Ca' Foscari
University
of Venice

Master's Degree programme in
Science and Technology of Bio and Nanomaterials

Computational Study of Solvent Effects on the Stability of Native Structures in Proteins

Supervisor
Ch. Prof. Achille Giacometti

Assistant supervisor
Dr. Tatjana Škrbić

Graduand
Emanuele Petretto
Matriculation Number 855446

Academic Year 2016/2017

Contents

Introduction

Water is one of the main components of the all ecosystems and the basis for life-as-we-know-it. Some important properties of water, beside that of the other all simple organic compounds, has been fundamental to the abiogenesis. These are, for example, the wide range temperature in which it can be found in liquid state, the heat capacity, useful for thermoregulation, and the high solvent power that is useful to spread and diffuse solute. Polar character, due by the charge dislocation actuated by the electronegative oxygen atom, the capability to generate hydrogen bonds and the corresponding network, and the self-ionization that allows water to donate or accept H^+ ion, provides to water a large chemical activity. Metabolic pathway reaction are intimately connected to water as solvent, reactant or product. Evolution shaped biomolecule activity and functionality, from the first self-replicating and catalytic RNA molecules to the neurotransmitter receptor associated with the membrane in post synaptic cells, in water environment. Understanding how life could have been without water is one of the most challenging issues in astrobiology. This problem is of considerable interest even on earth, as other solvents might be competitive with water. A canonical example can be the methane (CH_4) composed by the to most common element in the universe: carbon and hydrogen. In this environment, soluted biomolecule do not occur into hydrolysis reactions, allowing a larger chemical stability, but subsequently, this solvent is not able behave as the solvent, reactant or product feature in the water. Second example is provided by protein living in membrane within hydrophobic environment. Focusing on the electrostatic interaction, a completely apolar solvent as methane, could make the intrasolute or solute-solute electrostatic interaction more favorable than in water since this could disrupt these interaction. A possible way to understand this kind of dynamics is study the stability and solubility of molecule in this new solvents. Use proteins to proceed on this exploration could permit to have a multiple point of view to understand the physico-chemical basis of protein. Protein folding and stability are intimate correlated since the native conformation of a protein, capable to provide a wide diversity of function, is even the most stable. Exploring proteins stability in other polar or non-polar solvent such as ethanol and cycloehexane, respectively, have relevance, not only in biology but even in pharmacology, nanotechnology, or industrial chemistry. Numerical simulation plays an important role in the study of biological world as an additional tools alongside the experiment since that is able to investigate a variety of effects that are non-accessible with the experimental methods. Amongst the many numerical approach molecular dynamics has a relevant role due by the ability to predict the general dynamics. In this work we explore the solvation free energy using a dual approach. The first one is a bottom up method, focused on understand the free energy of solvation contribute of single amino acids using a molecular dynamics and thermodynamic integration. We performed such calculation and compared with past studies with positive results. Taking this into account we proposed to validate the solidity of molecular dynamic confronting results with literature and, with the support of the data generated by the thermodynamic integration, study how chemical groups present in the amino acids take part in this process. A direct scale up of this calculation to a full protein is limited by the large computational required , as well as by the unreliability of the numerical precision. This first approach can be used to test the reliability of an alternative approximate approach, that as been used in the second part.

The second approach is top down method, where free energy calculation is not calculated by molecular dynamics but using a morphometric approach able to calculate free energy considering the excluded volume generated by the presence of the protein in the solute and by the hydrogen possible bonds pattern in the system. This method is able to rapidly, respect of the molecular dynamics, the protein stability and is devised by a group at the University of Kyoto whom we have been collaborating with. Due the rapidity of calculation this method is able to compare a native state protein with a wide range of other structures. Rather than compare different biological protein, compare slightly or considerably different geometrical conformation of the same protein to understand how geometrical, and the derived change in H-bonds pattern, contribute change the stability. This study can be actuated in different solvent model. This alternative geometrical structure, defined decoy, take part in the validation of several numerical methods. Generally provided decoy algorithm generate this structure with lightly changing the input structure, since the main intent is compare the studied protein with a distorted one. This thesis work contributes aim to develop a bioinformatic tool capable to build up several decoys with widely distributed topologies and predefined content of secondary structure.

Thesis plan

In the first chapter focuses on protein starting with amino acid description, the secondary structures,protein folding, solubility, and stability in water both in water than other solvents. Second chapter describes numerical methods, in particular molecular dynamics and and Monte Carlo, pointing out the aspect used for this work. Third chapter reports methods and results of the analysis made on amino acids side chain analogs. Analysis points out the validity of data calculated and the detailed analysis given by the thermodynamic integration. Fourth chapter describes the basis of the morphometric methods and an accurate description of the the bioinformatic tool able to generate different geometrical conformation of a studied protein. This chapter is closed with the analysis of the protein structure generated. Last part is dedicated to the conclusion and perspective.

Chapter 1

Proteins

Proteins are the most common polymers in the biological world and, furthermore, these macromolecules exhibit huge diversity in function as a result of their variety in three-dimensional structure and chemical activity. Chemico-physical distinctive features are given by the monomeric forming subunits: *amino acids*. Cells or metabolically inert structures, like viruses, genetic heritage are expressed with the purpose to build up the metabolic machinery which is mainly composed by proteins. Protein are assembled in a specific amino acid sequence by the cell through *transcription*, *maturation* and *traduction* processes (Fig. 1.1). Transcription is the first step of gene expression, in which a particular segment of DNA is transcripted and copied into RNA. Maturation process lead a full functional RNA from a precursor messenger RNA (pre-mRNA). DNA is responsible of storage of the biological information whereas RNA convey, specifically as mRNA, the genetic information to the protein. Translation, finally, is the process in which ribosomes decode the information coded in mRNA by the binding of complementary tRNA anticodon sequences to mRNA codons. tRNAs, the RNA transfer, are specific RNA sequences, with an amino acid physically binded in the structure, that works as adapter from the RNA codons code to the amino acid sequences. Amino acids are chained together by the ribosome machinery into a polypeptide and are released from the ribosome in the cell environment where reach the mature conformation through the protein folding processes. Protein function arise from the specific three-dimensional structure defined by this process. Depending on the length of the protein this process can be molecular driven by molecular chaperones that are proteins capable to assist the covalent folding or unfolding of a single protein or the assembly and disassembly of multiple folded protein subunits.

Protein native state can be divided in different categories according to size and

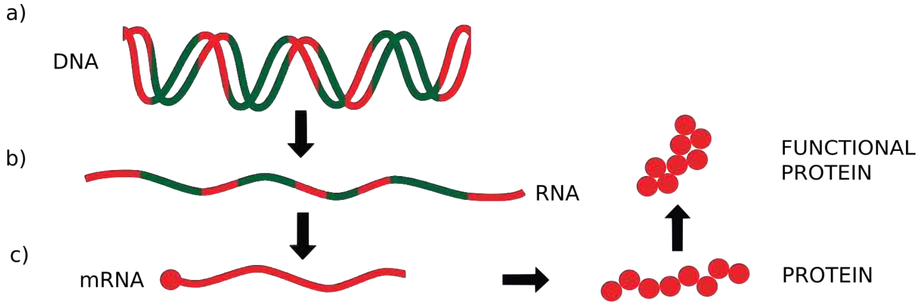


Figure 1.1: Protein synthesis scheme: a) transcription; b) maturation; c) traduction.

shape. This morphologic characteristics vary over the functionality and the working environment: *globular proteins*, are spheroidal shaped and mainly soluble in the physiological condition, where forms colloids in water. Globular proteins assume different roles in the cell organization taking part in biological catalyst of chemical reactions, as enzymes, transmission of informations, regulation biological processes as messengers, structural scaffolds or transporter. *Fibrous proteins* are generally insoluble, usually assembled into bundles that make possible to play the structural and supporting role as in the connecting connective tissue, tendons, and bone matrices. This family includes proteins like keratin, collagen, elastin, and fibroin. Another class is represented by proteins that are able to interact with the biological membranes. This class of protein, the *membrane proteins*, is divided in *integral membrane proteins*, permanently embedded in the bilayer, and the *peripheral membrane proteins* that are associated temporarily. This class include proteins that have different functions as the membrane receptor, like the protein G family, protein able to transport molecules, or adhesion structures able, via a stimulus, to start a biochemical cascade that amplify the signal of the interaction and define a response of the cell.

1.1 Amino acids

Proteins are a single linear polymer chain of amino acids building-blocks. Only 22 amino acid appear in the universal genetic code from the plethora of 500 naturally occurring known molecules. Biologically active amino acids are α-amino acids. The 20 proteinogenic amino acid (Fig 1.3) share a common structure formed by a carbon substituted with a *carboxyl group* $-$COOH, an *amino group* $-\mathrm{NH}_2$, and a characterizing group defined as *side chain* **R**. This group changes amino acid properties by its size, charge and chemical-physical properties. The two remaining, selenocysteine and pyrrolysine, are encoded via variant codons, and occurring rarely in nature. The *pro-*

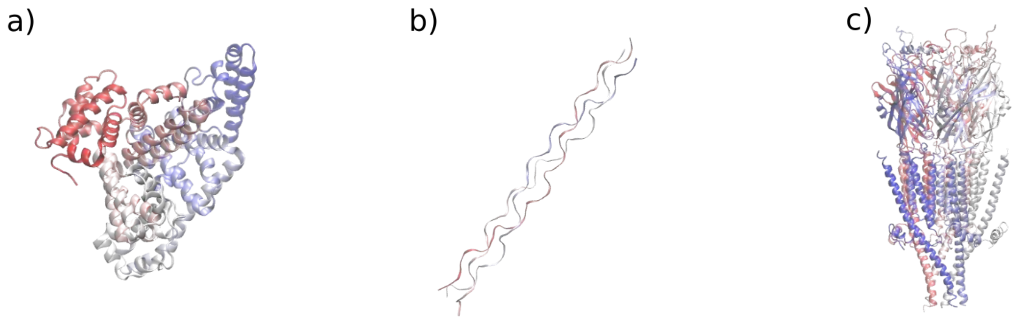


Figure 1.2: Protein morphology: a) globular protein, caprine serum albumin(5ORI); b)Fibrous proteins, a collagen-like protein (1CAG); c) Memebrane protein, structure of serotonin receptor.

line amino acid have a peculiar structure as the side chain **R** bonds the amino group forming a *imino group*. Side chains are mainly composed by other carbon atoms that are nominated β, γ, δ, ϵ-carbon taking as reference the α-carbon. α-carbon is a *chiral center* because the tetrahedral configuration of the orbitals as well as, consequently, the bonds geometry, are substituted with four different groups, hence, molecules like amino acids are optically active and capable to rotate the plane of polarized light. The enatiomers are specified as Fisher's convention for simple sugars by the D,L system. In terms of a living system D and L stereoisomers are absolutely different and, in almost universal way, the biologic amino acid are L stereoisomers. A particular case is carried by the *glycine*, since the side chain is composed by only a hydrogen which makes this amino acid not chiral. The different chemical groups in the amino acids side chains ensure three states: cationic, protoned or a zwitterionic. The zwitterionic state can be found when a molecule, with two or more functional groups, has a positive a negative electrical charge in different regions making the net charge of the entire molecule zero. Under physiological conditions (pH 7) the most common state is the zwitterionic, since the amine group deprotonates the carboxyl acid via a kind of intramolecular acid–base reaction:

$$\mathrm{NH_2RCHCO_2H} \rightleftharpoons \mathrm{NH_3{}^+RCHCO_2{}^-}$$

Amino acid can be categorized in four groups: hydrophobic side chain, polar uncharged side chain, electrically, positive or negative, charged side chain.

Hydrophobic side chain. The hydrophobic amino acids are glycine, alanine, valine, leucine, isoleucine, methionine, proline, phenylalanine, and tyrosine. Noteworthy amino acids in this category are the proline, methionine and tyrosine. Proline is characterized by pyrrolidine side chain and it is the only amino acid with a secondary amine bonded directly to amino group, making the α-carbon a side chain atom. Due of the stiffness of this amino acid, proline is commonly found as the first residue of an α-helix, in the edge strands of β-sheets, and in turns structure of the proteins. Methionine, besidesthe polar uncharged cysteine, is one of two sulfur-containing amino acids. Tyrosine, despite the presence of an hydroxyl group ($-$OH) can be considered partially hydrophobic because of the aromatic group that is significantly less soluble in water.

Polar uncharged side chain. These amino acid are the serine, threonine, tryptophan, cysteine, asparagine and glutamine. Serine and threonine have a hydroxyl group ($-$OH), asparagine and glutamine an amidic group ($-CONH_2$). Cysteine is characterized by a thiol group ($-$SH). This amino acid has a pivotal role in protein structure since thiol group can be oxidated to give the disulfide derivative cystine.

Electrically positive charged side chain Amino acids in this category are lysine, arginine and histidine. These side chain ends with an aminic group, a guanidino group and a imidazole group, respectively. These end groups make the side chain hydrophilic. The imidazole have a pKa of 6 and in physiological environment can be protoned or deprotonated in function of the chemical makeup.

Electrically negative charged side chain Aspartate and glutamate shares a carboxyl ($-$COOH) common feature.

