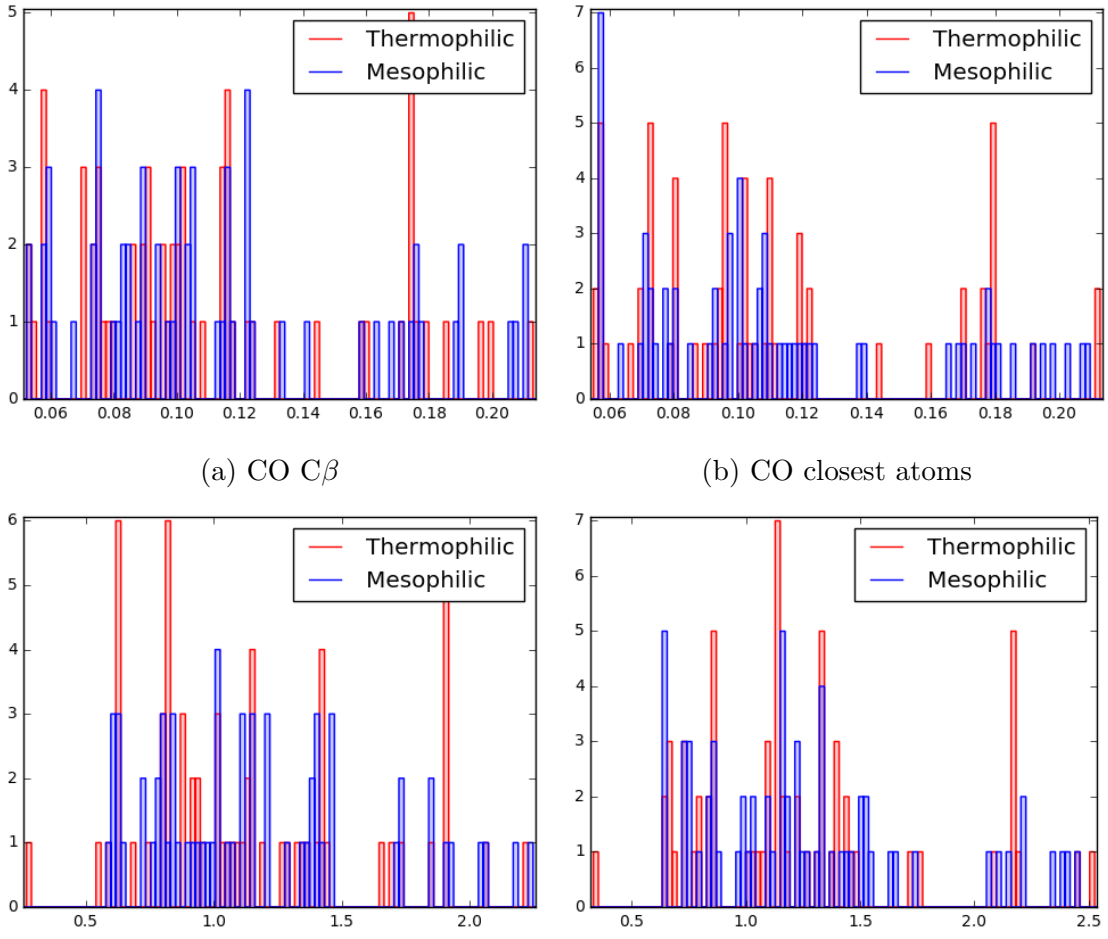
$$CO = \frac{1}{NL} \sum_{ij}^{N} a_{ij} \cdot |i-j|, \tag{4.3}$$

where a_{ij} is the element of the contact map, N is the length of the map and L is the number of links. From the definition it follows that the CO is an index that is higher for maps that have more contacts far away from the main diagonal, and lower for those whose contacts concentrate in the first diagonals.

It is also possible to modify the CO to obtain a Long Range Contact Order (LRCO) to highlight the contribute of the links between distant residues, neglecting the first diagonals in the PCM. The LRCO is therefore defined similarly as the CO, but for residues that are at least 12 spots apart in the primary structure:

$$LRCO = \frac{1}{N^2} \sum_{ij,|i-j|>12}^{N} a_{ij} \cdot |i-j|. \tag{4.4}$$

The resulting CO values are in the range 0.044 , to 0.213 , while the LRCO 0.23 to 2.51 . The distribution of the values of CO and LRCO for the Cβ and closest atoms maps can be seen in figure 4.5. From the graphs it is clear that is not possible to differentiate the thermophilic from the mesophilic distribution. To check whether it is possible to differentiate not in the whole database, but only between the two



(a) CO Cβ

(b) CO closest atoms

(c) LRCO Cβ

(d) LRCO closest atoms

Figure 4.5: Distribution of the CO and LRCO values of the Cβ and closest atoms maps.

proteins in the couple which one is thermophilic and which one is mesophilic, the differences between the CO values of each couple have been calculated. The average difference is, for all the maps, in the interval $[-2.6 \times 10^{-3}, -3.6 \times 10^{-3}]$, that is less than 10% of the smallest CO value. Even if the value is very close to zero, its sign is reflected in the fact that between the 52% and the 60% of the couples, depending on the type of map, presents a CO value higher for the mesophilic PCM. For the LRCO the averages for the various maps are in the range $[-2.4 \times 10^{-2}, -3.8 \times 10^{-2}]$; in this case, even if the averages of the differences are negative, the percentages of couples with a higher LRCO values in the mesophilic contact map is not always higher than 50%, being 48% for the Cα, 59% for the Cβ, 60% for the center of mass and 49% for the closest atoms.

Summing up, there are no major differences in the contact distributions between the two subsets, thermophilic and mesophilic proteins, in our database. The probability distribution of the contacts depending on the distances of the residues in the primary structure.

4.3 The Laplacian and its spectrum

The Laplacian spectrum of a contact map, as discussed in section 2.3, is strongly related with the vibrations of the system. Therefore it was chosen as mean of analysis of the two subsets in the database, hoping to find a clear difference between the thermophilic proteins that endure higher temperatures, and therefore bigger amplitudes of the thermal vibrations for its atoms, than the mesophilic counterparts. As a matter of fact, the *hypothesis of equivalent*, or *corresponding, states* states that the thermophilic proteins are more rigid than mesophilic ones at room temperature, reaching the same flexibility only at higher temperatures where they are fully functional. This is the reason why the spectrum of the Laplacian matrix was analysed, looking for differences between the thermophiles and the mesophiles, as described in the next section.

4.3.1 Spectrum analysis

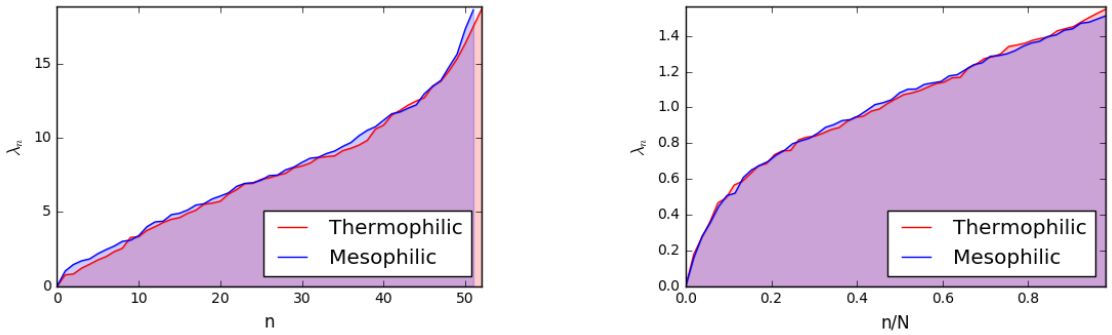
From the contact maps of the proteins in the database the protein backbone was removed to eliminate the trivial contacts on the first diagonals. This was done in order to give more weight to the long range contacts where we expect to find the biggest differences between thermophilic and mesophilic PCM. The backbone was eliminated by removing the contacts from the diagonals. The number of diagonals whose contacts were removed, was chosen on the basis of the connectivity propriety of the graph. If d was the smallest number of diagonals with no contacts that caused the network to become disconnected, the PCM A without backbone, called B, had a

general element of the kind:

$$b_{i,j} = \begin{cases} a_{i,j} & \text{if } |i-j| > d \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

In this way the graph remains connected and the number diagonals to remove depend on the propriety of the graph itself. Having a connected graph is useful since it minimizes the number of null eigenvalues of the graph Laplacian, that are as many as the connected components of the network.

The Laplacian matrix, which is defined in equation (2.11), presents a spectrum with eigenvalues that are *real*, since the Laplacian matrix is symmetric, and *nonnegative*, since the Laplacian is positive semi-definite. Similarly, the normalised Laplacian, defined in equation (2.12), has a spectrum that is real, nonnegative and also with values always in the interval $[0, 2]$.



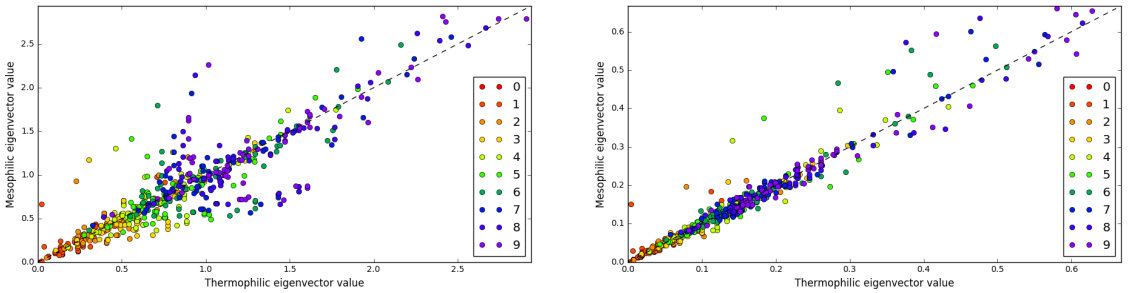
(a) Laplacian

(b) Normalised Laplacian

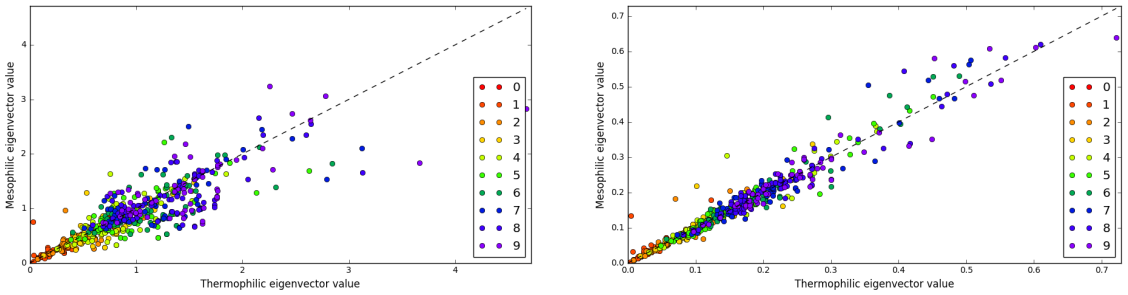
Figure 4.6: The eigenvalues of the closest atoms contact map Laplacian and normalised Laplacian for the couple of proteins 1caa.A and 8rxn.A.