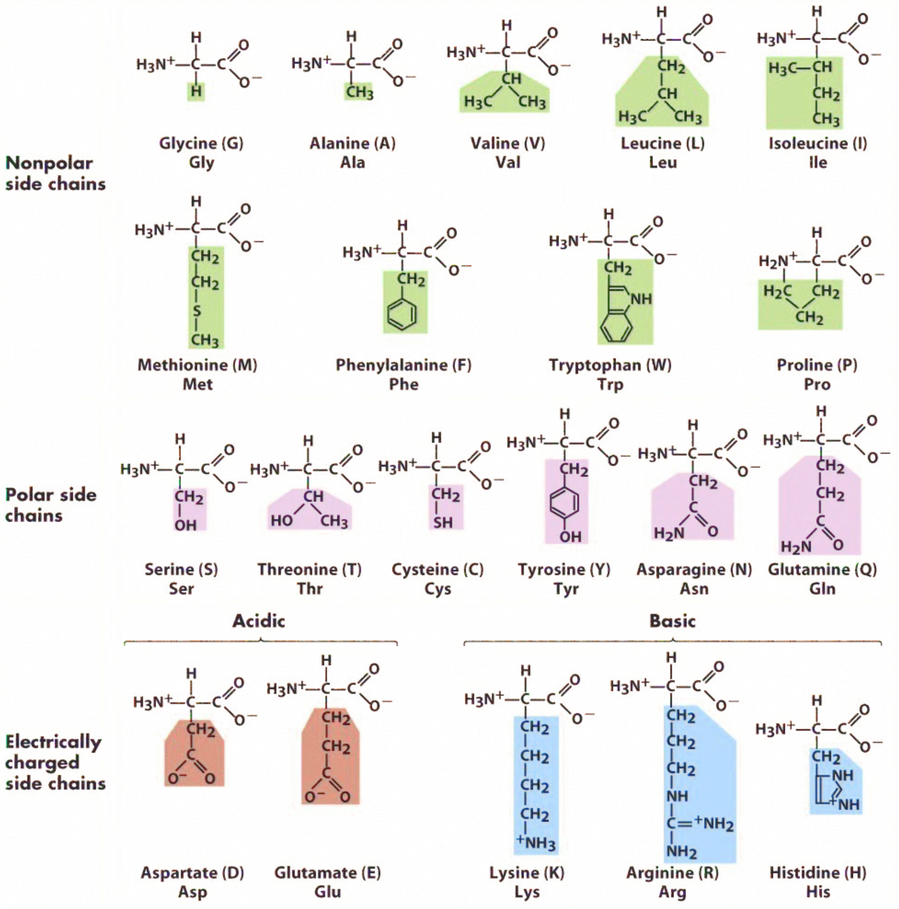


Figure 3-5 Biological Science, 2/e © 2005 Pearson Prentice Hall, Inc.

Figure 1.3: The 20 proteinogenic amino acid. Image taken from [4].

1.2 Forces

Protein folding, peptide interaction, enzymatic activity are driven by a different non-trivial, considering the complexity of the protein chemical characteristics, interactions. The forces interplay in protein stability are prevalently non-covalents forces, with the exception of the sulfur bridges formed by cystine, the oxidized dimer form of the amino acid cysteine, that is covalent. These non-covalent interaction are the hydrogen bonds, the electrostatic interactions, and Van der Waals interactions.

Hydrogen bonds. The Hydrogen bond is an high directional electrostatic bond where an electronegative atom, like oxygen or nitrogen, is able to exert charge delocalization on the bonded hydrogen making this positive. This positive partial charge is able to bind an negative charged atom in the environment. This kind of bond plays a focal role in protein interaction and stabilization. The electronegative atom not covalently attached to the hydrogen is named proton acceptor, instead the one covalently bound to the hydrogen is named the proton donor. This kind of bond is often described as an electrostatic dipole-dipole interaction. H-bonds share some features with covalent bond since it is directional and produces interatomic distances shorter than the sum of the van der Waals radii. Hydrogen bond is highly directional and, consequently, the strongest interaction is due when the the acceptor is aligned to the covalent bond between the donor and the hydrogen atom. Because of this geometric feature, a deviation from linearity leads to a decrease in energy of binding. Typical length of this kind of bonds is 2.5 Å to 3 Å. As explained later on, the *peptide hydrogen bonds* are involved in the structure stabilization.

Electrostatic interaction. This force is generated between the charged or partially charged group, typical of the electrically charged amino acids, can be described as a coulombian interaction between two point charges in a dielectric environment. Electronegativity cause delocalization of electron density towards the more electronegative atoms, generating a dipole moment able to interact electrostatically.

van der Waals. Pairs of atoms interact with a van der Waals energy potential in function of the distance r. The most common interaction potential employed is the in *Lennard-Jones* form

$$V_{\mathrm{LJ}} = \frac{C^{(12)}}{r^{12}} - \frac{C^{(6)}}{r^{6}} \tag{1.1}$$

The first term, the repulsive one, is generated by the overlapping on the electronic orbitals, that is, commonly, modeled as hard spheres with van der Waals radius. This repulsive interaction is one of the main force that generate the constraint interaction. Typical radius of carbon atom is 1.55 Å, while, for nitrogen is 1.75 Å. When two atoms are covalently bonded the distance between the center of the species is less the sum of radii. So Van der Waals radius of a defined element is relative to the bond co-partecipant.

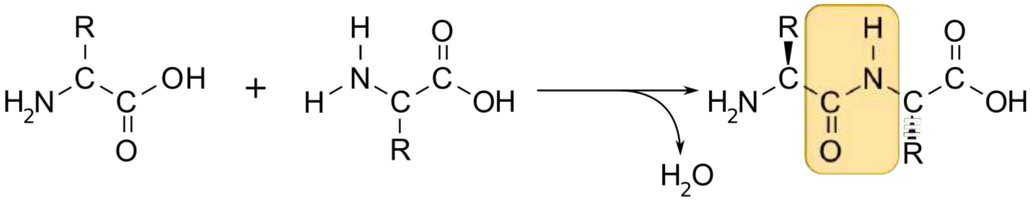


Figure 1.4: Peptide bond condensation reaction. In highlighted are the peptide bond.

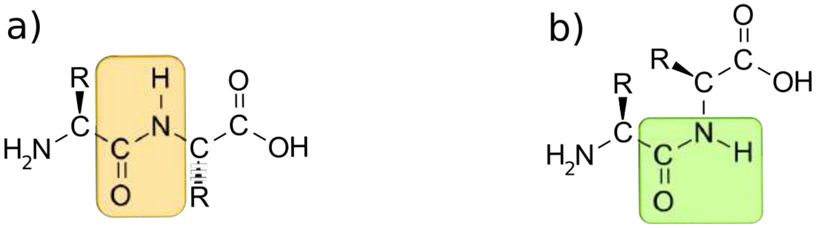


Figure 1.5: Different amino acid dimers geometrical organization. (a) *cis* configurational isomerism; (b) *trans* configurational isomerism.

1.3 Primary structure of protein

The primary structure is the linear sequence of amino acids in a protein. The covalent bond that links consecutive amino acid is a peptide bond formed through a condensation reaction. Peptide bond is formed when the carboxyl group of one amino acid reacts with the amino group of another amino acid , causing the release of a molecule of water (Fig. 1.4). This reaction is energetically unfavorable and, in living systems, is fueled by ATP. In aqueous environment the hydrolysis of the peptide bond is extremely slow, with half life, at 298,15K, of between 350 and 600 years per bond. Typical energies of covalent bonds are 300 kJ mol^{-1} to 700 kJ mol^{-1} at room temperature. This peptide bond construction define the backbone structure in the protein.

The typical length of the bond in peptide bond is 1.32 Å as opposed to the typical C−N and C=N that are, respectively, 1.49 Å and 1.27 Å. This is because of the lone pair of electrons on the nitrogen atom that gives a partial double bond character. The amide group, even though is a single bond, holds a planar behavior occurring in either the *cis* or *trans* isomers (Fig. 1.5). In trans conformation, distance between α-carbon atoms is fixed to 3.80 Å. Peculiar behavior is noticeable in peptide bond of a X-proline, where X is an arbitrary amino acids, since the peptide bond character is lost and the cis-trans form cannot be accomplished. Despite this, trans form is

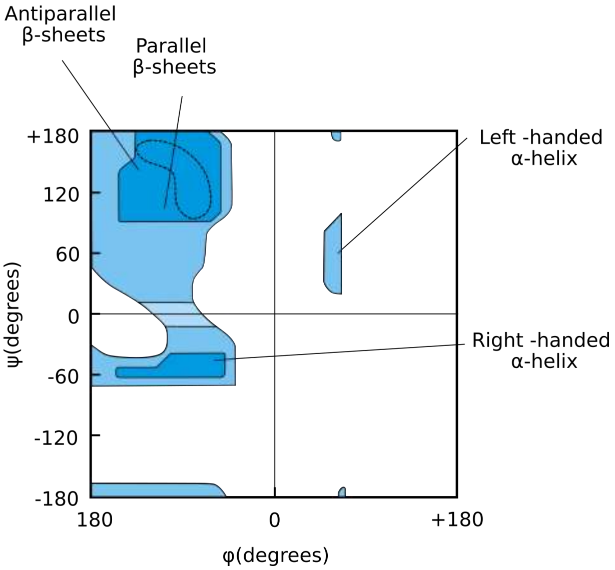


Figure 1.6: Ramachandran plot: in blue the regions where is possible to find the α-helix and the β-sheet secondary structure.

preferred in most peptide bonds. This dihedral angle is labeled as ω. The rotation of the dihedrals formed in the backbone can be defined using the angle ϕ for the rotation around the N$-$C$_\alpha$ axes and angle ψ or the rotation around the C$_\alpha-$C. The peptide bond rotation angle ω is generally 180°. During protein folding not all the variation of ϕ and ψ are allowed because of the presence of steric hindrance due by the side chains. An exception is due by the proline where, because of the side chain structure, ω is fixed. Beyond this, even electrostatic interactions play a role in this angle restriction. It is possible to visualize energetically allowed regions for backbone dihedral angles ψ against ϕ is called Ramachandran plot (Fig. 1.6), introduced by G. N. Ramachandran. Using ω fixed to build up a representation, the backbone can be pictured as a rigid planar structure. This kind of plot can empirically show the distribution of data points observed, to discriminate favored regions. A X-glycine peptide bond in a protein, ψ spread from $\psi = 180°$ to $\psi = 0°$, and gives the condition to cluster formation in Ramachandran plot [5].

The local clusters generated by the ϕ and ψ allowed angle values observable in Ramachandran plot corresponds to the secondary structure of the protein: α-helix structures and the β-sheets.

1.4 Secondary structure of protein

Local conformation as a consequence of protein folding are referred as *secondary structure*. This kind of organizations are particularly stables and are a peculiarity of protein structure. Secondary structure are stabilized by the H-bond formation between the H-bond donor and acceptor species in the backbone, since hydrogen atom in the secondary amine group $-\mathrm{R_2NH}$ is partially positive charged by the electronegativity of the nitrogen atom, and chemical groups like the carbonyl $-\mathrm{C{=}O}$ can form two non-linear H-bonds with the covalent bond. The most important secondary structures are the α-helix and the β-sheet. Other structure, like β-turn or ω-loop, are stabilized by this kind of interaction but play a minor role in protein. Pioneering works in protein analysis has been made by Pauling e Corey that predicted the secondary structure from the optimization of H-bond in proteins [6] , and by Ramachandran that studied the sterically accepted torsion angle in proteins second structures formation.

1.4.1 α-helix

The α-helix (Fig. 1.7(a)) is the simplest arrangement in which a protein can occur during protein folding. This is due for the optimal use of H-bonds between the nitrogen atom and the carbonyl oxygen, located on the fourth amino amino acid away ($i + 4 \rightarrow i$ H-bonds). α-helix backbone proceeding recall a coil springs or the handrails of spiral staircases, side chains in this representation go outwards respect of the central pole. Each turn of the helix extend the total length of the helix of 5.4 Å using 3.6 amino acids. The ϕ and ψ variates of $-45°$ to $-50°$ and $-60°$, respectively, forming prevalently right-handed structures. Mixed stereoisomers are not able to form α-helix structures. The former simplified model of α-helix has been build without take into account the effect of the side chain steric hindrance or charged species that not promote the structure formation. For instance, if a protein has a secondary structure formed by a sequence of amino acid belonging to the same electrically charged side chain family, α-helix will be not built up because of the electrostatic interaction between these side chains. Branched amino acids like valine, isoleucine and threonine destabilize this structure the for steric interaction. Amino acids that are able to form H-bonds can interact with the backbone and, subsequently, unpromote the α-helix formation. As precedently explained, proline, if involved in a peptide bond, due is peculiar side chain structure, is not able to form H-bonds, since the secondary amine, canonically present in the backbone, is a tertiary amine $-\mathrm{R_3N}$.