The graphs showing the Laplacian and normalised Laplacian spectrum can be seen in figure 4.6 for the proteins 1caa.A and 8rxn.A, the same couple shown in figures 4.1 and 4.3. For the normalised Laplacian, the x axis has been normalised on the length of the protein chain and therefore its values are $x = n/N_t$ for the thermophilic chain of length N_t and $x = n/N_m$ for the mesophilic one of length N_m. In this way it is possible to compare the different values of the thermophilic and the mesophilic Laplacian spectrum.

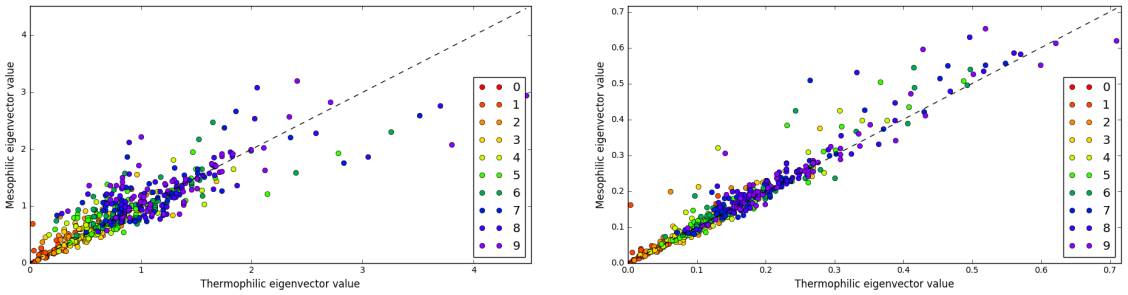
From the two graphs it is possible to notice that again the thermophilic and mesophilic values are very similar and it is difficult to discriminate between the two groups. This is not always true in the whole database; for some couples of proteins it is possible to see that either the thermophilic or mesophilic values are higher than the others. Unfortunately, it is not always the same one of the two subsets that is higher than the other, as it is noticeable from figures 4.7 and 4.8, where respectively the initial and final thermophilic eigenvalues against the mesophilic ones are plotted.



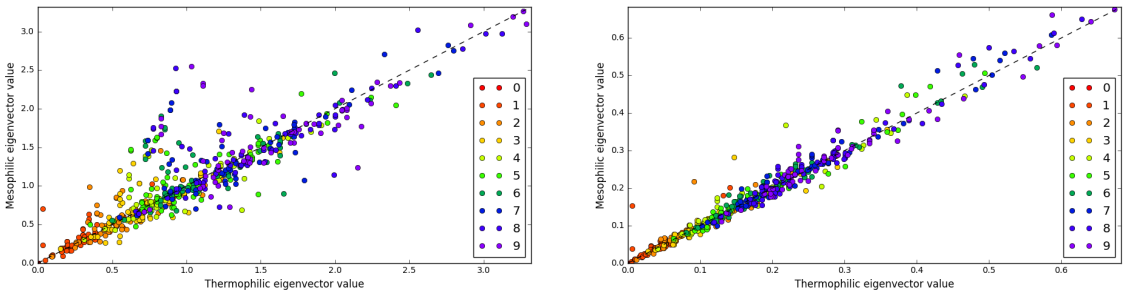
(a) Cα



(b) Cβ

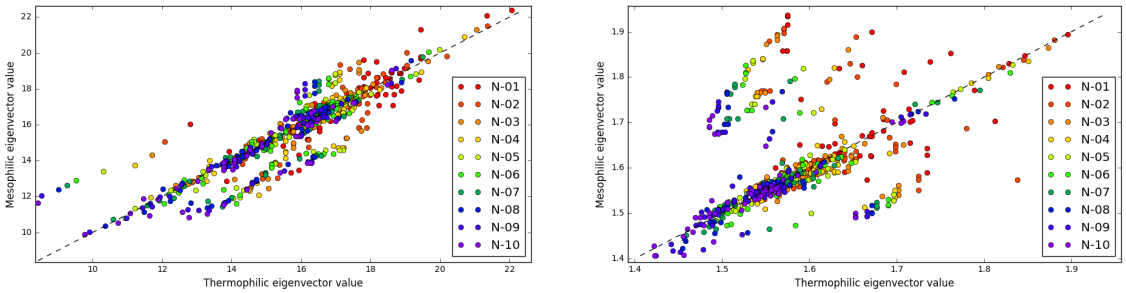


(c) Center of mass

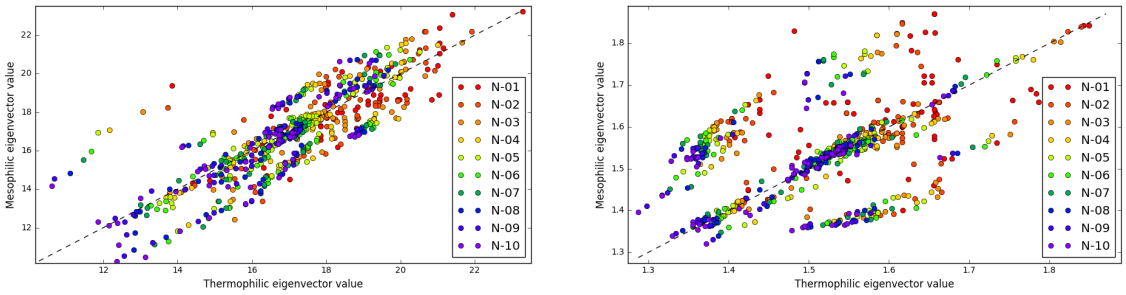


(d) Closest atoms

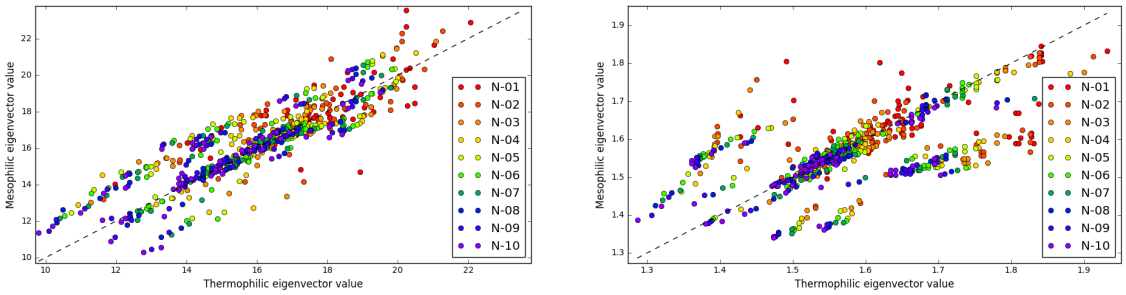
Figure 4.7: The first eigenvalues of the Laplacian on the *left* and of the normalised Laplacian on the *right*, plotted thermophilic vs mesophilic.



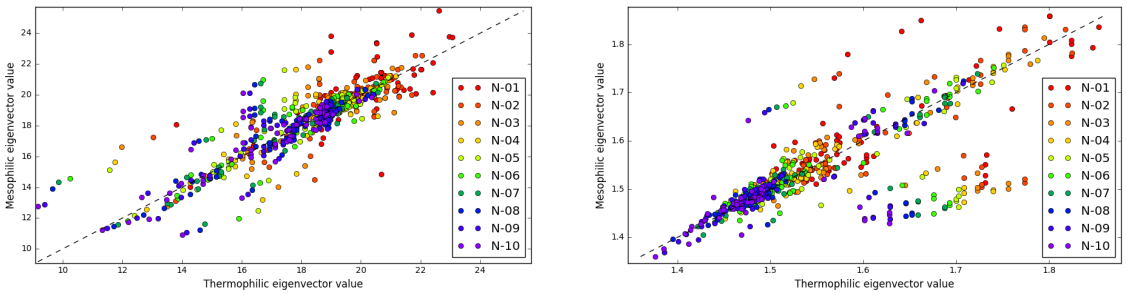
(a) Cα



(b) Cβ



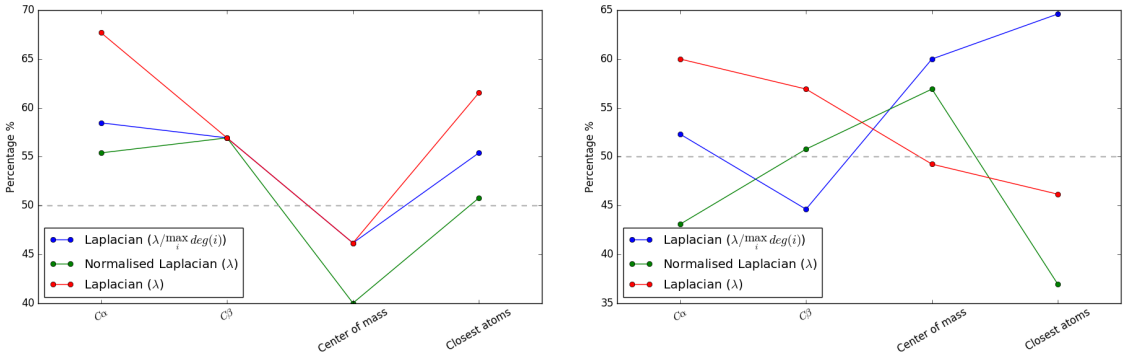
(c) Center of mass



(d) Closest atoms

Figure 4.8: The last eigenvalues of the Laplacian on the *left* and of the normalised Laplacian on the *right*, plotted thermophilic vs mesophilic.

In figures 4.7 and 4.8 the values of the first ten eigenvalues and last ten eigenvalues are plotted, thermophilic vs mesophilic, for the Lapacian and the normalised Laplacian. The eigenvalues of the graph Laplacian have been divided by $\max_i deg(i)$, the highest node degree of each graph, in order to be comparable between each others. From the graphs it is possible to notice that the values are in general very close to the diagonal, meaning that the thermophilic and mesophilic values are very similar. In figure 4.9 it is possible to appreciate the percentage of couples whose thermophilic protein has the sum of the first or last ten eigenvalues higher than the mesophilic one.



(a) Initial (b) Final

Figure 4.9: The percentage of pairs whose sum of the first or last ten Laplacian eigenvalues is higher for the thermophilic.