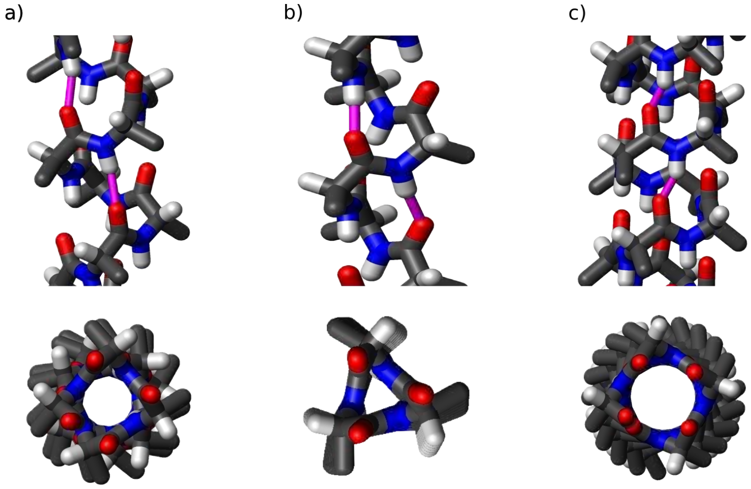


Figure 1.7: Side view (top) and top view (bottom) of several helix type in stick representation. In purple are highlighted the H-bonds: (a) α-helix; (b) 3_{10}-helix; (c) π-helix. Images taken from [8–13].

The distribution of the amino acids participating in the α-helix formation plays a role in the global stability of this structure. Generally, a small electric dipole exists in each peptide bond because of the partial negative charge of the carbonyl oxygen and partial positive charge of the amide nitrogen. A net dipole is formed by the H-bonds and extended along the helix. Dipole moment increases with helix length, therefore, α-helix is negative in the carboxyl-terminal ends and positive in the amino-terminal end. Since the H-bond formation needs a fourth amino acid away in the amino acid structure the last four amino acid in the helix do not participate in the H-bonds network. Negatively charged amino acids are usually found in the amino terminal end, since, this amino acid stabilize the interaction with the positive charge of the helix dipole. Conversely, a positively charged amino acid at the amino-terminal end is destabilizing. Other α-helix like structure are the 3_{10}-helix ($i + 3 \rightarrow i$ H-bonds) and the π-helix ($i+5 \rightarrow i$ H-bonds) (Fig. 1.7(b)-(c)). The 3_{10}-helix constitute nearly 10–15% of all helices types and are mostly located as portion of α-helix structures at amino and carboxyl terminals. The π-helix ($i + 5 \rightarrow i$ H-bonds) constitute the 15% of the secondary structure. The π-helix is typically short, 7-10 residues, and the presence of this structure is associated by an underrated insertion mutation of an amino acid in an α-helix structure.

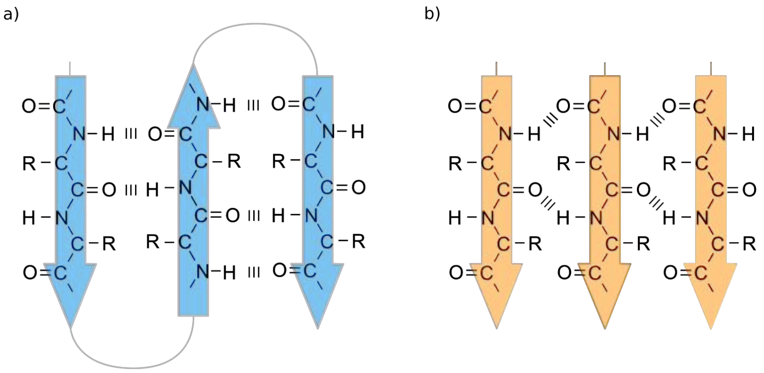


Figure 1.8: Antiparallel and parallel β-sheets. Hydrogen bonding patterns is represented by dotted lines. (a) Antiparallel β-sheet; (b) Parallel β-sheet.

1.4.2 β-sheet

β-sheet are structures composed by portion protein, in some case nearby in the polypeptide chain or even quite distant from each other in the linear sequence of the polypeptide, called β-strands, mutually combine via the H-bonds interaction. This arrangement form an extensive H-bonds network. The side chains in a peptide strand protrudes, from a residue to the following, in opposite directions, creating the alternating pattern. This pattern is followed therefore by the side chain and even the peptide bond is mirrored respect to the direction of the backbone every amino acid. In adjacent β-strands forming the β-sheets, the α-carbon atoms are lined up. Due the tetrahedral geometry of the amino acid this conformation looks pleated like an open folding fan. Because of the directionality of proteins, the adjacent polypeptide chains in a β-sheet can have the same or opposite amino-to-carboxyl orientation, generating parallel or antiparallel β-sheet (Fig. 1.8). These structures are quite similar but the parallel or antiparallel orientations change the repeat period, from 6.5Å for parallel configuration to 7Å for antiparallel, and the H-bond pattern. Steric clash can occur when two or more β-sheets are close together within a protein, therefore the side chains of the amino acid residues needs to be small enough.

In compact protein structure like the globular one, β-turns are quite common motif. In this case one-third of amino acid residue change abruptly direction forming a loop or a turn structure. β-turn connect frequently adjacent segments of antiparallel β-sheet and are generally composed by four amino acid within H-bond is formed between the carbonyl of the first amino acid and the second amine of the fourth.

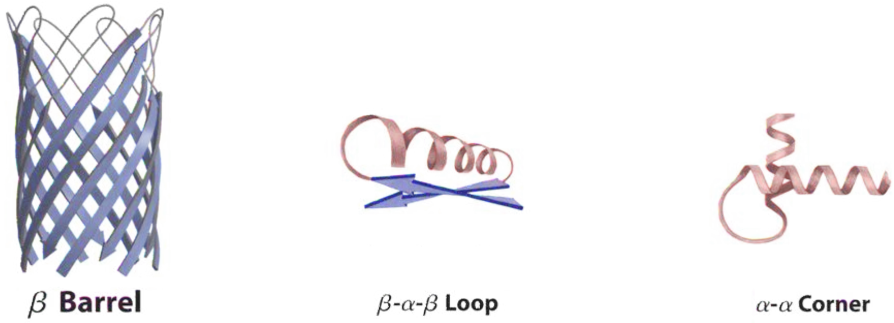


Figure 1.9: Different secondary structures: (a) β-barrel; (b) β-α-β; (c) α-α corner.

Glycine and proline usually take parts in β-turns for different function: glycine is small and flexible, proline, instead, for the peculiarity of the amino acid structure, forces backbone to assume a cis configuration which permit a tight turn. Considerably less common is the γ-turn, a three residue turn with a hydrogen bond between the first and third residues.

1.5 Tertiary and quaternary structure

The arrangement of several secondary structures are called *supersecondary structures*, also called *folds* or *motifs* (Fig. 1.9). This complex structures cluster shares similar function in different proteins. Protein with more than 300-400 amino acid folds in specific globular sub-regions called *domains*. This subdivision arise by an evolutionary advantage given by the quicker folding and the possibilities of each domain to fold individually. Domains can be characterized by the fold. Domains have the same fold if can share similar supersecondary structures.

The *tertiary structure* represent the three-dimensional shape of a proteins, whereas the arrangement of a protein composed from different polypeptide chains is defined as *quaternary structure*. In this hierarchic level, same multiple sub-unit can interact and, usually, are arranged with rotational or helical symmetry. This kind of subunits super structures are preferred, in evolutionist terms, because built up a fully functional multi-polypeptide protein can lead to translations error or misfoldings. Further than this, use single chain protein to build up complex structure is more efficient from a genetic heritage usage since the information for each sub-unit need to be stored just one time.

Protein can be classified by the *The Structural Classification of Proteins* (SCOP) database. The hierarchical organization is defined as, follows

Class Type of folds.

Fold The different three dimension structure of domains in a class.

Superfamily Domain in a fold grouped in superfamilies, which have a distant common ancestor.

Family Domain in a superfamily grouped in families, which have a recent common ancestor.

Protein Domain Domain in a fold grouped in protein domains, same protein.

Species Domain in protein domain grouped in species

Domains Part of a protein, for small proteins can be the whole one.

The classes groups the domains eleven categories depending on the second structure contents, type of protein, or quality. Four on eleven are dedicated for the second structure content: all-α proteins are composed prevalently by α-helix, all-β, controversially, are defined by the prevalence of β-strands, α/β include $\beta - \alpha - \beta$ where β-sheets are surrounded by α-helices. Last category is $\alpha + \beta$ where no evident motif arise.

1.6 Protein folding

The three-dimensional unique structure in which a protein can fold is defined by the interaction of a specific polypeptide sequence and environment contribute. In the last decades many works tried to clarify this process, using both experimental and theoretical approaches, in order to reveal the intimate interplays from which arise a defined native state. Protein folding structure prediction as the Chou-Fasman [14] or the Garnier-Osguthorpe-Robson [15] methods, focus on this kind of possibility. The knowledge of the physical connection between a defined primary sequence and the final geometrical configuration have not only basic knowledge purpose but even strictly practical. Genetic engineering has been involved the modification of naturally occurring proteins and the ability to manage the protein folding precess should be a possible way to design new functional proteins that can be used in pharmacological, industrial and general nanotechnology field. Bioinformatics use data bases of protein

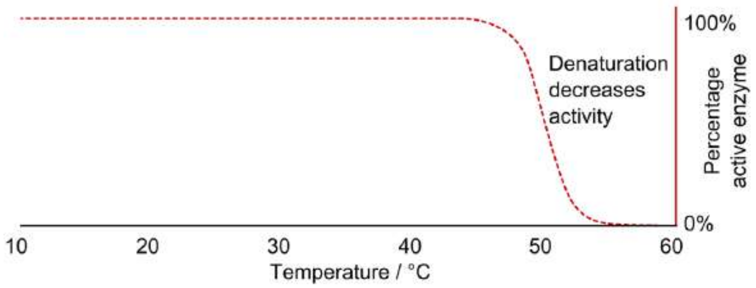


Figure 1.10: General proceedings of an enzymatic activity as function of temperature.

structures and relative amino acids or nucleotide sequences to find , using a statistical approach able to identify similarities, a possible structure, function or, from a genetics point of view, a phylogenetic path to a common ancestor.

The understanding of the denaturarion, the loss of three-dimensional structure and function, and renaturation has been a starting point for the thermo analytical protein studies. Protein structure evolved to have a specific function and activity in a well defined cellular environment, due the definition of denatured state, emerge that a precise polypeptide could cover different paths of folding in function of the interacting solvent. Denatured state is non unique and is better described as a family of structure that occur in a random linear configuration and in a, subsequently, loss of function. Since the folded state is defined by a complex interplay due by weak interaction, provide heat to the system can be one of the way to denature a protein. If the change in temperature is controlled and is increased slowly, protein function and structure remains stable until an a sudden loss of activity. This sigmoid-shaped change (Fig. 1.10) implies that a small lost in structure stability destabilize the other parts.

The tertiary structure is determinate, as the secondary, by the amino acid sequences. This statement can be proved because denaturation, as shown by some experiments, can be reversible. This process is called *renaturation*. First evidence that the amino acid sequence holds all the information necessary for the folding was carried out by Anfinsen in the 1950s [16]. This shows that protein folding is a thermodynamically driven processes where the ground state of this system is the native structure and the solvent molecules. The reversibility of unfolding and refolding, using thermodynamic stability, is a two-state process:

$$P^{(\mathrm{Fold})} \xrightleftharpoons{k_u} P^{(\mathrm{Unfold})} \tag{1.2}$$

The reversibility observed by the Anfinsen experiment is enshrined in this type of two

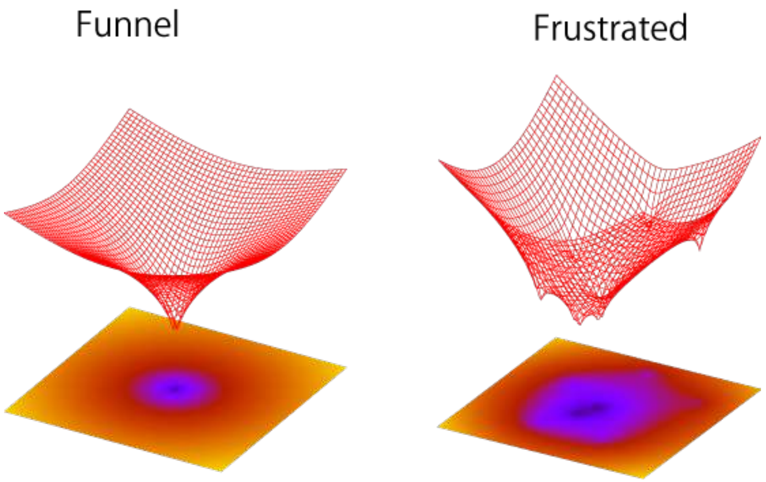


Figure 1.11: Different types of energy landscape. On the left an ideal funnel shaped energy landscape. On the right a frustrated system, characterized by many local minima.

step process. Purified ribonuclase, a type of nuclease that catalyzes the degradation of RNA into smaller components, can be denatured in urea solution in the presence of 2-mercaptoethanol as reducing agent. This molecule is able to reduce the disulfide bond, meanwhile urea minimize the hydrophobic interaction pushing structure to unfold. After the removal of urea and 2-mercaptoethanol, protein refolds in the active native structure. If the process was random the eight cysteine could recombine to build up the four disulfide bonds in 105 different ways.