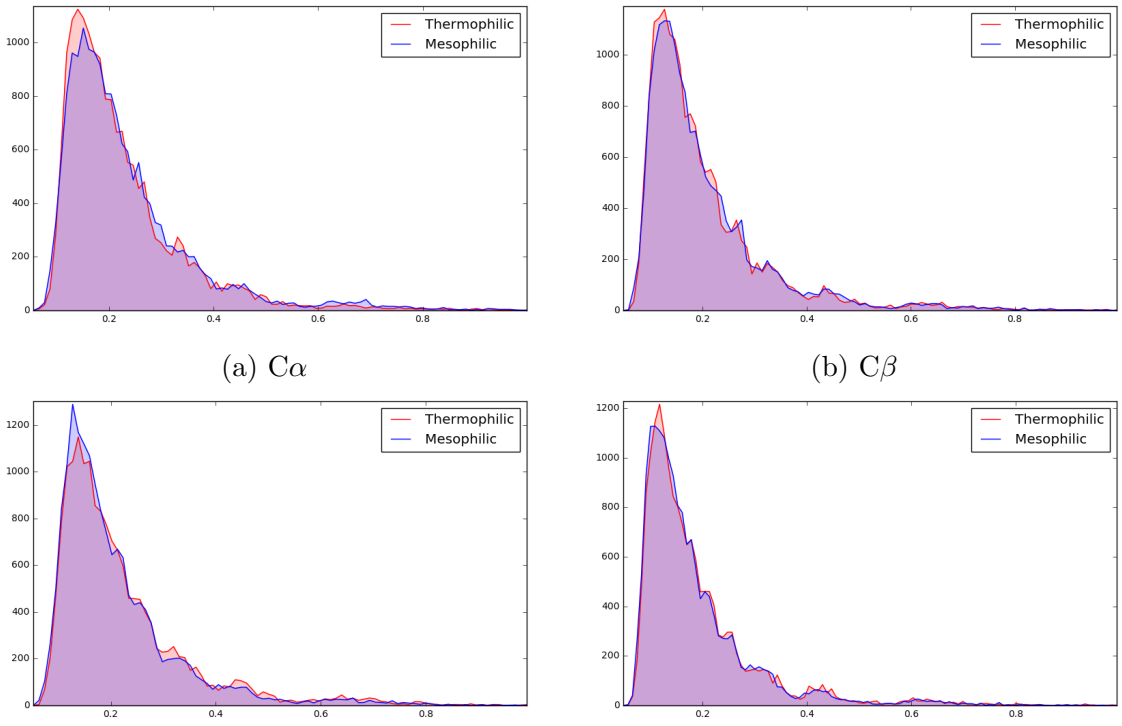
4.3.2 Vibrations

Using the equation 2.14, it is easy to obtain a distribution of Δs, a quantity proportional to what was there defined as Δx, the magnitude of vibrations in the protein, without having to assume values for the spring constant k:

$$\Delta s_i = \sqrt{l^+_{ii}}, \tag{4.6}$$

where i indicates the i-th residue in the chain.

The distributions of Δs for all the proteins in the database and all the maps can be seen in figure 4.10. The different kinds of contact maps exhibit slightly different distributions. The great majority of the Δs is concentrated on the beginning of the distributions, in fact the medians are between 0.15, for the closest atoms, and 0.20, for the Cα, even if the highest values, not shown in the graphs, are 2.6 for the Cβ, 3.06 and 3.37 for the Cα and closest atoms, respectively, and 4.13 for the center of mass.



(a) Cα

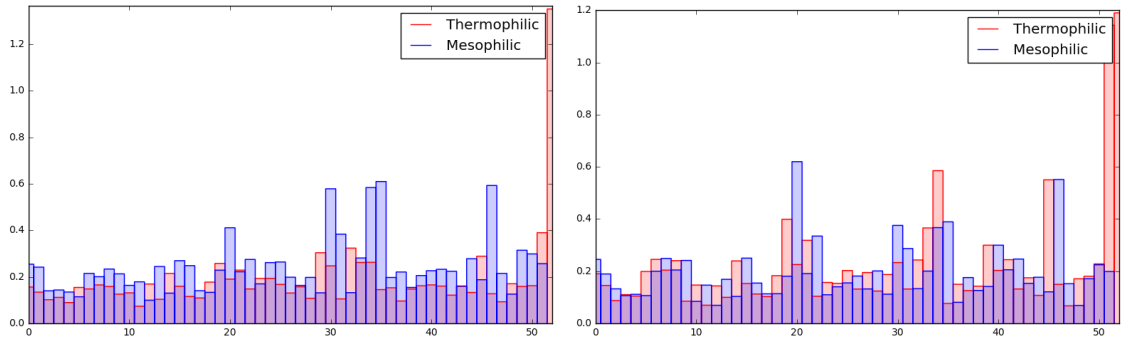
(b) Cβ

(c) Center of mass

(d) Closest atoms

Figure 4.10: Histogram showing the distributions of the values < 1.0 of Δs of the thermophilic and mesophilic proteins for the various type of contact map. The x axis has been limited between [0,1] to zoom in on the initial distribution.

The distributions of the thermophilic and mesophilic proteins are, once more, very similar. It is an interesting fact to notice that the Δs are pretty similar between the two subgroups of proteins, even for the highest values. The hypothesis of equivalent states can not find a confirmation from these distributions.



(a) Cβ (b) Closest atoms

Figure 4.11: The Δs value of the each residue for the couple of proteins 1caa.A/8rxn.A.

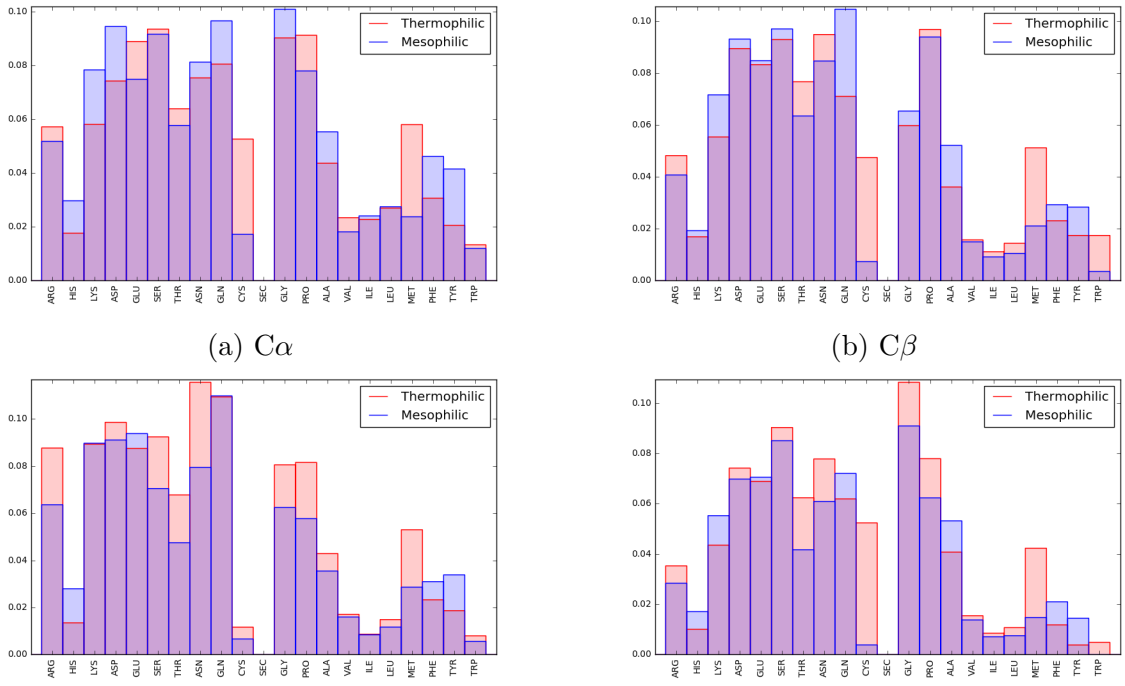
Being possible to associate a Δs to each residue, as shown in 4.11 for the couple 1caa.A/8rxn.A, more information can be gathered from this analysis, to look for clues of thermostability that up to now have been evasive. In particular, interesting questions that can be answered are whether there are some residues that are more likely to have a high value of Δs than others and what is the probability that residues that have a high value of Δs in the thermophilic protein are substituted with other amino acids in the mesophilic.

Firstly, to answer those questions, a more formal definition of high values of Δs is requested. The Δs_i, of the i-th residue, is considered to be high if:

$$\Delta s_i \geqslant \frac{M_t + S_t + (M_m + S_m)}{2}, \tag{4.7}$$

where M_t and M_m are the arithmetic mean and S_t and S_m the corrected sample standard deviation of the distribution of Δs for, respectively, the thermophilic and mesophilic protein in each couple. With this definition there is no common threshold for all the Δs in the database, but the different couples have different thresholds. In this way, the variability between the various protein chains is taken into account.

The number of occurrences for thermophilic and mesophilic chains of high Δs values by residues, normalised with the number of occurrence of that residue in the thermophilic or mesophilic proteins respectively, can be seen in the graphs 4.12. It can be observed that some residues are favoured by thermophilic and other by mesophilic proteins, but there are some changes in the different type of contact map.



(a) Cα

(b) Cβ

(c) Center of mass

(d) Closest atoms

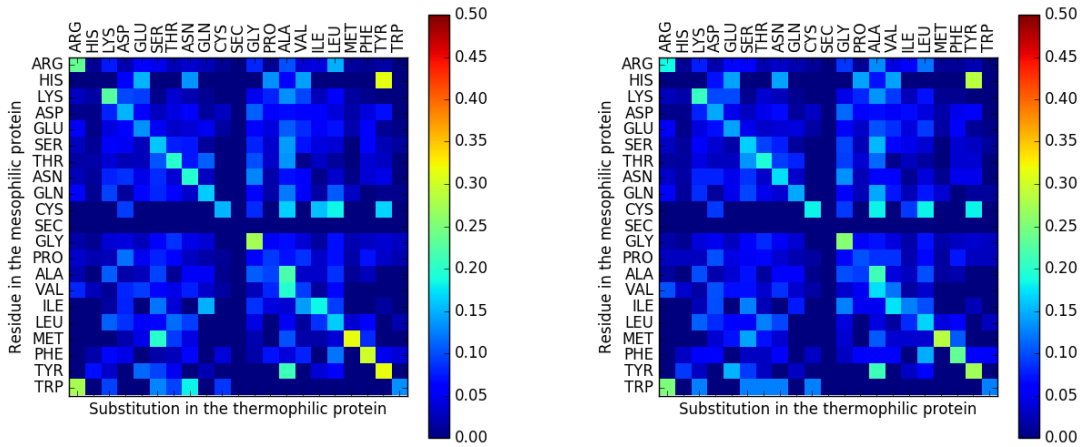
Figure 4.12: The times each amino acid presents a high value of Δs in the database, divided by the number of times that amino acid is present in the database.

In table 4.2 it is possible to see the difference between the thermophilic and mesophilic occurrences divided by the lowest number of occurrences. From the table is possible to better appreciate the differences between the various contact maps. Some residues present the same sign between the different type of maps; in particular Arginine, Threonine, Cysteine, Proline, Valine, Methionine and Tryptophan are preferred in thermophilic structures, while Histidine, Lysine, Glutamine, Phenylalanine and Tyrosine in the mesophilic ones.

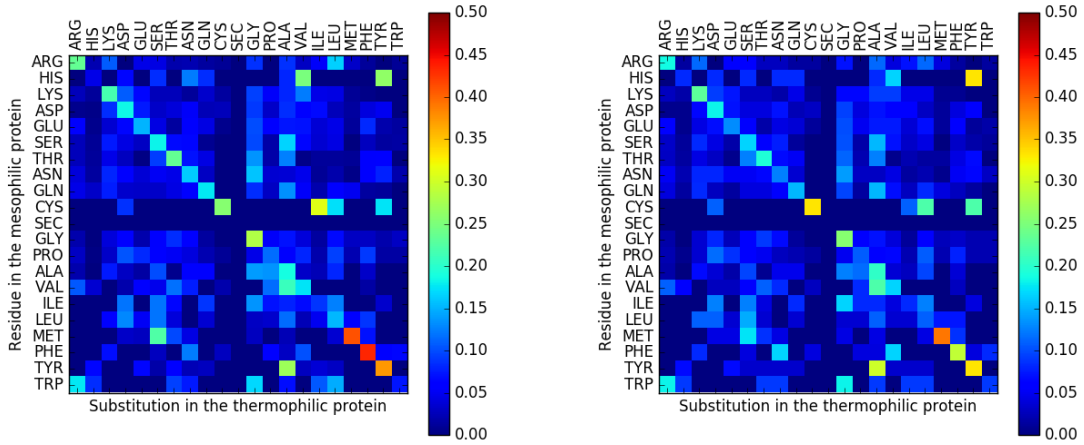
Map	ARG	HIS	LYS	ASP	GLU	SER	THR	ASN	GLN	CYS
Cα	0.10	-0.68	-0.35	-0.27	0.19	0.02	0.11	-0.08	-0.20	2.05
Cβ	0.18	-0.14	-0.29	-0.04	-0.02	-0.04	0.21	0.12	-0.47	5.43
CM	0.38	-1.07	-0.00	0.08	-0.07	0.31	0.43	0.45	-0.00	0.75
CA	0.24	-0.70	-0.27	0.06	-0.02	0.06	0.50	0.28	-0.16	12.57
Map	GLY	PRO	ALA	VAL	ILE	LEU	MET	PHE	TYR	TRP
Cα	-0.12	0.17	-0.27	0.29	-0.06	-0.02	1.44	-0.50	-1.02	0.11
Cβ	-0.09	0.03	-0.44	0.05	0.22	0.37	1.43	-0.27	-0.62	3.88
CM	0.29	0.41	0.21	0.07	0.03	0.27	0.85	-0.33	-0.81	0.41
CA	0.19	0.25	-0.30	0.12	0.20	0.42	1.86	-0.77	-2.77	inf

Table 4.2: The differences between thermophilic and mesophilic maps in the number of occurrences of the residues with the highest Δs divided by the lower of the two values. In the map column the kind of contact map is indicated; CM is an abbreviation for center of mass and CA for closest atoms.