Proteins generated by the metabolic machinery are assembled in a very high rate. A complete, biologically functional protein of 100 amino acids can be synthesized in roughly 5 second at 37 °C. This 100 residues protein have 99 bonds and 198 possible ϕ and ψ angle value. Hypnotizing, that only three configuration are accepted by a random process and each possible conformation is tested until it finds the native state this require 3^{198} steps. Assuming the shortest possible time ($\approx 10 \times 10^{-13}$ sec) for bond rotation, the time required for the process would amount 10^{72} years, more that the age hypotized for the universe (13.8×10^{9} years). This statement, called Levinthal's paradox, was proposed by Cyrus Levinthal to point out that protein folding cannot be a random process [17]. Many models has been proposed to elucidate this process [24–26]. In all of them he common feature lies in the *Principle of minimal frustration* proposed by Joseph Bryngelson and Peter Wolynes in the '90s [21]. The energy landscape, a representation of energy function across the configuration space of the system, of a protein is generally coarse, so a change in the conformation define a jumpy and many minima profile in energy, generating multiple u-shaped energetic wells directed to the minimum (Fig. 1.11). This roughness of the energy landscape generate in a many interaction in energy function. This interaction is defined *frustation.* A canonical example of frustrated system is the *spin glass* [18], a disordered magnetic system in which spins of the component atoms are randomly directed. The analogy with glass arise from the disorder position pf the magnetic component as bonds in an amorphous solid. Spins in this kind of system interact as consequence of the random position, pointing in the same direction (ferromagnetic) or in opposite direction (anti-ferromagnetic). These local behavior is not able to satisfy completely a spin general arrangement of the system that is, consequently, frustrated. Complex systems, like protein folding need to be described adding a stochastic component to the Hamiltonian and gaining the system fuzziness. Frustration, in polymers, arise from the inability to satisfy electrostatic interaction due the unfavorable path needed to this hypothetical optimization since the system cannot satisfy all the energetic and geometrical constraint simultaneously.

The principle of minimal frustration has been stated to solve this discrepancies. The idea of this principle defines that evolution selected that amino acid sequence ables to fold in a stable and functional form. In parallel, evolution, operates a counter-selection against the amino acids that disfavored an appropriated folding. Richard Dawkins in "The Blind Watchmaker" wrote: "*Mutation is random; natural selection is the very opposite of random* " [22]. Nature uses mutation as material for change, and even in the principle of minimal frustration changes are feasible as local minima in the energy landscape of protein folding. Jose Onuchic proposed that protein folding energy landscape are funnel like, with local minima but strictly directed in a global minimum or ground state where lies the native protein [23]. Three-dimensional energy landscape consist in the configuration space of the protein (x, y) and the energy (z). At higher energy system, near the upper-edge of the funnel-shaped potential, as reported in Fig. 1.12, the correlated configuration space, the protein is unfolded and reach the maximum number of conformation, indicating that the unfolded state is a family of possible configuration that converge with different path of folding in a unique native configuration. If the denatured state were a unique configuration the energy landscape, using a geographical analogy, should be on the on the tip of a high ground. These pathways can be used more or less frequently depending the thermodynamic favorability. During the intermediate state, or *molten globule*, protein starts to assume more thermodynamically favorable structure until do not attain the native state [28]. The funneled energy landscape and the principle of minimal frustration are consequences defined by a "top-down" approach for the protein folding.

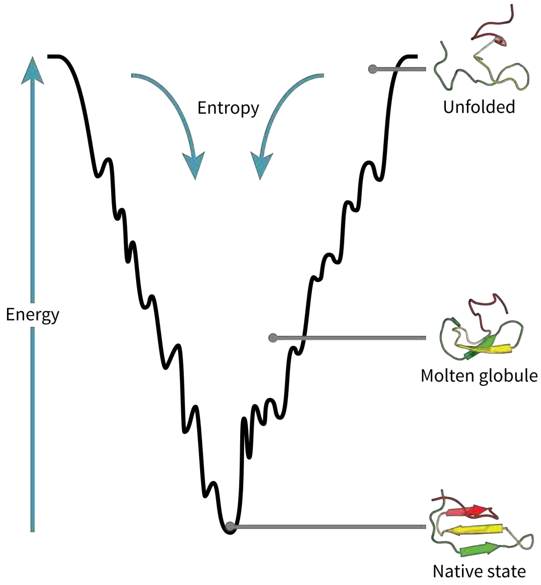


Figure 1.12: The energy folding funnel landscape for protein denaturation. The width of funnel is related with the entropy of the system.

In literature the two most supported models are the hierarchical [34] and the HP model [26]. In the hierarchical model, local second structure are the first structure formed by the primary sequences since some amino acid fold more specifically in α-helix or β-sheets. Supersecondary structure are built up subsequently by long-range interaction of the former. This process continues until complete folding. Another model, HP model, states that protein collapse due the cooperation of hydrophobic interaction with aqueous environment, taking in to account only the difference between hydrophobic (H) and polar (P) amino acid residues. The interaction energy between two hydrophobic amino acids V_{HH} is negative, instead interaction between two polar amino acids V_{HP} or mixed types V_{PP} a larger energy. The HP model has been used as interacting parameter in lattice protein simulation, where interaction of hydrophobic residues are the driving force of protein folding using an explicit kind of solvent.

1.7 Protein solubility and stability

The *solubility* of proteins is correlated with the distribution of hydrophilic and hydrophobic amino acids in the protein surface. Globular protein carry out function

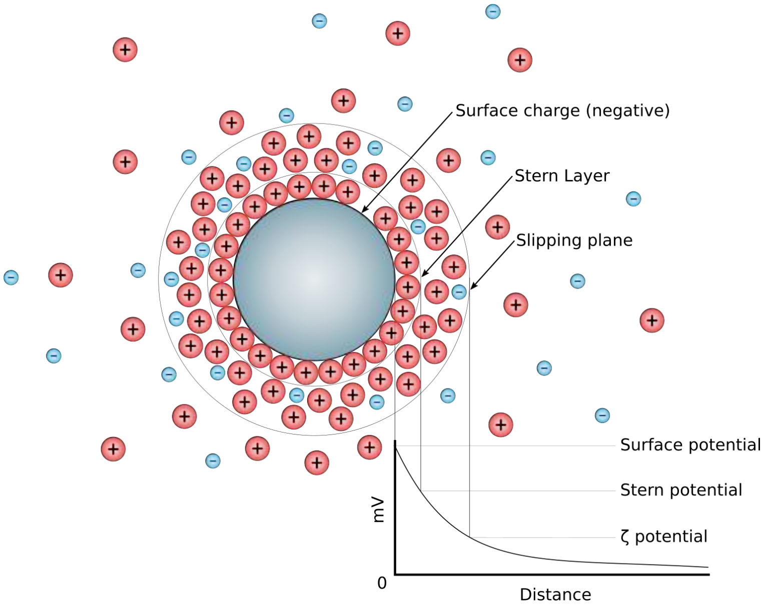


Figure 1.13: Interface double layer diagram showing the electrolyte concentration an potential as function of distance from the charged surface.

in aqueous environment and, consequently, hydrophilic residues occur prevalently in the protein surface. Protein with high presence of hydrophobic amino acids have low solubility in water and prefer to bind or be completely surrounded by polar environment. Trans-membrane proteins are an example of this case, since are partially or totally immersed in the phospholipidic bilayer. Charged surface interaction with solvent is fundamental for protein activity since solvation shell around the protein (≈10 Å) have a completely different behavior respect the bulk. When proteins are solubilized in physiological environment the electrolyte couterions associate to the proteins surface forming a shell, named as *Stern layer* (Fig. 1.13). Next to this shell water molecules form a solvation layer able to diffuse from the shell to the bulk solvent with a gradient that contains a decreasing contraction of couterions and an increasing concentration of ions. The thickness of this layer is called *Debye length* and the outer layer is the *diffuse layer*. The Debye layer is characterized by the presence of a *slipping plane* that separates mobile fluid from fluid that remains attached to the surface. On this plane the electric potential, the ζ*-potential*, can be see as the potential between solvent and dispersed protein. The precedence of this layers decrease the aggregation because of the decrease in ionic interactions. The ζ-potential is used as indicator of the colloidal stability of a solution. This description

H-bonds	Energy (kJ mol^{-1})
F−H···F	161.5
O−H···N	29.0
O−H···O	21.0
N−H···N	13.0
N−H···O	8.0

Table 1.1: Typical H-bonds bond energy in vapor phase.

needs to take into account that proteins have not a uniform charge distribution since this is dependent on the amino acids on the surface, consequently, interfacial double layer potential is expressed in a more complex function. Proteins are able to associate through dispersive and attractive force generated by permanent and induced dipoles. Electrically charged amino acids are able to have electrostatic interactions and in this case electrolyte cause the a decrease in protein-protein interactions. The precipitation of a colloidal protein solution can be modulated changing the pH of the solution. The isoelectric point (pI) is the pH at which the primary charge of the proteins is set to zero. Charged surfaces have an repulsive effect on protein due the interaction former described. Evaluating that, setting the solution pH as the pI of protein the net charge in the latter will be zero, decreasing the repulsive forces and the attractive force can dominate giving to the aggregation of protein.

The protein *stability* in a solution can be described in terms of thermodynamic stability or terms of kinetic stability. The thermodynamic stability is due by the difference in Gibbs free energy ΔG between the free energy of the folded state $G^{(\text{Fold})}$ and the free energy of the unfolded state $G^{(\text{Unfold})}$.

$$\Delta G^{(\text{Unfold})} = G^{(\text{Unfold})} - G^{(\text{Fold})} = -RT \ln k_u = -RT \frac{[P^{(\text{Unfold})}]}{[P^{(\text{Fold})}]} \tag{1.3}$$

Folding free energy $G^{(\text{Fold})}$ is usually in the order of $-20\,\text{kJ mol}^{-1}$ to $-65\,\text{kJ mol}^{-1}$, therefore, relatively small respect to the the typical atoms interaction energy as reported in table 1.1 The change in Gibbs free energy ΔG, is the maximum amount work that can be extracted from a thermodynamically closed system at constant temperature and pressure.

$$G = H - TS \tag{1.4}$$

where the H is the enthalpy, T is the temperature, and S is the entropy in the system.

The kinetic stability, controversially, describe how fast a protein goes to the folded state to the unfolded state. A kinetically stable protein unfold more slowly respect to a kinetically not stable one. This case is not an equilibrium process since the protein unfold to a denatured state in irreversibly way. The energy needed for the unfolding process from the kinetic point of view is the amount of energy necessary to pass from the folded state to the transition point. This energy is well knows as the *activation energy*. The irreversible unfolding is described by the relation:

$$P^{(\mathrm{Fold})} \overset{k_u}{\rightleftharpoons} P^{(\mathrm{Unfold})} \xrightharpoondown{k_i} P^{(\mathrm{Inactive})} \tag{1.5}$$

where the energies barrier to the inactivated state $P^{(\mathrm{Inactive})}$ is smaller the the activation energy needed the transition point to the folded state $P^{(\mathrm{Fold})}$.

The main contributions to protein stability, most widely accepted in literature, are the *hydrogen bonds* and the *hydrophobic effect*. These contributions are cooperative since iteration generated by each residue are generally small. On the other hand, the unfolded state is stabilized by the *conformational entropy*. The rotation of a bond in denaturated state is much more favorite then in folded state. This define a strong entropic driving force for unfolding. Conformational entropy can be figured, likewise, as that increasing the stiffness of a polypeptide in the unfolded state, decreasing the possible configuration in the unfolded state, implies in a stability gain for the folded stare respect on the unfolded state [30]. Therefore, the unfolded protein is a limit case where all the residue interact only with the water molecules and do not interact as intra-protein bonds. Using this definition the unfolded state is, in opposite way to the folded state in a given solvent, a collection of extended conformations. This picture is not able to fit always the experiment evidence. Rarely, unusual behavior has been reported in literature [31]. For the present aim, we assume that all interaction that stabilize the native state, i.e the H-bonds, are completely removed in unfolded state.

As previously described the H-bonds have a noteworthy role protein folding. Beyond the peptide hydrogen bond, H-bonds behavior of liquid water, and the global uniqueness respect to other liquid, take a relevant contribute. In a water molecule, using oxygen as reference, there are four regions of excess of charge in tetrahedral arrangement. Two on four are generated by the hydrogen atoms due the high electronegativity of the oxygen that attract the shared pair of electrons towards itself. The negative excess of charge, otherwise, is located in two ion-pairs on the oxygen. This charge disposal and the electronegativity of the oxygen make the H-bonds the main interaction between water molecules. The lowest arrangement is reach when the donor molecule, respect of the hydrogen-oxygen axis, is tilted of 57° on the orthogonal plane to the plane of the hydrogen-oxygen-hydrogen of the acceptor [33]. Taking into account a water molecule this generates four H-bond interaction with the four nearby water molecule disposed in a tetrahedral arrangement. This dynamic

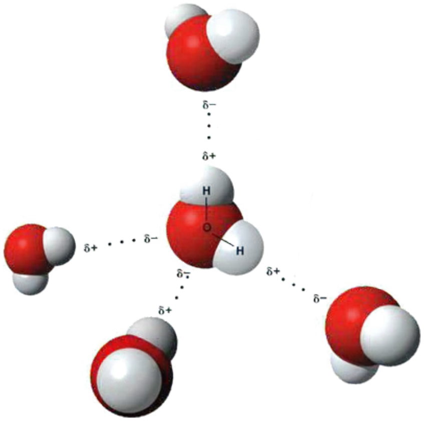


Figure 1.14: The tetrahedral arrangement of the liquid water.

configuration, where hydrogen atoms behave like donor and the ion-pairs as acceptors, is called *Walrafen pentamer*. Since that, is possible to fill the space tessellation with tetrahedral cells making the water capable to build a crystal typical of the solid ice.

The H-bonds generated interaction in liquid water can be outlined by clustering models where different number of water molecules interact in non canonical ideal tetrahedral configuration. As reported in table 1.1 the strength of an hydrogen bonds is $8\,\mathrm{kJ\,mol^{-1}}$ to $29\,\mathrm{kJ\,mol^{-1}}$. Considering protein, McDonald at al. in 1994 showed that for high resolution structure, buried nitrogen and oxygen atoms fail for the 9.5% and 5.8%, respectively, to form a H-bond but, relaxing the protein structure, roughly the 1-2% of the backbone H-bond donors and acceptors fails to bonds with the counterpart. Carbonyl, more specifically, fails with 80% the second H-bond. [32] From an entalpic point of view the H-bonds contribute is unfavorable for protein folding. Hydrogen bonds, instead, stabilize protein with an entropic effect due by the disrupted H-bonds in water bulk by the presence of the protein, leading an entropy gain in solvent.

The main driving force for protein folding as been conferred the to hydrophobic effect proposed by Kauzmann et al (1959) but, recently, literature rebalanced this contribute significance [35]. The hydrophobic effect is given by the aggregation of nonpolar substances in aqueous solution in order to not interact with water molecules.

This behavior characteristic of mixed nonpolar/aqueous solution is lead by H-bonds formation between molecules of water in the way to minimizes the area of contact between water and nonpolar molecules. This interaction is typical in aqueous environment since dipole-dipole interaction in water are stronger than the dipole-induced between water and protein. From a general point of view, in a multi-phase solvent the energy needed to transfer a non-polar molecule from an organic phase to polar one is the energy of transfer $\Delta G_{\mathrm{Tr}}^{(\mathrm{Polar})}$

$$\Delta G_{\mathrm{Tr}}^{(\mathrm{Polar})} = \Delta H_{\mathrm{Tr}}^{(\mathrm{Polar})} - T\Delta S_{\mathrm{Tr}}^{(\mathrm{Polar})} \tag{1.6}$$

In normal condition of temperature (300 K), the enthalpy of transfer $\Delta H_{\mathrm{Tr}}^{(\mathrm{Polar})}$ is negligible. Entropy, in contrast, is negative since water molecules interact, with H-bonds, with other water molecules. Locally, water nearby the non-polar compound form H-bonds with subsequent lose of energy. Consequently, water balance lose energy due the generation of water-water H-bonds near the solute molecule where water-water H-bonds pattern is similar to that of solid water. Rising the temperature up to the boiling point of water, tetrahedral configurations are disrupted and, by the relation of entropy with the temperature, the entropy of transfer $\Delta S_{\mathrm{Tr}}^{(\mathrm{Polar})}$ tends to zero. At the same time the enthalpy of transfer $\Delta H_{\mathrm{Tr}}^{(\mathrm{Polar})}$is positive. Even enthalpy is correlated with the temperature by the *Kirchoff's Law*. Enthalpy increase with temperature, leading to a gain in enthalpies of product and reactants. At constant temperature, the heat capacity is

$$C_p = \left(\frac{\Delta H}{\Delta T}\right) \tag{1.7}$$

Kirchoff's Law can be used only for small temperature change (±100 K) since for greater range the correlation the heat capacity began not constant. In this lower range of temperature, the enthalpy change is proportional to the product of the change in temperature and the variation in heat capacity for product and reactants. The final enthalpy is defined as

$$H_{T_f} = H_{T_i} = \int_{T_i}^{T_f} dT C_p \tag{1.8}$$

or for systems where the heat capacity is temperature independent for a defined range of temperature:

$$H_{T_f} = H_{T_i} = c_p(T_f - T_i) \tag{1.9}$$

from the 1.7 is evident that the entropy and the enthalpy are temperature correlated in different ratio. Because of this temperature correlation can be defined a specific temperature in which the hydrophobic effect reach an optimum point and below and above this temperature the hydrophobic effect decreases. Another result of this water lattice reorganization caused by the hydrophobic effect is the gain in thermal capacity a result of protein unfolding, calorimetry routines (see A.2) use this variation to calculate the free energy. In spite of this, apolar side chain can be found inside the globular structure of the proteins where mutually interact to contribute for protein folding. In protein folding this separation leads to a compact structure where apolar side chains are buried inside the globular structure to avoid the interaction with water. Most recent theory compare the hydrophobic effect with the hydrogen bonds and the van der Waals interaction. The burial of a $-\mathrm{CH_2}-$ contributes $\approx 5.0\,\mathrm{kJ\,mol^{-1}}$ to the stability instead the $\Delta G_{\mathrm{Tr}}^{(\mathrm{Polar})}$ is $\approx 4.0\,\mathrm{kJ\,mol^{-1}}$, suggesting that the 80% of the hydrophobic effect is generated by the hydrophobicity and the 20% from the packing of amino acid side chain inside the globular structure [36].

Other minor factor are the charge-charge interactions, salt bridges, the aromatic stacking, metal binding and disulfide bonds. The distribution of charge in a protein surface, at physiological pH, arise to be more attractive than repulsive, so this interaction stabilize the folded state. Charge-charge interactions, on the other hand tend to stabilize even the denaturated state making this contribution small. Charge-charge interactions have a dependence with the temperature and thermophilic organisms seems to adopt this strategy to stabilize proteins [37]. The salt bridges are a particular type of H-bonds strengthened by the combined effect of an electrostatic interaction, typical the electrically charged amino acid positioned at less then 5 Å as glutamate and lysine. This interactions contribute less than $\approx 5\,\mathrm{kJ\,mol^{-1}}$ on the protein surface. On the other hand, buried salt bridges have a contribute more than $\approx 17\,\mathrm{kJ\,mol^{-1}}$.

Aromatic side chains as phenylalanine, tyrosine and tryptophan participate to the protein stability with the *aromatic stacking*. This non-covalent interaction is presented between the aromatic rings. The aromatic rings have more p-orbitals due by the covalent bonds, the superposition of these orbitals generate a so called π-orbital conjugated. The interaction of more π-orbital in stacked configuration strengthen this bond more stable since there is a gain in shared electron.

1.8 Protein stability in solvent other than water

As precedently reported protein stability is in range of the order of $20\,\mathrm{kJ\,mol^{-1}}$ to $65\,\mathrm{kJ\,mol^{-1}}$. Biological evolution use random mutations as bifurcation point for a possible more efficient, in some task point of view, product. This idea of evolution can be observed in different complexity level as species, organism or molecules. Proteins perform a widely array of function, from the structural protein, to the catalysis, passing through stocking, transport, receptor and so forth. All this classes have in common the capability of interaction with other molecules by a particular moiety of the protein as a catalytic or a binding site. This interaction is strictly correlated with the structure of the protein and the working environment: water. The evolutionary pressure for a protein needs to be well balanced between the functional state and the possibilities to not accumulate a mutation with a prominent impact for the biological function. In general, a mutation that leads to a production of a damage for the native protein stability is less probable than a decrease in stability. The low range energy in protein stability is used to control this type of event and prevent the accumulation of lethal mutation. Mutations can effect in different ways considering the amino acid substituted, in fact , mutations in amino acid that have similar chemico-physical properties, compared with the wild-type amino acid, have less influence on a large change in stability. Stability can be used as different point of view of the functionality, since these two features are superimposed by the structure in native state. Mutation can lead to misfolding, aggregation with other proteins, or non interaction with the proposal molecule.

During the folding process, protein lose entropy ($\Delta S_{\text{prot}} < 0$) meanwhile solvent gains entropy ($\Delta S_{\text{solv}} > 0$). Throughout this process protein break protein-water (P-W) H-bonds to form protein intramolecular (P-P) H-bonds. In parallel, water break water-protein(W-P) H-bonds formed to restore the water tetrahedral H-bonds network. In this case, the global balance of the components is not clear and cannot be defined in advance. The H-bonds binding energy is defined by the reaction environment. Another contribute is due by the electrostatic interaction, generated, in physiological environment by electrically, positive or negative, charged side chains. This interaction are well defined in water environment. When a protein in transferred in another solvent, as ethanol, cycloehexane or vacuum, all the interaction precedently described needs to be re described, Pace in a 2004 work, described this kind of interactions [38].

In this new solvent the native state needs to share the same features of the native state in water, so, must be favored as in water and the conformational stability needs to be in in range of the order of $20\,\mathrm{kJ\,mol^{-1}}$ to $65\,\mathrm{kJ\,mol^{-1}}$. In the other hand, each state, the unfolded and the folded, needs to be soluble in the new solvent. Ethanol (C_2H_6O) is completely miscible in water, polar organic solvent, aliphatic hydrocarbons, and aliphatic chlorides. This molecule have a dipole moment (1.69 D) making this molecule polar. The presence of the hydroxyl group ($-OH$) able to make H-bonds, permit to the ethanol to be less volatile than similar molecular weight compound as the propane (C_3H_8). Controversially the cyclohexane (C_6H_{12}) is non

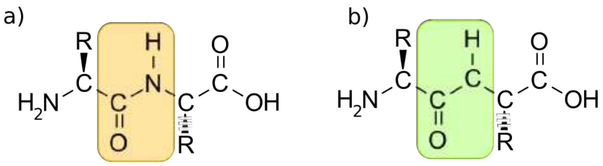


Figure 1.15: (a) Peptide bond; (b) Pseudo peptide with ester bond instead of the canonical peptide bond.

miscible in water, but soluble in ether, ethanol and acetone.

In ethanol, protein stability is decreased by the exposure of a peptide bond group and the gained for the exposure of a non-polar side chain. Since this groups are buried in equal amount this contribution the ΔG_{Tr} suggest that the protein unfold in ethanol. This behavior is usually observed in water-ethanol solvent solution where the concentration depends on the specific protein. α-helix is quite stable in ethanol since the backbone is unexposed and the the non polar side chain free to interact with the solvent. In this solvent protein structure defold completely and folded in to a new all α-structure. Refolding could be modulated with a less polar backbone replacing the peptide bond with an ester bonds (Fig. 1.15).

Cycloehexane free energy of transfer ΔG_{Tr} from water to cycloehexane is greater then ethanol. Cycloehexane is not polar and cannot form H-bonds with the backbone, controversially, cycloehexane molecules have an advantageous interaction with the backbone due van der Waals interactions as the non polar side chains. In this environment protein is stable and functional. In non-polar environment solubility of protein and substrate is, in the most of the case, very low. Protein solubility can be gained changing all the amino acid side chains in to non-polar.

Protein, counterintuitively, are not able to unfold in vacuum. In vacuum, free energy of transfer from water for the backbone is more unfavorable than cycloehexane and the non-polar side chain are less favorable. Backbone prefer the aqueous environment since in vacuum cannot formed H-bonds or van der Waals interaction. Controversially, non-polar side chains are more stable in vacuum because the the unfavorable hydrophobic effect given with this groups in water.

Chapter 2

Computational methods

In this thesis has been used numerical methods like *molecular dynamics* (MD) and *Metropolis Monte Carlo*. Molecular dynamics is a computational method able to simulate physical interaction in the way to explore the conformation space generated moving atoms of atoms and molecules. Atoms and molecules can interact for a defined amount of time with the purpose of obtaining view of the dynamic evolution of the system. First condensed phase simulation has been done by Alder and Wainwright in 1957 with an IBM 704 and using a hard-sphere model [39]. In this simplified model spheres move in linear trajectories between the collision. Each sphere is defined by his center of mass and a collision occur when to center of mass distance equals with the sphere diameter. The pair potential was defined as square-well potential where the interaction between two particles is zero beyond a cutoff value σ_2 and infinity if below the value σ_1 and equal to a predefined potential v_0 between these cutoff value. Highly simplified model like these had a key role giving an idea of the nature of this system. During the evolution this kind of scheme of MD simulations has been preserved reaching the possibilities to be tool appropriate to biochemical and biophysical simulations. Proteins and other macromolecules can be simulated using tools based on experimental data from X-ray crystallography and NMR spectroscopy to observe phenomena that cannot be explored directly. MD simulations are useful to study the motions of macromolecules, for interpreting the results of experiments and for modeling interactions with other molecules like the ligand docking.

Metropolis Monte Carlo algorithm calculate a sequence of random samples from a probability distribution for which direct sampling is difficult to calculate, this method was developed in the late 1940s by Stanislaw Ulam.

2.1 Molecular dynamics

Molecular dynamics use classic mechanics to solve Newton's equation of motion of N interacting atoms:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = F_i = -\nabla_i V(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, ..., \mathbf{r}_N). \tag{2.1}$$

Forces are the negative derivatives of a potential function $V(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, ..., \mathbf{r}_N)$ of N molecule with mass m and position $(\mathbf{r}_i, ..., \mathbf{r}_N)$. The trajectories of each molecule are rappresentated as a function of time. Molecular dynamics provide a quantitative prediction taking into account have some limitations defined by the approximation of the calculation. The main nature of a particles system can be indeed described by classic mechanics but very light atoms, like hydrogen atoms, has quantum mechanical character. The classic harmonic oscillator, pivoting elements in molecular dynamics calculation, has a substantial difference from the real quantum oscillator when frequency ν are multiples of the Planck constant $\hbar$. Typical vibration frequencies for hydrogen atoms are at 300 K are roughly $200\,\mathrm{cm}^{-1}$, so all bond-angle vibration cannot be computed in classical way and to an acceptable value we can perform different strategies. To perform a molecular dynamics simulation using classical harmonic oscillator we need to correct the total energy $U = E_k + E_{pot}$ and specific heat C_v or using *constraints* in the equations of motion. The idea is based on that the quantum oscillator describe in a better way a constrained bond than a classical one. Use of constraints bear the simulation algorithm to use a larger time step without losing accuracy in calculation.