To check whether there are in our database some preferred substitutions between residues in thermophilic maps that present a high value of Δs, and therefore can be considered the weak points of the protein chain, to different kinds of residue in the mesophilic ones the substitution matrices. The substitution matrix is a 21×21 matrix whose rows represent the mesophilic residues and the columns the thermophilic ones. For every mesophilic residue that presents a high value of Δs, c is added in the cell on the row of that residue and on the column of the corresponding mesophilic residue. In figures 4.13 and 4.14 it is possible to observe the substitution matrices for all kinds of contact map. In the figures, the matrices have been normalised by rows, meaning that the sum of each row is equal to 1. There are two matrices of each type of map, since two choices have been made as the value of c. For the matrices on the right $c = 1$, while for the ones on the left c is weighted with the Δs value of the thermophilic protein.

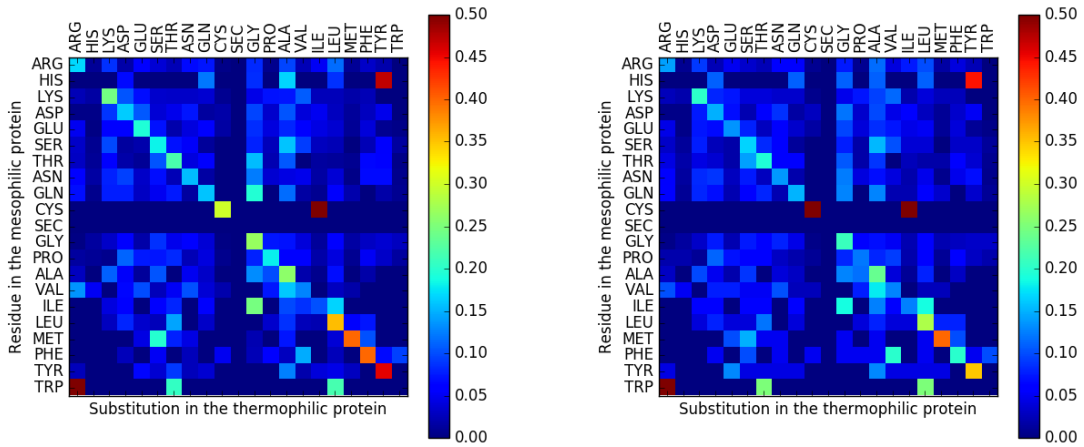


(a) Cα

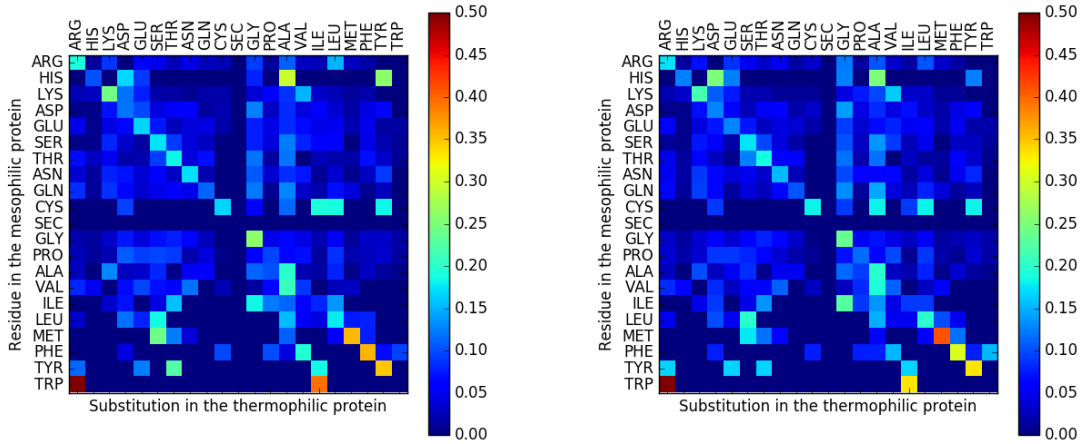


(b) Cβ

Figure 4.13: Occurrences of substitutions in the database for residues in the mesophilic proteins with high values of Δs for Cα and Cβ contact maps, normalized by lines. The maps on the *left* have been weighted with the magnitude of the Δs of the thermophilic residue.



(a) Center of mass



(b) Closest atoms

Figure 4.14: Occurrences of substitutions in the database for residues in the mesophilic proteins with high values of Δs for center of mass and closest atoms contact maps, normalized by lines. The maps on the *left* have been weighted with the magnitude of the Δs of the thermophilic residue.

5

Conclusions

The aim of this thesis was to study with a network approach the structure of thermophilic and mesophilic proteins, trying to find visible differences that would allow to discriminate thermostable proteins from the other group. The method chosen to perform this investigation allowed to see the 3D protein structure as a 2D matrix. On this matrix, the Protein Contact Map, it is possible to perform interesting analysis, in particular it was chosen to focus on the Laplacian and its spectrum as it is strictly related with normal modes. In fact, the *hypothesis of equivalent states*, that proposes an explanation for the higher thermostability of the extremoximes that maintain their functionality at high temperature, states that the thermophilic proteins are more rigid at room temperature than their mesophilic counterparts; they become more flexible as the temperature rises thanks to the thermal vibrations, allowing them to reach their maximal activity at temperature higher than the melting one for the mesophilic proteins. The first step, in order to perform those analyses, was to construct a befitting database of couples of proteins, one thermophilic and the other one mesophilic.

The database characterisation

A thorough search in the literature for couples of homologous thermophilic and mesophilic couples led to a set of 447 pairs. The data of the proteins were retrieved from the PDB archive, as PDB files. The files in this set were then analysed and, according to the quality parameters that were needed in order to obtain meaningful contact maps, many of them were discarded. The final database on which the analyses were performed consisted of 130 proteins, 65 pairs. The quality check that removed the highest number of couples from the database was the number of missing residues in the file. In fact, with the X-ray crystallography, the method used to obtain the proteins structures, it may happen that the atoms of some residues are

not located in the reconstructed model. Skipping a residue in this case would mean to remove a node in the network, eliminating a row and a column in the contact map; for this reason it was decided to keep only the proteins that had no missing residues, lowering the number of suitable pairs with this decision alone to 89.

The analysed database with its 65 couples is a well characterised set of thermophilic and mesophilic proteins, that comply with high quality requirements. These high quality PDB files selected for the database should adhere closely to the physical protein structures, limiting the errors in defining the distances between residues, and therefore enhancing the quality of the analyses performed.

The downside of this selection is that many proteins were discarded, reducing the statistics. For future works, it would be ideal to enlarge the database, possibly constructing it using the criteria that Glyakina [19] or Taylor [42] used to create their database starting from an up-to-date PDB archive. The PDB archive is constantly enriched by new entries and improved versions of proteins that are already there. Having a bigger database could give the possibility to divide the couples in subsets depending on their evolutionary history, which plays a central role in the differences between thermophilic and mesophilic homologues, and would provide more confidence in the results obtained.

The analyses and the Laplacian spectrum

Two different kind of analyses were performed on the database. The first used more "traditional" observables; for instance, the distances between residues and the Contact Order. The second one, instead, took advantage of the network approach and focused on the proprieties of the graph Laplacian.

The first set of observables provided some disappointing results, since it is not possible to discriminate the thermophilic and mesophilic proteins. The distance matrices, obtained with four different rules to define the position of the residue, are very similar one another, producing comparable distance histograms. It is not possible to say therefore that the thermophilic are packed closer together than the mesophiles. The thresholds to consider if two residues are in contact are individuated and chosen from the peaks in the distance histograms. The total number of contacts obtained is approximately the same for the thermophilic and mesophilic proteins, as is the same the frequency of contacts in the diagonals of the maps. Other two parameters are calculated to deepen the knowledge of the contact distribution: the Contact Order and the Long Range Contact Order. This two quantities have values in the same range for the thermophilic and the mesophilic chains and the two distributions are not separable.

The second part of the analysis focused on the Laplacians and the normalised Laplacian of the contact maps. At a first look, the thermophilic and mesophilic spectra look very similar and, even if some of the couples show some variation, this is not always in favour of the thermophilic or always the mesophilic one. It was then

calculated the percentage of couples whose thermophilic graph Laplacian has the sum of the first or last ten eigenvalues higher than the mesophilic. For the last eigenvalues, associated with high frequency, a high variation of results was obtained, while for the sum of the first eigenvalues, associated with low frequencies, the outcomes were more consistent. This is a promising result, as those values are higher than 50% for all the maps but the center of mass type. What is more, for the Cα contact maps it rises up to almost 67%. This percentage can be compared with a value found in literature where, studying the normal modes of couples of thermophilic and mesophilic proteins using the Cα positions to approximate each residue, it reported to be possible to discriminate correctly the 59% of the proteins. The criticism of this result is that there is no error bar associated with it. To further explore this path as a way to discriminate thermophilic and mesophilic proteins, a null model should be devise in order to obtain a confidence interval.

To deepen the analysis on the vibrational proprieties of the proteins, for every residue in the chain Δs, a value of a quantity proportional to its vibration's magnitude, was calculated. Once more the distributions of these values are very similar between thermophilic and mesophilic maps, but the idea this time was to look at the kind of residue that was linked to the vibration. The weak points of the chain were pinpointed as the residues whose value of Δs was higher than the average plus one standard deviation of the Δs values of that protein. In this way it was possible to observe that between those some kind of residues are preferred by the thermophilic and others by the mesophilic. In particular, once normalised by the occurrency of those amino acids in the thermophilic and mesophilic groups, more Arginine, Threonine, Cysteine, Proline, Valine, Methionine and Tryptophan are found in highly flexible spots for the thermophilic structures, while more Histidine, Lysine, Glutamine, Phenylalanine and Tyrosine for the mesophilic ones. It is possible to speculate that this result could be an interesting indication for protein engineers when they devise a thermostable proteins, as it could indicate what kind of residue to place in the weak points of the structures.

Future works

This thesis started from the construction of a high quality database and aimed to provide some insight on the structure strategies of the thermophilic proteins to increment their thermostability. The results are promising, since it was found that there is a variation in the Laplacian spectrum, in particular in the smallest eigenvalues, that allows to discriminate almost the 70% of the proteins. This value is still far from telling apart all the cases, and the possible explanations for this are two. Either the parameter is a good one for the discrimination, but the accuracy in the contact map is too low or either the parameter itself has a limit. To explore the first hypothesis, the model could be enhanced, for example by weighting the networks, with weights proportional to the inverse of the residues' distance or to

the interaction energy between the two residues. But the second one should not be ignored, as the greater rigidity of the thermophilic proteins at room temperature is still debated. It was reported that single point mutations that lead to enhance unequivocally the thermostability of the protein were not all linked to an increase of rigidity [42].

Nonetheless, starting from what it was learned from this thesis, more work should be done. From the obtained results it is possible to further seek a pattern of differences between the two protein groups by using all the obtained parameter as input to a Neural Network, or other machine learning algorithm. It is possible that the distinction between thermophilic and mesophilic can be seen in a high dimensional space and the various parameter obtained have to be put together to find it.

What is more, to add some new analysis, the residues in the protein chains could be to grouped with a spectral clustering. Interesting would be to look at the dimensions of the clusters and their composition, looking for a favourite combination of amino acids and at the probability of, given one residue, finding another in the same cluster. As it has been seen that the kind of residue under high vibrational stress in the chain has different occurrences between thermophilic and mesophilic proteins, it is plausible that one can find some interesting results also from the composition of the clusters.

Appendices

A

Couples of proteins

The following tables give details about the list of couples of proteins from which the proteins in the analysed database were selected, as described in section 3.2. In table A.1 a list of all the PDB IDs of all the couples of thermophilic and mesophilic proteins found in the literature, whose assortment is illustrated in 3.1 can be found. In addition to this, table A.2 presents more details on the couples that were not recovered from the two big databases by Glyakina [19] and Taylor [42], but from various other papers.

	T ID	M IDS		T ID	M IDS		T ID	M IDS
1.	1a1s.A	1dxh.A	2.	1a2z.A	1aug.A	3.	1a8h.A	1f4l.A
4.	1a8l.A	1hyu.A	5.	1adj.A	1kmm.A	6.	1aip.A	1efu.A
7.	1aj8.A	1a59.A	8.	1amu.A	1mdb.A	9.	1anu.A	1g1k.A
10.	1aoh.A	1g1k.A	11.	1ati.A	1g5h.A	12.	1b04.A	1ta8.A
13.	1b4a.A	1f9n.E	14.	1b4b.A	1xxa.A	15.	1b78.A	1k7k.A
16.	1baw.A	7pcy.A	17.	1bdm.B	1b8p.A	18.	1bmd.A	4mdh.A
19.	1bq8.A	1iro.A	20.	1bqc.A	1a3h.A	21.	1brw.B	2tpt.A
22.	1bxb.A	1xlh.A	23.	1bxb.A	2gyi.A	24.	1bxb.A	5xin.B
25.	1bxy.A	1nwy.X	26.	1c3r.A	1t64.A	27.	1caa.A	8rxn.A
28.	1ciu.A	1cdg.A	29.	1ciu.A	1cdg.A	30.	1cz3.A	1qzf.A
31.	1d3u.B	1c9b.A	32.	1dd3.A	1ctf.A	33.	1e19.A	1b7b.A
34.	1ebd.A	1lvl.A	35.	1eft.A	1efu.A	36.	1eg5.A	1p3w.A
37.	1eh1.A	1is1.A	38.	1ej2.A	1kaq.D	39.	1ep0.A	1oi6.A
40.	1ep0.A	1dzt.A	41.	1ep0.A	1upi.A	42.	1ewq.A	1w7a.A
43.	1eys.M	1dxr.M	44.	1eys.C	1prc.C	45.	1eys.H	1prc.H
46.	1f5j.A	1igo.A	47.	1f5s.A	1l8l.B	48.	1fc3.A	1lq1.C

	T ID	M IDS		T ID	M IDS		T ID	M IDS
49.	1fnm.A	1zm9.A	50.	1fsz.A	1ofu.A	51.	1fxq.A	1g7v.A
52.	1g0h.A	1imb.A	53.	1g2w.A	1iye.A	54.	1g5c.A	1ylk.A
55.	1g61.A	1g62.A	56.	1g9x.A	1g6h.A	57.	1gbg.A	2ayh.A
58.	1gd7.A	3ers.A	59.	1gd7.A	1euj.A	60.	1gd9.A	1j32.A
61.	1geq.B	1rd5.A	62.	1gku.B	1mw8.X	63.	1go3.M	1y14.D
64.	1gt6.A	4tgl.A	65.	1gtd.A	1t4a.A	66.	1gtf.A	1wap.A
67.	1h0b.A	2nlr.A	68.	1h1n.A	1egz.C	69.	1h1n.A	1edg.A
70.	1h98.A	7fd1.A	71.	1hbn.B	1e6y.B	72.	1hh2.P	1k0r.A
73.	1hjz.A	1spv.A	74.	1hqk.A	1rvv.1	75.	1hv8.A	1xti.A
76.	1hvx.A	1e40.A	77.	1hvx.A	1bli.A	78.	1hvx.A	1e43.A
79.	1i1x.A	1qnr.A	80.	1i1x.A	1fob.A	81.	1i2s.A	1ylp.A
82.	1i5f.A	1mjc.A	83.	1i6m.A	1yi8.B	84.	1im5.A	1j2r.A
85.	1inl.C	1iy9.A	86.	1ipd.A	1cm7.A	87.	1iq0.A	1f7u.A
88.	1iq8.A	1wke.A	89.	1iqp.A	1sxj.B	90.	1iqr.A	1owl.A
91.	1iqr.A	1tez.B	92.	1iqr.A	1dnp.A	93.	1iqz.A	1fca.A
94.	1iua.A	1cku.A	95.	1iug.A	1vjo.A	96.	1iv3.A	1h48.A
97.	1iw7.A	1bdf.A	98.	1ixr.A	1d8l.A	99.	1iy2.A	1lv7.A
100.	1iz6.A	1xtd.A	101.	1iz9.A	2cmd.A	102.	1iz9.A	4mdh.B
103.	1iz9.A	1civ.A	104.	1j0a.A	1f2d.A	105.	1j1y.A	1psu.A
106.	1j2p.A	1ryp.G	107.	1j33.A	1l4n.A	108.	1j33.A	1l5o.A
109.	1j3b.A	1oen.A	110.	1j3l.D	1nxj.A	111.	1j3n.A	2buh.A
112.	1j3n.A	1e5m.A	113.	1j3n.A	1oxh.A	114.	1j3n.A	1b3n.A
115.	1j6o.A	1xwy.A	116.	1j6u.A	1gqy.B	117.	1jbm.A	1n9r.G
118.	1jg2.A	1i1n.A	119.	1ji1.A	1ea9.D	120.	1jji.A	1lzl.A
121.	1jnr.A	1nek.A	122.	1ka9.H	1ox6.A	123.	1kei.A	1esp.A
124.	1ki9.A	1kht.A	125.	1kij.A	1ei1.A	126.	1kkj.A	1eqb.B
127.	1kl1.A	1dfo.A	128.	1kl1.A	1bj4.A	129.	1kl1.A	1eqb.A
130.	1km2.A	1eix.A	131.	1knv.A	1cfr.A	132.	1kor.B	1k92.A
133.	1ktp.A	1jbo.A	134.	1ku0.A	1tah.A	135.	1ku0.A	1ex9.A
136.	1l0w.A	1il2.A	137.	1l0w.A	1eqr.B	138.	1l0w.A	1c0a.A
139.	1l1j.A	1te0.A	140.	1l8q.A	1j1v.A	141.	1ldn.A	1lth.R
142.	1lf6.A	1ulv.A	143.	1lfp.A	1mw7.A	144.	1lk5.A	1m0s.A
145.	1lnf.E	1bqb.A	146.	1lnf.E	1npc.A	147.	1lnq.A	1id1.A
148.	1loj.A	1n9r.A	149.	1lqy.A	1g27.B	150.	1lqy.A	1lm4.B
151.	1lss.A	1lsu.A	152.	1lvw.A	1mp3.A	153.	1lxn.A	1lxj.A
154.	1m4y.A	1e94.A	155.	1m5h.A	1m5s.A	156.	1mgt.A	1sfe.A
157.	1miw.A	1ou5.A	158.	1mkm.A	1tf1.A	159.	1mow.A	1g9z.A
160.	1mp9.A	1qna.B	161.	1mqq.A	1gqi.A	162.	1mro.C	1e6y.C
163.	1mro.A	1e6y.A	164.	1mtp.A	1lq8.E	165.	1my6.A	1qnn.D

	T ID	M IDS		T ID	M IDS		T ID	M IDS
166.	1my6.A	1isa.B	167.	1my6.A	1dt0.C	168.	1mz4.A	1f1c.A
169.	1n1q.A	1jig.A	170.	1n97.A	1izo.A	171.	1n97.A	1bu7.A
172.	1nbc.A	1g43.A	173.	1nj1.A	1nyr.A	174.	1nog.A	1rty.A
175.	1nox.A	1vfr.A	176.	1nox.A	1icu.A	177.	1nrf.A	1xa1.B
178.	1nvt.A	1npd.A	179.	1ny5.A	1pey.A	180.	1ny5.A	1ojl.B
181.	1nyk.A	1g8k.D	182.	1nz0.A	1a6f.A	183.	1o0w.A	2a11.A
184.	1o12.A	2vhl.B	185.	1o12.A	2vhl.B	186.	1o1x.A	1nn4.B
187.	1o20.A	1vlu.A	188.	1o4u.B	1qpo.A	189.	1o4v.A	1xmp.C
190.	1o54.A	1i9g.A	191.	1o5z.A	1jbw.A	192.	1o6d.A	1ns5.A
193.	1oao.A	1jqk.D	194.	1obr.A	1m4l.A	195.	1ode.A	1dbf.A
196.	1odk.B	1ov6.C	197.	1odk.B	1vhw.A	198.	1odl.A	1k9s.A
199.	1oi7.A	1jkj.A	200.	1omo.A	1x7d.A	201.	1onl.A	1dxm.A
202.	1ov8.A	1cuo.A	203.	1oy5.A	1p9p.A	204.	1oz9.A	1xm5.A
205.	1p1l.A	1osc.A	206.	1p6r.A	1okr.B	207.	1pg5.B	1nbe.B
208.	1phn.A	1cpc.A	209.	1phn.B	1cpc.B	210.	1php.A	1hdi.A
211.	1php.A	1qpg.A	212.	1pvt.A	1gt7.A	213.	1pyb.A	3ers.A
214.	1pyb.A	3ers.X	215.	1pzn.A	1xu4.A	216.	1q6w.A	1iq6.A
217.	1q7z.A	1lt8.A	218.	1qdl.B	1i7q.B	219.	1qho.A	1cxh.A
220.	1qvr.A	1r6b.X	221.	1qvr.A	1khy.D	222.	1qyp.A	1tfi.A
223.	1qzt.A	1td9.A	224.	1r0r.E	1sbh.A	225.	1r2z.A	1pjj.A
226.	1r3e.A	1k8w.A	227.	1rbl.A	1rbo.B	228.	1rfk.B	1czp.A
229.	1rfz.A	1y9i.A	230.	1ril.A	1jxb.A	231.	1ril.A	2rn2.A
232.	1rlk.A	1q7s.A	233.	1rq0.B	1zbt.A	234.	1rqg.A	1qqt.A
235.	1rrs.A	1wef.A	236.	1rvg.B	1gvf.B	237.	1rwz.A	1sxj.G
238.	1rxv.A	1ul1.Z	239.	1s4e.F	1pie.A	240.	1sau.A	1yx3.A
241.	1sei.A	1s03.H	242.	1sei.A	1s03.G	243.	1sg9.A	1t43.A
244.	1sj1.A	1fxd.A	245.	1sky.B	1w0j.A	246.	1snn.A	1k4i.A
247.	1su7.A	1jqk.A	248.	1t7l.A	1u1h.A	249.	1t8h.A	1rw0.A
250.	1tf4.A	1kfg.B	251.	1tf4.A	1g87.B	252.	1thl.A	1npc.A
253.	1thm.A	1ic6.A	254.	1thm.A	1yjb.A	255.	1thm.A	2prk.A
256.	1tib.A	1uza.B	257.	1tib.A	1tgl.A	258.	1tib.A	1lgy.B
259.	1tig.A	2ife.A	260.	1til.F	1auz.A	261.	1tml.A	1cb2.A
262.	1tml.A	1qk0.A	263.	1tqg.A	1i5n.C	264.	1tqh.A	1auo.B
265.	1tux.A	1ta3.B	266.	1twi.A	1hkv.A	267.	1tzv.A	1eyv.A
268.	1u0l.A	1t9h.A	269.	1u1i.A	1p1j.A	270.	1u4b.A	2kfn.A
271.	1u9c.A	4qyx.A	272.	1uar.A	1orb.A	273.	1uay.A	1geg.B
274.	1uay.A	1gee.A	275.	1uay.A	1fmc.A	276.	1uay.A	1k2w.A
277.	1ub3.A	1ktn.A	278.	1ub7.A	1hnj.A	279.	1udd.A	1yaf.A
280.	1udn.A	1r6l.A	281.	1udx.A	1lnz.A	282.	1uek.A	1oj4.A