Molecular dynamics use conservative force fields that is strictly described by the atoms position. Electrons use the *Born-Oppenheimer* approximation and stay in their ground state denying the electronic transfer processes and exited states. From this, and other issues, ensue chemical reaction cannot be computed. To define force between interacting bonded or non bonded atoms we need to define a force field able to parametrize constant specific for each interaction. All the Non-bonded forces result from the sum of non-bonded iterations by an effective potentials.

2.1.1 Force field

Force fields methods use the nucleus position to calculate the energy of a system. This methods, respect to a quantum mechanical approach, allow to calculate system with large amount of molecules without using increasing the computer performance. Force field define interaction between particles through three macro-contribution:

bonded, non-bonded and restraints. Bonded contributes are defined as function that describe the amount of energy change respect to a distance r between two bodies (Fig. 2.1), an angle θ in a three bodies system (2.2), and a dihedral angle ω that describe the mutual orientation between two planes generated by four objects (Fig. 2.3). Angle and bond interaction are calculated with an harmonic potential respect of a reference point of equilibrium. The dihedral component is calculated as a linear combination of periodic function where ω is the angle and ϕ is the phase. The non-bonded contributions are defined by a Coulombian potential, Lennard-Jones potential, and a restraint potential that impose some rigidities to the system. Total potential can be written as:

$$V(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) = \sum_i \frac{1}{2} k_{ij}^b (r_{ij} - b_{ij})^2 \qquad \text{Harmonic potential} \tag{2.2}$$

$$+ \sum_i \frac{1}{2} k_{ijk}^\theta (\theta_{ikj} - \theta_{ikj}^0)^2 \qquad \text{Harmonic angle potential} \tag{2.3}$$

$$+ \sum_{i,j,k} \frac{1}{2} k_{ijk}^\omega \left[(1 + \cos(n\omega_i - \phi_i)) \right]^2 \qquad \text{Diedrals based angle potential} \tag{2.4}$$

$$+ \sum_{i<j} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \qquad \text{Lennard-Jones potential} \tag{2.5}$$

$$+ \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \qquad \text{Coulomb potential} \tag{2.6}$$

$$+ V_{\text{res}} \qquad \text{Restrained contribute} \tag{2.7}$$

2.1.2 Bonded interaction

Bonded interaction are settled by a fixed list of atoms interacting on not-only neighbor atoms but even three and four body as well. Mainly we can define:

1. Bond stretching (2-atoms)
2. Bond angle (3-atoms)

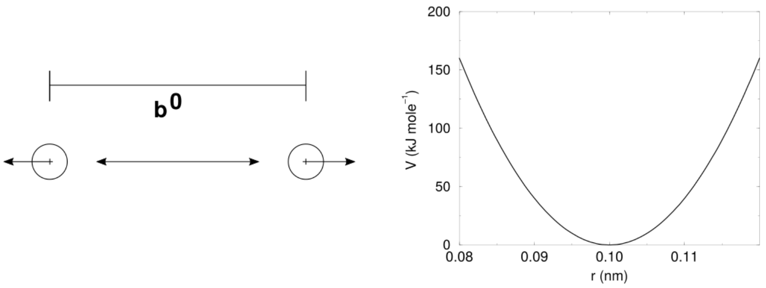


Figure 2.1: Bond stretching: The interaction model on the left, on the right the bond stretching potential.

3. Dihedral angle (4-atoms)

Dihedral angles, in such peculiar way, can be used as *improper dihedrals* to force atoms to lay in a defined plane or to prevent a change of chirality. Bond stretching between a pair of atoms is described by an harmonic oscillator (Fig. 2.1):

$$V_{\mathrm{b}}(r_{ij}) = \frac{1}{2}k^{b}_{ij}(r_{ij} - b_{ij})^2 \tag{2.8}$$

While force

$$\mathbf{F}_i(\mathbf{r}_{ij}) = k^{b}_{ij}(r_{ij} - b_{ij})\frac{\mathbf{r}_{ij}}{r_{ij}} \tag{2.9}$$

Bond angle vibration in a three atoms i,j,k covalently bonded defined by an harmonic angle potential and force are:

$$V_a(\theta_{ikj}) = \frac{1}{2}k^{\theta}_{ijk}\left(\theta_{ikj} - \theta^{0}_{ikj})\right)^2 \tag{2.10}$$

$$\mathbf{F}_i = -\frac{dV_a(\theta_{ijk})}{d\mathbf{r}_i} \tag{2.11}$$

$$\mathbf{F}_k = -\frac{dV_a(\theta_{ijk})}{d\mathbf{r}_k} \qquad \theta_{ijk} = \arccos\frac{\mathbf{r}_{ij}\mathbf{r}_{jk}}{r_{ij}r_{kj}} \tag{2.12}$$

$$\mathbf{F}_j = -\mathbf{F}_i - \mathbf{F}_k \tag{2.13}$$

The dihedral angles are divided in propers and improper dihedral angles. The proper dihedrals are the angle ϕ defined between ijk and jkl planes, the zero point is settled as the cis conformation.

$$V_{\mathrm{d}}(\phi_{ijk}) = k_{\phi}(1 + \cos(n\phi - \phi_s) \tag{2.14}$$

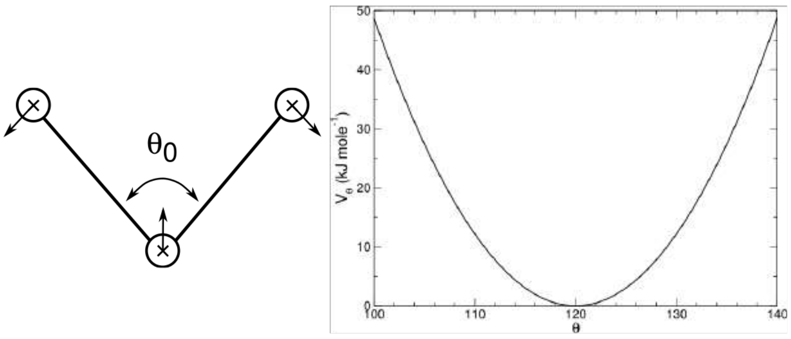


Figure 2.2: Bond angle vibration model and potential.

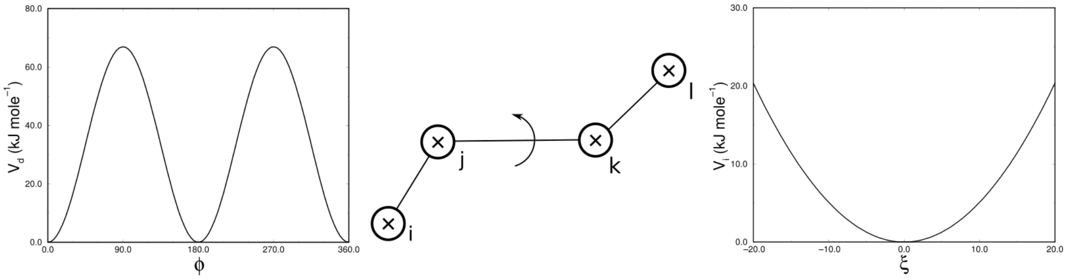


Figure 2.3: From the left to right: the improper dihedral potential; a model of the rotation on a dihedral angle; the proper dihedral potential

Improper dihedrals have the function to take keep aromatic rings, or other kind of planar group, planar. Improper dihedrals are able to make molecules flipping over the bonds conformation.

$$V_{\mathrm{id}}(\xi_{ijk}) = \frac{1}{2}k_{\xi}(\xi_{ijk} - \xi_0)^2 \tag{2.15}$$

2.1.3 Non-Bonded interaction

Non-bonded interaction are pair-additive and centro-symmetric:

$$V(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) = \sum_{i<j} V_{ij}(\mathbf{r}_{ij}) \tag{2.16}$$

$$\mathbf{F}_{ij} = -\sum_j \frac{dV_{ij}(r_{ij})}{dr_{ij}} \frac{\mathbf{r}_{ij}}{r_{ij}} \tag{2.17}$$

Lennard-Jones potential (Fig. 2.4) between a couple of atoms is:

$$V_{\mathrm{LJ}}(r_{ij}) = \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}} \tag{2.18}$$

or

$$V_{\mathrm{LJ}}(r_{ij}) = 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.19}$$

where in Eq. 2.18, the parameters $C_{ij}^{(12)}$ and $C_{ij}^{(6)}$ are defined on pair of atom, meanwhile in Eq. 2.19, ε_{ij} is the depth of the potential well and σ_{ij} is the finite distance at which potential is zero. Straightforwardly, force derived by potential is:

$$\mathbf{F}_i(\mathbf{r}_{ij}) = \left(12\frac{C_{ij}^{(12)}}{r_{ij}^{13}} - 6\frac{C_{ij}^{(6)}}{r_{ij}^{7}} \right) \frac{\mathbf{r}_{ij}}{r_{ij}} \tag{2.20}$$

Coulomb potential (Fig. 2.4) generated between a couple of charged particles is given by:

$$V_{\mathrm{Cou}}(r_{ij}) = k_e \frac{q_i q_j}{\varepsilon r_{ij}} \tag{2.21}$$

Force derived is

$$\mathbf{F}_{\mathrm{Cou}}(\mathbf{r}_{ij}) = k_e \frac{q_i q_j}{\varepsilon r_{ij}^2} \frac{\mathbf{r}_{ij}}{r_{ij}} \tag{2.22}$$

where k_e is the Coulomb constant[1].

Restraint potential are useful when is necessary to impose stiffness to the system to reproduce particular behavior i.e. restrain motion in a protein to force its native conformation. There are different type of restraint like: *position restraint*, *flat-bottomed position restraints*, *angle restraints*, *dihedral restraints*, and *position restraint*.

2.1.4 Free energy

For the aim of the thesis, the non-bonded interaction between the initial state A and the dummy state B (Chapter 3.1) and the soft-core interaction were interpolated

[1] $k_e = \frac{1}{4\pi\varepsilon_0} = 8.99 \times 10^9\,\mathrm{Nm^2C^2}$

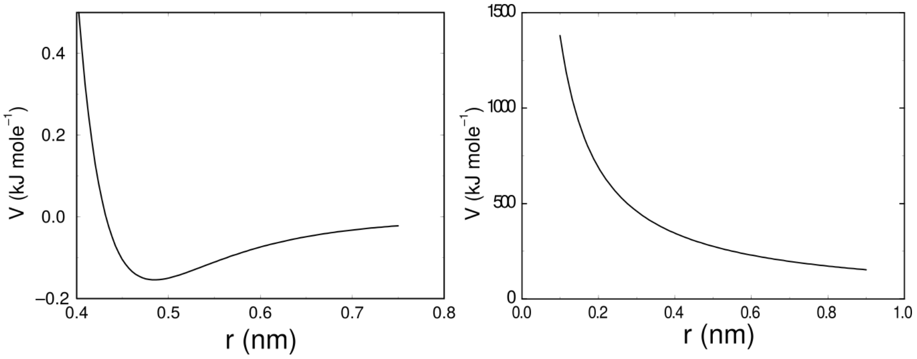


Figure 2.4: Non bonded interaction. On the left, the Lennard-Jones potential; on the right Coulomb potential.

using a soft-core interaction. Consequently, bonded and non-bonded potential need to be restated taking the λ parameter in to account. The harmonic potential, the angle potential, and the improper dihedral angle are calculated in the spirit of the bond potential:

$$V_b = \frac{1}{2}\left[(1-\lambda)k_b^A + \lambda k_b^B\right]\left[b - (1-\lambda)k_0^A + \lambda k_0^B\right] \tag{2.23}$$

while for proper dihedral the equation is

$$V_d = \left[(1-\lambda)k_d^A + \lambda k_d^B\right]\left\{1 + \cos\left[n_\phi \phi - (1-\lambda)\phi_s^A + \lambda \phi_s^B\right]\right\} \tag{2.24}$$

In this thesis work, the non bonded interaction, as Coulomb and Lennard-Jones, between molecules and solvent box has been subjected on a free energy routine. The Coulomb and Lennard-Jones interactions variate on λ dependence as:

$$V_C = \frac{1}{\varepsilon_{rf} r_{ij}}\left[(1-\lambda)\, q_i^A q_j^A + \lambda q_i^B q_j^B\right] \tag{2.25}$$

$$V_{\mathrm{LJ}} = \frac{(1-\lambda)C_{12}^B + \lambda C_{12}^B}{r_{ij^{12}}} - \frac{(1-\lambda)C_6^B + \lambda C_6^B}{r_{ij}^6} \tag{2.26}$$