	T ID	M IDS		T ID	M IDS		T ID	M IDS
283.	1uf9.A	1viy.A	284.	1ufy.A	1dbf.A	285.	1ug6.A	1qox.A
286.	1ug6.A	1cbg.A	287.	1ug6.A	4pbg.A	288.	1ug6.A	1tr1.A
289.	1ugs.A	1ahj.A	290.	1ui0.A	1mug.A	291.	1uir.B	1iy9.B
292.	1uir.B	1xj5.A	293.	1uiy.A	1dci.A	294.	1uj5.A	1m0s.B
295.	1ukk.A	1nye.A	296.	1ukw.A	3mdd.A	297.	1ulq.A	1m3k.A
298.	1ulu.B	1qsg.A	299.	1ulz.A	1dv1.A	300.	1umd.B	1dtw.B
301.	1umd.B	1x7y.B	302.	1umd.A	1qs0.A	303.	1up7.A	1u8x.X
304.	1urd.A	3mbp.A	305.	1uso.A	1dco.C	306.	1usy.E	1nh8.A
307.	1uxx.X	1uyx.B	308.	1uzb.A	1ag8.B	309.	1uzb.A	1cw3.A
310.	1uzb.A	1bxs.A	311.	1uzb.A	1a4s.A	312.	1v1a.A	1bx4.A
313.	1v1a.A	1gqt.D	314.	1v2z.A	1r5q.A	315.	1v30.A	1xhs.A
316.	1v3w.A	1xhd.A	317.	1v47.A	1i2d.A	318.	1v4n.A	1cb0.A
319.	1v4v.A	1o6c.B	320.	1v5x.A	1pii.A	321.	1v6s.A	16pk.A
322.	1v6s.A	1hdi.A	323.	1v6t.A	1xw8.A	324.	1v7c.A	1pwh.A
325.	1v7c.A	1d6s.A	326.	1v8f.A	1mop.A	327.	1v8f.A	1iho.A
328.	1v8m.A	1mqw.A	329.	1v8q.A	1y69.U	330.	1v8z.A	1a5a.B
331.	1v93.A	1b5t.A	332.	1v93.A	1b5t.B	333.	1v9c.B	1f2v.A
334.	1vbi.A	1z2i.A	335.	1vc4.A	1pii.A	336.	1vc4.A	1q6l.A
337.	1vco.A	1s1m.B	338.	1ve1.A	1y7l.A	339.	1ve2.B	1s4d.B
340.	1vf5.A	1q90.B	341.	1vf5.C	1q90.A	342.	1vim.D	1m3s.B
343.	1vjr.A	1ys9.A	344.	1vkn.A	1xyg.A	345.	1vku.A	1t8k.A
346.	1vkz.A	1gso.A	347.	1vl1.A	1y89.A	348.	1vl4.A	1vpb.A
349.	1vla.A	1ml8.A	350.	1vlg.A	1eum.A	351.	1vlh.C	1qjc.A
352.	1vlj.A	1oj7.A	353.	1vlq.A	1l7a.A	354.	1vm7.B	1rkd.A
355.	1vma.A	1fts.A	356.	1vmd.B	1s89.B	357.	1vp5.A	1vbj.A
358.	1vpa.A	1i52.A	359.	1vph.A	1xbf.A	360.	1vpk.A	1ok7.A
361.	1vpq.A	1vpy.A	362.	1vq0.B	1vzy.A	363.	1vrg.A	1on3.A
364.	1w2i.A	2acy.A	365.	1wa3.A	1eua.A	366.	1wdv.A	1dbu.A
367.	1we3.O	1pcq.O	368.	1we3.A	1pcq.A	369.	1wek.A	2a33.B
370.	1whi.A	1ffk.H	371.	1whi.A	1jj2.J	372.	1wkc.A	1sbq.A
373.	1wki.A	1y69.K	374.	1wls.A	1nns.A	375.	1wos.A	1yx2.B
376.	1wp5.A	1zi0.B	377.	1wub.A	1y0g.A	378.	1ww1.A	1y44.B
379.	1wwr.C	1p6o.A	380.	1wx0.A	1l6w.A	381.	1wy5.A	1ni5.A
382.	1x87.A	1uwl.A	383.	1xaa.A	2ayq.A	384.	1xaa.A	1cm7.A
385.	1xex.B	1w1w.B	386.	1xgs.A	1r58.A	387.	1xhk.A	1rr9.D
388.	1xi3.A	2tps.A	389.	1xi8.A	1g8l.A	390.	1xqu.A	1kpf.A
391.	1xrg.A	1oni.A	392.	1xtt.A	1bd3.A	393.	1xx7.A	2paq.B
394.	1xx7.A	2paq.B	395.	1y1l.A	1ljl.A	396.	1y4y.A	2hpr.A
397.	1y51.A	1sph.A	398.	1y80.A	1bmt.A	399.	1ybx.A	1pug.A

	T ID	M IDS		T ID	M IDS		T ID	M IDS
400.	1ycg.A	1e5d.A	401.	1yfz.A	1tc1.B	402.	1ykf.A	1jqb.A
403.	1yna.A	1xnb.A	404.	1yna.A	1enx.A	405.	1yna.A	1pvx.A
406.	1yna.A	1ree.A	407.	1yna.A	1c5h.A	408.	1yna.A	1bk1.A
409.	1ynr.A	351c.A	410.	1yya.A	1tph.1	411.	1z82.A	1n1e.A
412.	1z85.B	1nxz.A	413.	1zh8.A	1h6d.B	414.	1zin.A	1aky.A
415.	1zin.A	1e4y.A	416.	1zin.A	1ank.A	417.	1zin.A	1zak.A
418.	1zin.A	1s3g.A	419.	1zip.A	1p3j.A	420.	2a61.A	1s3j.A
421.	2bj7.A	1q5y.A	422.	2bm3.A	1qzn.A	423.	2bty.A	2buf.A
424.	2bty.A	2buf.A	425.	2cev.A	1t4t.A	426.	2cv4.A	1prx.A
427.	2cv4.A	1prx.A	428.	2hax.A	1csp.A	429.	2ng1.A	1fts.A
430.	2pfk.A	3pfk.A	431.	2pjr.A	1uaa.A	432.	2prd.A	1ino.A
433.	2prd.A	1i40.A	434.	2prd.A	1obw.C	435.	2prd.A	1ypp.B
436.	2tlx.A	1bqb.A	437.	2tlx.A	1npc.A	438.	2tlx.A	1u4g.A
439.	2ts1.A	1x8x.A	440.	3hpd.A	1c3q.A	441.	3mds.A	1gv3.A
442.	3mds.A	1vew.A	443.	3mds.A	1vew.B	444.	3mds.A	1qnm.A
445.	3pva.A	2bjf.A	446.	3tgl.A	1lgy.A	447.	4pfk.A	1pfk.A

Table A.1: All the 894 PDB IDs of the proteins considered for the creation of the database: the list from Glyakina et al. [19], the list from Taylor and Vaisman [42] and the other couples found in literature, whose more detailed information can be found in the next table A.2. For every couple, the first one is the ID of the thermophilic protein followed by the chain , followed by the mesophilic.

Protein	Identity e Similarity	Cited in	T/M	PDB ID	Length	Resolution
Cold-shock protein	81.6% 89.5%	[7]	T	**2HAX**	66	1.29 Å
			M	1CSP	67	2.45Å
C-phycocyanin (beta subunit)	76.2% 91.9%	[41]	T	**1PHN.B**	162	1.65 Å
			M	1CPC.B	162	1.66 Å
Adenylate Kinase	75.0% 88.7%	[2]	T	**1ZIP**	217	1.85 Å
			M	1P3J	217	1.9 Å
C-phycocyanin (alfa subunit)	73.5% 85.2%	[41]	T	**1PHN.A**	162	1.65 Å
			M	1CPC.A	162	1.66 Å
Neutal Proteases	72.9% 83.9%	[7] and [45]	T	**1THL**	317	1.7 Å
			M	1NPC	317	2.0 Å
Cyclodextrin glycosyltransferase	69.7% 83.3%	[41]	T	**1CIU**	683	2.3 Å
			M	1CDG	686	2.0 Å
HPr proteins	69.0% 85.0%	[37]	T	**1Y4Y**	87	2.0 Å
			M	2HPR	87	2.0 Å
Rubredoxin	66.7% 78.4%	[7], [41] and [37]	T	**1CAA**	53	1.8 Å
			M	8RXN	52	1.0 Å
Elongation factor	63.8% 74.8%	[7]	T	**1EFT**	405	2.5 Å
			M	1EFU	363	2.5 Å
Xylanase I	59.8% 74.6%	[41]	T	**1YNA**	194	1.55 Å
			M	1ENX	190	1.5 Å
Phosphofructo Kinases	57.3% 75.7%	[7] and [45]	T	**2PFK**	301	2.4 Å
			M	3PFK	319	2.4 Å
RibonucleaseH 2RN2	56.3% 72.2%	[7]	T	**1RIL**	147	2.8 Å
			M	2RN2	155	1.48 Å
Triacylglycerol acylhydrolase	56.0% 70.7%	[41]	T	**3TGL**	265	1.9 Å
			M	1LGY	265	2.2 Å

Malate dehydrogenase	54.4% 70.4%	[7] and [41]	T	**1EFT**	405	2.5 Å
			M	4MDH	334	2.5 Å
Ribonuclease H	54.2% 70.1%	[37]	T	**1RIL**	147	1.6 Å
			M	1JXB	152	2.8 Å
Manganase superoxide dismutase	52.2% 62.3%	[41]	T	**3MDS**	203	1.8 Å
			M	1VEW	205	2.1 Å

Table A.2: In the table sixteen pairs of proteins are listed. Letters T/M indicate whether the protein is thermophilic (T) or mesofilic (M). The parameters *Identity* and *Similarity* are referring to the results obtained by the jFATCAT alignment tool on the PDB website[4].

B

Couples of proteins in the analysed database

In table B.1 a list of all the proteins in the database is reported. Alongside the PDB ID, there are information on the length of the selected chain of each protein, the values of the Identity, MaxSub and TM-score for each couple, the Resolution and R-factor of every file and the number of missing Cα and Cβ in every structure.

Protein	PDB ID	Chain	Length	Identity	MaxSub	TM-score	Resolution	R-factor
Aminopeptidase	T: 1xgs	A	295	0.373	0.69174	0.93792	1.75	0.187
	M: 1r58	A	369				1.9	NULL
βamylase	T: 1tml	A	286	0.234	0.61404	0.87976	1.8	0.184
	M: 1qk0	A	363				2.1	0.181
Cellulose degradation	T: 1nbc	A	155	44.8	0.88247	0.93812	1.75	0.193
	M: 1g43	A	160				2.2	0.19
Cohesin	T: 1anu	A	138	32.3	0.83323	0.88976	2.15	0.197
	M: 1g1k	A	143				2.0	0.23
Electron transport	T: 1PHN	B	171	76.0	0.96212	0.9783	1.65	0.183
	M: 1CPC	B	171				1.66	0.181
Electron transport	T: 1sj1	A	66	42.9	0.90087	0.85757	1.5	0.194
	M: 1fxd	A	57				1.7	NULL
Electron transport	T: 1CAA	A	53	66.7	0.95493	0.92644	1.8	0.178
	M: 8RXN	A	52				1.0	NULL
Electron transport	T: 1iqz	A	81	22.2	0.80218	0.72675	0.92	0.094
	M: 1fca	A	55				1.8	0.143
Electron transport	T: 1iua	A	83	89.2	0.9267	0.94337	0.8	NULL
	M: 1cku	A	85				1.2	0.124
Electron transport	T: 1ynr	A	80	56.2	0.94395	0.96181	2.0	0.171
	M: 351c	A	82				1.6	0.195
Electron transport	T: 1PHN	A	162	73.5	0.9762	0.98669	1.65	0.183
	M: 1CPC	A	162				1.66	0.181
Gene regulation	T: 2HAX	A	66	81.6	0.52941	0.52221	1.29	0.13

Protein	PDB ID	Chain	Length	Identity	MaxSub	TM-score	Resolution	R-factor
	M: 1CSP	A	67				2.45	NULL
Glycosidase	T: 1ciu	A	683	70.2	0.95749	0.99146	2.3	NULL
	M: 1cdg	A	686				2.0	NULL
Hydrolase (Glucanase)	T: 1gbg	A	214	76.6	0.97611	0.98808	1.8	NULL
	M: 2ayh	A	214				1.6	NULL
Hydrolase (Metalloprotease)	T: 1lnf	E	316	73.1	0.94359	0.98296	1.7	NULL
	M: 1npc	A	317				2.0	NULL
Hydrolase (Metalloprotease)	T: 1lnf	E	316	48.1	0.91783	0.95466	1.7	NULL
	M: 1bqb	A	301				1.72	0.176
Hydrolase (Serine Protease)	T: 1thm	A	279	45.1	0.89767	0.93943	1.37	NULL
	M: 1yjb	A	274				1.8	0.177
Hydrolase (Serine Protease)	T: 1thm	A	279	33.8	0.80011	0.86839	1.37	NULL
	M: 2prk	A	279				1.5	NULL
Hydrolase (Serine Protease)	T: 1thm	A	279	33.3	0.79882	0.86893	1.37	NULL
	M: 1ic6	A	279				0.98	0.114
Hydrolase Inhibitor	T: 1THL	A	316	73.1	0.95357	0.98624	1.7	0.162
	M: 1NPC	A	317				2.0	NULL
Hydrolase	T: 2prd	A	174	48.5	0.90102	0.9393	2.0	NULL
	M: 1i40	A	175				1.1	0.09
Hydrolase	T: 1yna	A	193	50.3	0.87904	0.91736	1.55	0.155
	M: 1c5h	A	185				1.55	0.194
Hydrolase	T: 1h1n	A	305	17.8	0.63558	0.76352	1.12	0.1198
	M: 1egz	C	291				2.3	0.179

Protein	PDB ID	Chain	Length	Identity	MaxSub	TM-score	Resolution	R-factor
Hydrolase	T: 1bqc	A	302	17.6	0.66515	0.79506	1.5	0.119
	M: 1a3h	A	300				1.57	0.14
Hydrolase	T: 1yna	A	193	50.8	0.87945	0.91757	1.55	0.155
	M: 1xnb	A	185				1.49	0.165
Hydrolase	T: 1qho	A	686	46.4	0.88092	0.95145	1.7	0.151
	M: 1cxh	A	686				2.41	NULL
Hydrolase	T: 1YNA	A	193	60.3	0.94994	0.97188	1.55	0.155
	M: 1ENX	A	189				1.5	0.193
Hydrolase	T: 2prd	A	174	48.5	0.90602	0.94222	2.0	NULL
	M: 1obw	C	175				1.9	0.176
Hydrolase	T: 2prd	A	174	48.5	0.89347	0.93488	2.0	NULL
	M: 1ino	A	175				2.2	NULL
Hydrolase	T: 1r0r	E	274	71.4	0.97703	0.98942	1.1	0.158
	M: 1sbh	A	274				1.8	0.185
Hydrolase	T: 1f5j	A	199	59.6	0.89619	0.95823	1.8	0.185
	M: 1igo	A	205				2.2	0.208
Hydrolase	T: 1wls	A	316	24.8	0.72559	0.86354	2.16	0.211
	M: 1nns	A	326				1.95	0.13
Hydrolase	T: 1yna	A	193	60.3	0.95119	0.9726	1.55	0.155
	M: 1ree	A	189				1.6	0.181
Hydrolase	T: 2tlx	A	316	73.1	0.94718	0.98415	1.65	0.163
	M: 1npc	A	317				2.0	NULL
Hydrolase	T: 1yna	A	193	87.6	0.98435	0.99509	1.55	0.155

Protein	PDB ID	Chain	Length	Identity	MaxSub	TM-score	Resolution	R-factor
	M: 1pvx	A	194				1.59	0.191
Hydrolase	T: 1h1n	A	305	0.177	0.56428	0.83706	1.12	0.1198
	M: 1edg	A	380				1.6	0.191
Hydrolase	T: 2tlx	A	316	48.1	0.91877	0.955	1.65	0.163
	M: 1bqb	A	301				1.72	0.176
Isomerase	T: 1lk5	A	229	43.6	0.91793	0.95851	1.75	0.191
	M: 1m0s	A	219				1.9	0.179
Kinase	T: 1php	A	394	50.9	0.84089	0.96461	1.65	NULL
	M: 1qpg	A	415				2.4	NULL
Kinase	T: 1php	A	394	45.9	0.84826	0.96312	1.65	NULL
	M: 1hdi	A	413				1.8	0.207
Ligase	T: 1l0w	A	580	49.3	0.72917	0.91262	2.01	0.217
	M: 1c0a	A	585				2.4	0.208
Lyase	T: 1vc4	A	254	0.346	0.46665	0.90004	1.8	0.186
	M: 1pii	A	452				2.0	NULL
Lyase	T: 1j0a	A	325	30.1	0.7885	0.91941	2.5	0.201
	M: 1f2d	A	341				2.0	0.221
Lyase	T: 1v7c	A	351	20.4	0.73594	0.85196	2.0	0.196
	M: 1d6s	A	322				2.3	0.174
Maltose-binding	T: 1urd	A	370	29.9	0.77857	0.89964	1.53	0.206
	M: 3mbp	A	370				1.7	NULL
Methanogenesis	T: 1mro	C	247	57.9	0.93956	0.96576	1.16	0.197
	M: 1e6y	C	246				1.6	0.16

Protein	PDB ID	Chain	Length	Identity	MaxSub	TM-score	Resolution	R-factor
oxidoreductase	T: 1ipd	A	345	50.4	0.8761	0.97308	2.2	NULL
	M: 1cm7	A	363				2.06	0.173
Oxidoreductase	T: 1BMD	A	327	54.2	0.87238	0.94571	1.9	NULL
	M: 4MDH	A	333				2.5	NULL
Oxidoreductase	T: 3MDS	A	203	54.5	0.9019	0.94335	1.8	NULL
	M: 1VEW	A	205				2.1	0.188
Oxidoreductase	T: 3mds	A	203	54.5	0.9047	0.9448	1.8	NULL
	M: 1vew	B	205				2.1	0.188
Oxidoreductase	T: 3mds	A	203	51.9	0.87255	0.90316	1.8	NULL
	M: 1qnm	A	198				2.3	0.199
Oxidoreductase	T: 1iz9	A	327	53.9	0.90513	0.96137	2.0	0.193
	M: 4mdh	B	333				2.5	NULL
Oxidoreductase	T: 1a8l	A	226	0.225	0.31877	0.80672	1.9	0.192
	M: 1hyu	A	521				2.0	0.182
Oxidoreductase	T: 1xaa	A	345	50.4	0.87534	0.97292	2.1	0.157
	M: 1cm7	A	363				2.06	0.173
Oxidoreductase	T: 1iz9	A	327	21.6	0.71655	0.84978	2.0	0.193
	M: 2cmd	A	312				1.87	0.188
Oxidoreductase	T: 1uzb	A	516	29.3	0.80148	0.91151	1.4	0.177
	M: 1a4s	A	503				2.1	0.223
Oxidoreductase	T: 1ykf	A	352	76.9	0.96778	0.99106	2.5	0.215
	M: 1jqb	A	351				1.97	0.219
Phosphotransferase	T: 1zin	A	217	46.2	0.82575	0.88645	1.6	0.173

Protein	PDB ID	Chain	Length	Identity	MaxSub	TM-score	Resolution	R-factor
	M: 1ank	A	214				2.0	0.201
Phosphotransferase	T: 1zin	A	217	65.3	0.89686	0.93772	1.6	0.173
	M: 1s3g	A	217				2.25	0.223
Phosphotransferase	T: 1zin	A	217	45.5	0.84455	0.90697	1.6	0.173
	M: 1e4y	A	214				1.85	0.178
Photosynthesis	T: 1ktp	A	162	99.4	0.9974	0.99858	1.6	0.216
	M: 1jbo	A	162				1.45	0.146
Ribosomal protein	T: 1whi	A	122	41.5	0.81955	0.90304	1.5	0.189
	M: 1jj2	J	132				2.4	0.189
Signal Recognition	T: 2ng1	A	293	34.2	0.70085	0.84566	2.02	0.2
	M: 1fts	A	295				2.2	0.222
Structural Genomics	T: 1lxn	A	96	27.7	0.81265	0.8585	2.3	0.216
	M: 1lxj	A	101				1.8	0.211
Transcription	T: 1i5f	A	66	57.6	0.88821	0.90412	1.4	0.129
	M: 1mjc	A	69				2.0	NULL
Transferase	T: 1m5h	A	297	68.4	0.9666	0.98932	2.0	0.229
	M: 1m5s	A	297				1.85	0.197
Transferase	T: 1v6s	A	390	44.4	0.71624	0.91527	1.5	0.201
	M: 16pk	A	415				1.6	0.188
Transferase	T: 1j3n	A	408	38.7	0.87416	0.94771	2.0	0.21
	M: 2BUH	A	406				1.9	0.217
Transferase	T: 1v6s	A	390	40.9	0.76508	0.93476	1.5	0.201
	M: 1hdi	A	413				1.8	0.207

Protein	PDB ID	Chain	Length	Identity	MaxSub	TM-score	Resolution	R-factor
Unknown function	T: 1n1q	A	149	57.5	0.98041	0.98639	2.2	0.196
	M: 1jig	A	146				1.46	0.186

Table B.1: The list of the couples of proteins in the final database. The letters T/M indicate whether it is a thermophilic or mesophilic protein.