Soft-core interaction and non-bonded interaction

The linear interpolation between of the non-bonded interaction as Lennard-Jones or Coulomb potentials weakly converge when particle are approaching to disappear as in dummy-particles formation. When λ value is near the limit to be zero or one the interaction energy can be so feeble to permit particles collapse one against the other. The soft-core potentials remove this kind of undesirable singularities in the potential. The specific non bonded interaction function between two atoms i and j is:

$$V_{ij}(r) = (1-\lambda)V_{ij}^{A}(r_{ij}^{A}) - \lambda V_{ij}^{B}(r_{ij}^{B}) \tag{2.27}$$

where

$$V_{ij}^{A}(r_{ij}^{A}) = \left(\frac{C_{ij}^{(12)}}{(r_{ij}^{A})^{12}} - \frac{C_{ij}^{(6)}}{(r_{ij}^{A})^{6}} \right) + k_e \frac{q_i q_j}{r_{ij}^{A}} \tag{2.28}$$

$$V_{ij}^{B}(r_{ij}^{B}) = \left(\frac{C_{ij}^{(12)}}{(r_{ij}^{B})^{12}} - \frac{C_{ij}^{(6)}}{(r_{ij}^{B})^{6}} \right) + k_e \frac{q_i q_j}{r_{ij}^{B}} \tag{2.29}$$

$$r_{ij}^{A} = (\alpha(\sigma_{ij}^{A})^{6}\lambda^{p} + r_{ij}^{6})^{1/6} \tag{2.30}$$

$$r_{ij}^{A} = (\alpha(\sigma_{ij}^{B})^{6}(1-\lambda)^{p} + r_{ij}^{6})^{1/6} \tag{2.31}$$

where $V_{ij}^{A}(r_{ij}^{A})$ and $V_{ij}^{B}(r_{ij}^{B})$ are the potential define the non-bonded interaction between atoms i and j separated by a distance r_{ij} relative to the state A or the state B. Van der Waals parameters for repulsions and the dispersion energy terms are, respectively, $C_{ij}^{(12)}$ and $C_{ij}^{(6)}$,k is equal to $1/(4\pi\varepsilon_0)$ and q_i and q_j are the atom charges. The soft-core interaction parameters are the soft-core parameter p, the soft-core λ power p, and the radius of the interaction σ, which is $(C^{(12)}/C^{(6)})^{1/6}$ or an input parameter when $C^{(6)}$ or $C^{(12)}$ is zero

2.1.5 General simulation routine

MD simulation can be divided in several steps as reported in Fig. 2.5.

Initial condition To actuate the simulation, algorithm needs the topology information, as which atoms and which combination of atoms are participating, and a description of the force field. The box size, coordinates and velocities are necessary. The box shape is defined by three vectors $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{b}_3$. For initiate the run the coordinates and $t = t_0$ must be known. Dedicated algorithm update the time step by Δt, also needs velocities at $t = t_0 - \frac{1}{2}\Delta t$. If velocities are unknown, those are generated with a given absolute temperature T:

$$p(v_i) = \sqrt{\frac{m_i}{2\pi k_B T}} \exp\left(\frac{m_i v_i^2}{2k_B T}\right) \tag{2.32}$$

The total energy will be different to the required temperature T, this step need to be recalculated removing the motion of the center-of-mass and rescaling all velocities so that the total energy correspond exactly to T.

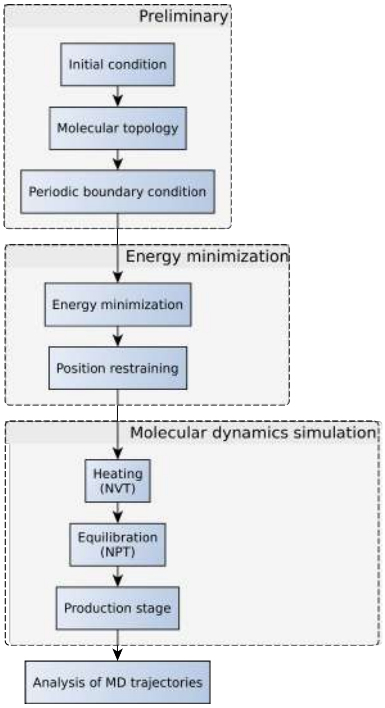


Figure 2.5: General scheme of a molecular dynamics simulation.

Neighbor searching Internal forces are defined by a tabulated list or from dynamic lists given by the non-bonded interaction between any pair of particles. The non-bonded pair interaction is calculated only if those pairs i, j is less than a given cut-off radius R_c The pair list include particles i, a displacement vector for i and j. This list is updated every n steps.

Energy minimization Potential energy of N-interacting particle can be mapped in a so called *potential energy surface* or *hypersurface*. In geometrical way the energy landscape is a representation of energy function across the configuration space of the system. In a system with N particles, energy is a function of $(3N-6)$ internal or $3N$ cartesian coordinates. Energy landscape has a global minimum, that correspond to the stable system, and a collection of local minima. To identify the configuration with minimum energy that correspond to the points in the configuration space is necessary use a minimization algorithm. Energy minimization process allow to find the *nearest local minimum* exploring the energy landscape. Minima are located using numerical methods changing the system coordinate gradually in the way to iteratively restart from a configuration with lower energy. Hence, from a starting configuration the nearest minima is that one can be achieve methodically by the steepest local gradient. The minimization problem can be stated as find a minimum value of a function f of a function which depends on more than one independent variables as $x_1, x_2, \ldots, x_N$. This topological places is defined when the first derivative of the function respect of all the variables is zero and the second derivatives are all positives:

$$\frac{\partial f}{\partial x_i} = 0; \qquad \frac{\partial^2 f}{\partial x_i^2} > 0 \qquad \text{where} i = \pm 1, 2, \ldots, N \tag{2.33}$$

With the second derivatives we can define a square $N \times N$ matrix called *Hessian matrix*. This matrix has non-negative eigenvalues at the local minima. Between local minima can be defined the *saddle point* where eigenvalue is zero and through this points system can migrate from one local minimum to another. Energy minimization using different algorithm find the minima in the energy landscape like steepest descent, conjugate gradients, or l-bfgs.

The steepest descent algorithm is a first-order minimization method that slowly change the coordinates of the atoms in the system in the aim to find a minima. Steepest descent move in the direction parallel to the net force calculated in the previously. This algorithm needs a initial displacement h_0 to start. Using a geographical analogy the steepest descent move in downhill direction. The vector $\mathbf{r}$ is defined by $3N$ coordinates. First forces $\mathbf{F}$ and potential energy are calculated.

$$\mathbf{r}_{n+1} = \mathbf{r}_n + \frac{\mathbf{F}_n}{|\max(\mathbf{F}_n)|} h_n \tag{2.34}$$

where h_n is the maximum displacement and $\mathbf{F}_n$ is the force calculated by the negative gradient of the potential V. The value $|\max(\mathbf{F}_n)|$ is the maximum of the absolute module of the force. The algorithm or when the maximum of the absolute value of the force is smaller than a specified value.

Periodic Boundary conditions MD simulation system space are usually defined in boxes with different shape and dimension. In the way to decrease the artifact generated by this finite box system, periodic boundary condition are applied. MD compute the interaction with particle taking in to account copy of itself translated in a space-filling box.

Heating The heating steps impose a temperature at the system as in real experiment is done using a thermostatted heat bath. In this condition, the probability to find the system in a defined energy is given by the Boltzmann distribution:

$$f(p) = \left(\frac{m}{2\pi k_B T}\right)^{3/2} \exp\left[-\beta \frac{p^2}{2m}\right] \tag{2.35}$$

in molecular dynamics temperature is generally calculate using the relation between the temperature T and the and the kinetic energy per particle:

$$\frac{3}{2} k_B T = \frac{1}{2} m \left\langle v_j^2 \right\rangle \tag{2.36}$$

where m is the mass v_j is the j-th component of the velocity. This calculation give the temperature per particle in the system. This value is not the the system temperature, as the constant temperature in the system is not equivalent to the the kinetic energy per particle. The variance of the kinetic energy per system in the system can be cancel out if the kinetic energy for particle is defined equal to the average value of kinetic energy. The kinetic energy variance is given by:

$$\frac{\sigma_{T_K}^2}{\langle T_K \rangle_{NVT}^2} = \frac{\langle T_K^2 \rangle_{NVT} - \langle T_K \rangle_{NVT}^2}{\langle p^2 \rangle_{NVT}^2} = \frac{2}{3N} \tag{2.37}$$

Constant temperature ensemble needs a thermostats algorithm designed to help the simulation by modulating the temperature of the system in some desired temperature. Thermostat does not keep the temperature constant in time but, instead, ensure that the average temperature of a system is correct. Sampling in some small region in the system the kinetic energy of this small number of particles fluctuate. Increasing the sample in a larger and larger number of particles, the fluctuations in the average get smaller and smaller, so for the whole system constant temperature. So the role of a thermostat as ensuring that we have the correct average temperature, and fluctuations of the correct size.

Equilibration The constant pressure (NPT) procedure regulates the system a pressure. The gain in kinetic energy defined in the heating step generate a pressure that can not be aligned with the prefixed experimental condition. In the same spirit of the thermostat algorithm a pressure coupling force the system to be coupled in a pressure bath. This type of algorithms rescale the box vector size in the aim of reduce the pressure fluctuation around a defined average pressure.

Production run In this step the equilibrated system is monitored for defined time of simulation. In this steps information about the trajectories and system property are analyzed.

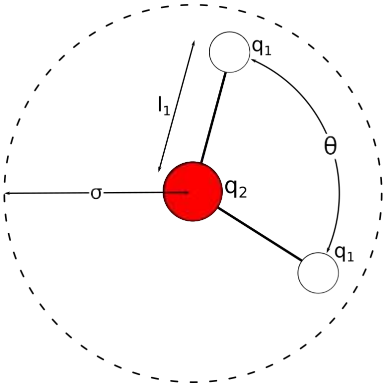


Figure 2.6: SPC water model. $q_1 = +0.41$ and $q_2 = -0.82$ are the charges of the hydrogen and oxygen, respectively. σ is the Lennard-Jones distance at which the inter-particle potential is zero, θ is the angle HOH, and l_1 is the distance oxygen-hydrogen.

2.1.6 Water models

All the simulations in this work has been done using water as solvent. More than 40 models has been developed in the aim to reproduce the properties of real water. Water models can be divided in four macro categories taking into account the geometry and the number and kind of parameters adopted. Model complexity improve the quality of calculation but needs more computational time due of the increased number of coordinates used to calculate the interaction energy. Water model potential energy is

$$V_{\text{Water}}(r_{ij}) = \sum_{i<j} k_e \frac{q_i q_j}{\varepsilon r_{ij}} + \sum_{i<j} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}^{(\text{O})}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}^{(\text{O})}} \right)^{6} \right] \tag{2.38}$$

where $r_{ij} = |\mathbf{r_j} - \mathbf{r_k}|$ is the distance between to atoms i and j, $r_{ij}^{\text{O}} = |\mathbf{r_j}^{(\text{O})} - \mathbf{r_k}^{(\text{O})}|$ is the distance between to oxygen atoms i and j. In this thesis has been used the SPC model for water. This model is a three sites model where each site is provided of charge and describe an atom in the molecule.

The SPC model use the tetrahedral angle 109.47° instead of the observed one 104.32°.

	SPC
σ (Å)	3.166
ε ($\mathrm{kJ\,mol^{-1}}$)	0.650
I_1 (Å)	1.0
q_1 (e)	$+0.41$
q_2 (e)	-0.82
$\theta_{\mathrm{H_2O}}$ (deg)	109.47

Table 2.1: Characteristic of SPC water. σ is the Lennard-Jones distance at which the inter-particle potential is zero, ε is the depth of the potential well, i_1 is the inter atomic distance O$-$H, q_1 and q_2 are the atoms charge and θ is the angle HOH.

2.2 Monte Carlo simulations

Monte Carlo simulation make random changes to position of a system with the aim of generate new appropriate configuration. The main difference with molecular dynamics is that Monte Carlo methods do not implement changes momenta but only samples from the $3N$-dimensional space of position in the system.