Bibliography

[1] Santiago Alvarez. "A cartography of the van der Waals territories". In: *Dalton Transactions* 42.24 (2013), pp. 8617–8636. ISSN: 1477-9234. DOI: 10.1039/c3dt50599e.

[2] Euiyoung Bae and George N Phillips. "Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases." In: *The Journal of biological chemistry* 279.27 (2004), pp. 28202–8. ISSN: 0021-9258. DOI: 10.1074/jbc.M401865200.

[3] Igor N Berezovsky and Eugene I Shakhnovich. "Physics and evolution of thermophilic adaptation." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.36 (2005), pp. 12742–7. ISSN: 0027-8424. DOI: 10.1073/pnas.0503890102.

[4] H M Berman et al. "The Protein Data Bank." In: *Nucleic acids research* 28.1 (2000), pp. 235–242. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235.

[5] Marco Berrera, Henriette Molinari, and Federico Fogolari. "in the Space of Contact Maps". In: *BMC Bioinformatics* 26 (2003), pp. 1–26.

[6] Steven Butler. "Eigenvalues and structures of graphs". PhD thesis. University of California, San Diego, 2008, p. 89.

[7] Changjun Chen, Lin Li, and Yi Xiao. "All-Atom Conctact Potential Approach to Protein Thermostability Analysis". In: 85.1 (2006), pp. 28–37. DOI: 10.1002/bip.

[8] F. Chung. "Eigenvalues and the Laplacian of a graph". In: *Spectral Graph Theory*. 2006. Chap. 1. ISBN: 978-0821803158. DOI: 10.1080/03081088508817681.

[9] R. Cohen and S. Havlin. *Complex Networks: Structure, Robustness and Function*. Cambridge University Press, 2010.

[10] David C. Demirjian, Francisco Morís-Varas, and Constance S. Cassidy. "Enzymes from extremophiles". In: *Current Opinion in Chemical Biology* 5.2 (2001), pp. 144–151. ISSN: 13675931. DOI: 10.1016/S1367-5931(00)00183-6.

[11] L Di Paola et al. “Protein contact networks: an emerging paradigm in chemistry.” In: *Chemical reviews* 113.3 (2013), pp. 1598–613. ISSN: 1520-6890. DOI: 10.1021/cr3002356.

[12] Reinhard Diestel. *Graph Theory*. 5th editio. 2016.

[13] Russell F Doolittle. “Similar Amino Acid Sequences: Chance Common Ancestry?” In: *Science* 214.October (1981), pp. 149–159.

[14] Jan Drenth. *Principles of Protein X-Ray Crystallography*. Second Edi. 1999.

[15] Jose M Duarte et al. “Optimal contact definition for reconstruction of contact maps.” In: *BMC bioinformatics* 11 (2010), p. 283. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-283.

[16] Ernesto Estrada and Naomichi Hatano. “A vibrational approach to node centrality and vulnerability in complex networks”. In: *Physica A: Statistical Mechanics and its Applications* 389.17 (2010), pp. 3648–3660. ISSN: 03784371. DOI: 10.1016/j.physa.2010.03.030.

[17] Vincent Frappier and Rafael Najmanovich. “Vibrational entropy differences between mesophile and thermophile proteins and their use in protein engineering”. In: *Protein Science* 24.4 (2015), pp. 474–483. ISSN: 1469896X. DOI: 10.1002/pro.2592.

[18] David Freedman and Persi Diaconis. “On the histogram as a density estimator:L 2 theory”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57.4 (1981), pp. 453–476. ISSN: 0044-3719. DOI: 10.1007/BF01025868.

[19] Anna V Glyakina et al. “Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms.” In: *Bioinformatics (Oxford, England)* 23.17 (2007), pp. 2231–8. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btm345.

[20] Thomas Hamelryck and Bernard Manderick. “PDB file parser and structure class implemented in Python”. In: *Bioinformatics* 19.17 (2003), pp. 2308–2310. DOI: 10.1093/bioinformatics/btg299.

[21] D A Hinds and M Levitt M. “Exploring conformational space with a simple lattice model for protein structure.” In: *J Mol Biol* 243.4 (1994), pp. 668–682.

[22] K. Kleppe et al. “Studies on polynucleotides: XCVI. Repair replication of short synthetic DNA’s as catalyzed by DNA polymerases.” In: *Journal of molecular biology* 56.2 (1971), pp. 341–361.

[23] Frances C. Lawyer et al. “High-level expression, purification, and enzymatic characterization of full-length Thermus aquaticus DNA polymerase and a truncated form deficient in 5’ to 3’ exonuclease activity”. In: *Genome Research* 2.4 (1993), pp. 275–287. ISSN: 10889051. DOI: 10.1101/gr.2.4.275.

[24] Guixia Liu et al. “Prediction of contact maps using modified transiently chaotic neural network”. In: *Advances in Neural Networks* (2006), pp. 696–701. ISSN: 16113349.

[25] Kelly S. Lundberg et al. “High-fidelity amplification using a thermostable DNA polymerase isolated from Pyrococcus furiosus.” In: *Gene* 108.1 (1991), pp. 1–6. ISSN: 0378-1119. DOI: 10.1016/0378-1119(91)90480-Y.

[26] Alex C.W. May. “Letter to the Editor Percent Sequence Identity : The Need to Be Explicit”. In: *Structure* 12 (2004), pp. 737–738. DOI: 10.1016/j.str.2004.04.001.

[27] Florence Mingardon et al. “Comparison of family 9 cellulases from mesophilic and thermophilic bacteria”. In: *Applied and Environmental Microbiology* 77.4 (2010), pp. 1436–1442. ISSN: 00992240. DOI: 10.1128/AEM.01802-10.

[28] E. V. Morozkina et al. “Extremophilic microorganisms: biochemical adaptation and biotechnological application (review)”. In: *Prikladnaia biokhimiia i mikrobiologiia* 46.1 (2010), pp. 5–20. ISSN: 0003-6838. DOI: 10.1134/S0003683810010011.

[29] K. Mullis et al. *Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction*. 1986. DOI: 10.1101/SQB.1986.051.01.032.

[30] K B Mullis. *The unusual origin of the polymerase chain reaction*. 1990. DOI: 10.1038/scientificamerican0490-56.

[31] “Protein Data Bank”. In: *Nature New Biology* (1971), 233:223.

[32] G P S Raghava and Geoffrey J Barton Barton. “Quantification of the variation in percentage identity for protein sequence alignments”. In: 7 (2006), pp. 1–4. DOI: 10.1186/1471-2105-7-415.

[33] Jane B. Reece and Neil A. Campbell. *Campbell Biology*. 9th. 2011.

[34] F T Robb and D S Clark. “Adaptation of proteins from hyperthermophiles to high pressure and high temperature.” In: *Journal of molecular microbiology and biotechnology* 1.1 (1999), pp. 101–5. ISSN: 1464-1801.

[35] R K Saiki et al. “Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase.” In: *Science (New York, N.Y.)* 239.4839 (1988), pp. 487–491. ISSN: 0036-8075. DOI: 10.1126/science.2448875.

[36] F Sanger. *The chemistry of insulin*. 1958. (Visited on 11/13/2015).

[37] J Martin Scholtz. “Lessons in stability from thermophilic proteins”. In: Jaenicke 1991 (2006), pp. 1569–1578. DOI: 10.1110/ps.062130306.Life.

[38] N. Siew et al. “MaxSub: an automated measure for the assessment of protein structure prediction quality”. In: *Bioinformatics* 16.9 (2000), pp. 776–785. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/16.9.776.

[39] Jagtar Singh et al. “A critical review on PCR, its types and applications”. In: *International Journal of Advanced Research in Biological Sciences* 1.7 (2014), pp. 65–80.

[40] Karl O. Stetter. “Extremophiles and their adaptation to hot environments”. In: *FEBS Letters* 452.1-2 (1999), pp. 22–25. ISSN: 00145793. DOI: 10.1016/S0014-5793(99)00663-8.

[41] a Szilágyi and P Závodszky. “Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey.” In: *Structure (London, England : 1993)* 8.5 (2000), pp. 493–504. ISSN: 0969-2126.

[42] Todd J. Taylor and Iosif I. Vaisman. “Discrimination and Classification of Thermophilic and Mesophilic Proteins”. In: *International Symposium on Voronoi Diagrams in Science and Engineering* (2007).

[43] B Van den Burg et al. “Engineering an enzyme to resist boiling.” In: *Proceedings of the National Academy of Sciences of the United States of America* 95.5 (1998), pp. 2056–60. ISSN: 0027-8424. DOI: 10.1073/pnas.95.5.2056.

[44] C Vetriani et al. “Protein thermostability above 100 degreesC: a key role for ionic interactions.” In: *Proceedings of the National Academy of Sciences of the United States of America* 95.21 (1998), pp. 12300–5. ISSN: 0027-8424. DOI: 10.1073/pnas.95.21.12300.

[45] M S Vijayabaskar and S. Vishveshwara. “Comparative analysis of thermophilic and mesophilic proteins using Protein Energy Networks.” In: *BMC bioinformatics* 11 Suppl 1 (2010), S49. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-S1-S49.

[46] Duncan J. Watts. “The “New” Science of Networks”. In: *Annual Review of Sociology* 30.1 (2004), pp. 243–270. ISSN: 0360-0572. DOI: 10.1146/annurev.soc.30.020404.104342.

[47] C R Woese and G E Fox. “Phylogenetic structure of the prokaryotic domain: the primary kingdoms.” In: *Proceedings of the National Academy of Sciences of the United States of America* 74.11 (1977), pp. 5088–5090. ISSN: 0027-8424. DOI: 10.1073/pnas.74.11.5088.

[48] Lee W. Yang et al. “Insights into Equilibrium Dynamics of Proteins from Comparison of NMR and X-Ray Data with Computational Predictions”. In: *Structure* 15.6 (2007), pp. 741–749. ISSN: 09692126. DOI: 10.1016/j.str.2007.04.014.

[49] Yang Zhang and Jeffrey Skolnick. “Scoring function for automated assessment of protein structure template quality”. In: *Proteins: Structure, Function and Genetics* 57.4 (2004), pp. 702–710. ISSN: 08873585. DOI: 10.1002/prot.20264.

[50] Yang Zhang and Jeffrey Skolnick. "TM-align: A protein structure alignment algorithm based on the TM-score". In: *Nucleic Acids Research* 33.7 (2005), pp. 2302–2309. ISSN: 03051048. DOI: 10.1093/nar/gki524.