Taking in account canonical ensemble partition function Q for N identical particles:

$$\mathrm{Q}(N,V,T)=\frac{1}{N!}\frac{1}{h^{3N}}\int\int d\mathbf{r}^N d\mathbf{p}^N \exp\left[-\beta H(\mathbf{r}^N,\mathbf{p}^N)\right] \tag{2.39}$$

where $N!$ indistinguishably of particles in the system. $H(N,V,T)$ is the Hamiltonian that correspond to the total energy in the system. This depends on $3N$ positions and $3N$ momenta, so the Hamiltonian can be written as the sum of the kinetic and potential contribute in the system.

$$H(N,V,T)=\sum_{l=1}^{N}\frac{|\mathbf{p}_l|^2}{2m}+V(\mathbf{r}^N) \tag{2.40}$$

The Eq. 2.39 can be splitted in two separate integrals since the kinetic and potential are independent each other. Kinetic contribute depends only in momenta and potential on position.

$$\mathrm{Q}(N,V,T)=\frac{1}{N!}\frac{1}{h^{3N}}\int d\mathbf{p}^N \exp\left[-\beta\frac{|\mathbf{p}_l|^2}{2m}\right]\int d\mathbf{r}^N \exp\left[-\beta V(\mathbf{r}^N)\right] \tag{2.41}$$

The integral over momenta results:

$$\int d\mathbf{p}^N \exp\left[-\beta\frac{|\mathbf{p}_l|^2}{2m}\right]=(2\pi m k_B T)^{3N/2} \tag{2.42}$$

The integral over positions denoted as configuration integral $Z(N,V,T)$ results :

$$Z(N,V,T) = \int d\mathbf{r}^N \exp\left[-\beta V(\mathbf{r}^N)\right] \tag{2.43}$$

For non interacting particles potential energy function is zero. The exponential, straightforwardly, equal to 1 taking the the integral of 1 over the coordinate of each atoms is the volume of the system for N non interacting particles. This result as:

$$Q(N,V,T) = \frac{V^N}{N!}\left(\frac{2\pi k_B T m}{h^2}\right)^{3N/2} \tag{2.44}$$

This can also written in terms of the Broglie thermal wavelength Λ [2]:

$$Q(N,V,T) = \frac{V^N}{N!\Lambda^{3N}} \tag{2.45}$$

From the Eq. 2.42 and 2.43 arise that the the partition function can be splitted in two contributes: the momenta contribute, due by the gas behavior, and a contribute generated by the interaction. The partition function for a system is settle up with a contribution due to momenta and a contribution between particles. All the deviation from an ideal gas are derived to the existence of interaction between the atoms in the system. This energy is correlated only with position and not with momenta and Monte Carlo method is able to calculate the interaction contribution. Random method of sampling can be a possible alternative. To determinate the area under function curve we need to fix bounding area in which operate sampling. The ratio of the number of points under the curve to the total number of points, multiplied the sampling area estimate the area under the curve.

For potential energy is not possible solve analytically integral commonly used in molecular modeling. In this limit case we need to find another path to evaluate the integrals using numerical methods like the trapezium rule or the Simpson's rule. Functions of two variables (x, y), for an accurate approximation need to square the number functions evaluations, so for a $3N$-dimensional integral the total number or evaluations will be in the order of j^{3N} where j is the particles number in a directions. For a 100 particles and 2 points for dimension we obtain 2^{200} evaluations!

2.3 Metropolis Method

In 1953 Metropolis et al. proposed a strategy to generate configuration that make a relevant contribution to the integral (Eq. 2.43). The Metropolis algorithm is a

[2] $\Lambda = \sqrt{h^2/2\pi k_B T}$

Markov chain Monte Carlo (MCMC) for obtaining a sequence of random samples. A Markov chain is stochastic model characterized by a finite set of events in which the probability of each event depends only on to the preceding one. This condition designate a difference with the molecular dynamics method in which all states are connected in time. This process is able to random generate a set of configuration with a probability proportional to the Boltzmann factor in a given temperature. Metropolis method routine change the phase space, from the configuration j in the system making some predefined moves and test the new configuration $j + 1$. Test result accept the generated configuration $j+1$ if the energy is lower than the configuration j, otherwise if the the energy increases, it accepted only with probability:

$$p(\Delta E) = \exp\left[-\frac{\Delta E}{k_B T^*}\right] \tag{2.46}$$

where ΔE is the energetic gain respect of the configuration j and T^* is a fictitious temperature.

Monte Carlo algorithm can be applicate to perform simulation of flexible molecules. The idea is based in applying random changes in the Cartesian coordinates of the atoms forming the molecule. This new configuration can be accepted by a acceptance ratio defined a priori. The simplest model use a lattice approximation where the molecule is represented by the interaction of connected center. This center are enforced to be in the vertices of lattice. A more complex model are the beads model where polymer is composed by subsequent sphere beads with defined radius bead and connection distance between consecutive object. Beads represent an effective monomer and interact with the first neighbor beads and with the nearby one. The new configuration is generated using a variant of the algorithm. This algorithm use a combination of crankshaft, reptation, and pivot end rotation move.

Crankshaft move select two random beads i,j where $i < j$, and with the condition that $j - i \leq k + 2$ where $k \ll N$. The number k represent the number of beads that will be rotate as the $i + 1, \ldots j - 1$ of an angle ϕ respect of the x_i-x_j axes. The angle is randomly chose in an given interval. Crankshaft move is a *local* move since only a small portion of chain is changed. Pivot move select a random bead j internal to the beads chain ($1 \leq j \leq N - 1$). This selected point act as fulcrum keeping fixed the the beads $1, \ldots, j - 1$ and moving the $j + 1, \ldots, N$ chain of a random, range chose, value.

Chapter 3

Solvation free energy of amino acids side chains

As precedently reported, classic Kauzmann et al. work point out that the transfer free energy between water and organic phase is favorable for hydrophobic side chain analogs. Translated into the so-called hydrophobic effects in protein folding that tends to bury hydrophobic side chain inside the protein shielded from water. In parallel experimental studies as reported in Chapter 1 underling the an additional interest in the understanding the effect of a specific solvent in protein stability. Molecular dynamics is a powerful tool that can provide useful and detailed information on both the hydrophobic effect and the effect of solvent on protein stability. On the other hand, solvation free energy is most conveniently tackled using thermodynamic integration rather then brute force molecular dynamics. In thermodynamics integration, the free energy change of a solute upon gradually turning on the presence of the solvent, is monitored. In this work, we have performed such calculation for 18 amino acids, ranging from the most hydrophobic to the charged ones, and compared with previously known results, both experimental ans numerical. This calculation could be repeated for different solvents and small peptides, and is related to the approximate method in Chapter 4. Thermodynamic free energy is the amount of energy available in a system that can be converted in thermodynamic work. This quantities are useful values to understand the system at equilibrium. As precedently described, protein folding, as binding between molecules, or diffusion through a membrane are regulated by this energy. Free energy profile are useful to intimate understand how a biological or chemical process take place and put in evidence the kinetic and dynamic property. This type of calculation combined with numerical simulation are used in several areas as pharmacology, biochemistry, biotechnology and so on.

3.1 Free energy

For a closed system σ with energy E, volume V, and number of particles N, the entropy S is at the maximum value only when the system is at equilibrium. Consequently, this system is not able to exchange heat, volume, or particles with a reservoir. Two different function of state can be defined applying either a thermostatted bath,fixing the temperature T, the volume V and the number of particle N so that the Helmholtz free energy $F = E - TS$ is at minimum at equilibrium, or couple the pressure to fix the pressure, and number of particles to set at minimum the Gibbs free energy $G = E + PV$. Comparing two system σ_{T_1} and σ_{T_2} with different temperature T_1 and T_2 defined coupling a thermostatted bath and calculating the Helmholtz free energy, know which system is more stable. In classical mechanics the Helmholtz free energy F is

$$F(N, V, T) = -k_B T \ln Q = -K_B T \ln \left(\frac{\int d^3p d^3r \exp[-\beta H(p, q)]}{\Lambda N!} \right) \tag{3.1}$$

where the Helmholtz free energy F is related with the partition function Q. Thermal quantities cannot be calculated directly neither in real experiment nor in numerical calculation. Experiment give as result the derivative of a free energy respect to the volume V or temperature T:

$$\left(\frac{\partial F}{\partial V} \right)_{NT} = -P \tag{3.2}$$

and

$$\left(\frac{\partial \frac{F}{T}}{\partial \frac{1}{T}} \right)_{VN} = E \tag{3.3}$$

The resulting pressure P and energy E can be measured in numerical calculation. Free energy of a system in a defined temperature and density can be calculated finding the path in V-T plane that connect the unknown state to a know one. The change in F is calculated with thermodynamic integration, integrating the equations 3.2 and 3.3. To do this we need to define a referring state. The idea is use an ideal gas phase for which free energy state as the thermodynamic state are always known. Numerical simulation permit to avoid this step building a state of reference modifying the parameter that define the model feature.

Generally speaking,another purpose of a free energy calculation is calculate relative energy between two different system like a receptor and a receptor with the binded molecule, the difference in free energy between these systems due by the docking is the energy of association. The difference in free energy between two states is

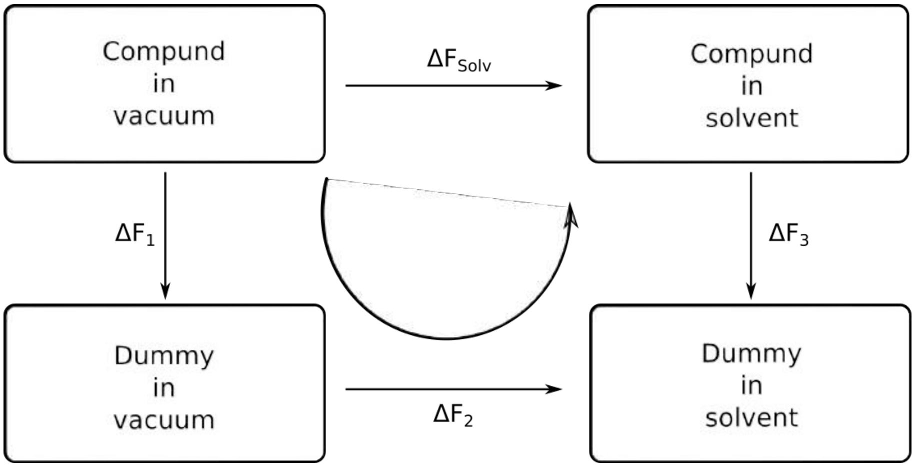


Figure 3.1: Thermodynamic cycle for the determination of solvation energy.

A and B is

$$\Delta F_{AB} = F_A - F_B = -\frac{1}{\beta} \ln \frac{Q_B}{Q_A} \tag{3.4}$$

where A and B are different conformations of the same system. The solvation free energy ΔF_{solv} can be defined as the work necessary to transfer a molecule from a gas phase to a solution. This value is computationally obtainable in a reasonable amount of time and with a reasonable accuracy. Situation aforementioned can be solved using a thermodynamic cycle and the Kirkwood's coupling parameter method [44].

In the thermodynamic cycle represented in Fig. 3.1 ΔF_3 is the work needed to remove the solute-solvent and solute-intramolecular interaction. The core idea resides in gradually mutate all the atoms in a given compound into a dummy atom. The dummy atom peculiarity is that has the non bonded interactions, as the Lennard-Jones, set to zero but bonded interaction were kept unaltered. ΔF_1 represent the energy required to remove all the internal non-bonded interaction of the compound in vacuo. ΔF_2 is the work necessary to transfer dummy from vacuo to the solvent phase. Difference between two states can be determinate using thermodynamic integration where Hamiltonian H is function of a parameter λ. The Hamiltonian of the system changes gradually from H_A to H_B. The coupling parameter of $H(p, q, \lambda)$ in order that $\lambda = 0$ describes system A and $\lambda = 1$ describes system B:

$$H(p, q; 0) = H^A(p, q) \tag{3.5}$$

$$H(p, q; 1) = H^B(p, q). \tag{3.6}$$

As precedently stated the partition function $Q(N, V, T)$ cannot be evaluated but it is possible to evaluate the derivative with respect to λ as an ensemble average

$$\Delta F(0 \rightarrow 1) = \int_0^1 d\lambda \frac{\partial F(\lambda)}{\partial(\lambda)} = -\int_0^1 d\lambda k_B T \frac{1}{\mathrm{Q}} \frac{\partial \mathrm{Q}}{\partial(\lambda)} \tag{3.7}$$

$$\frac{\partial \mathrm{F}}{\partial(\lambda)} = \frac{1}{N! h^{3N}} \iint dp\, dq \left(-\beta \frac{\partial H}{\partial \lambda}\right) \exp\left[-\beta H(p, q, \lambda)\right] \tag{3.8}$$

$$\frac{\partial F}{\partial \lambda} = \frac{\iint dp\, dq\, (\partial H / \partial \lambda) \exp\left[-\beta H(p, q, \lambda)\right]}{\iint dp\, dq \exp\left[-\beta H(p, q, \lambda)\right]} = \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{N,V,T,\lambda} \tag{3.9}$$

The difference in free energy between A and B can be found by integrating the derivative over λ:

$$F^B(V, T) - F^A(V, T) = \int_0^1 d\lambda \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{N,V,T,\lambda} \tag{3.10}$$

The λ-dependence of the potential in bonded interaction is liner while non-bonded interaction can be described with linear dependence or with softcore interaction.

3.2 Bennett's acceptance ratio method

Bennet in 1976 proposed a methodology to calculate the validate the free energy between two state A and B using Metropolis Monte Carlo simulation [45]. The same approach called Bennett's acceptance ratio (BAR) can be applied in thermodynamic integration [46].

$$\Delta G_{l \rightarrow m}^{(BAR)} = k_B T \left(\ln \frac{\langle f(H_l - H_m + C) \rangle_m}{\langle f(H_m + H_l - C) \rangle_l} \right) + C \tag{3.11}$$

where $f(x)$ is the Fermi function

$$f(x) = \frac{1}{1 + \exp\left(\frac{x}{k_B T}\right)} \tag{3.12}$$

H_l and H_m are the Hamiltonian in the states l and m. The constant C is calculated every iteration to fulfill the equivalence:

$$\langle f(H_l - H_m + C) \rangle_m = \langle f(H_m + H_l - C) \rangle_l \tag{3.13}$$

The free energy difference can be calculated as follows:

$$\Delta G^{(BAR)}_{m\to l} = -k_B T \ln \frac{N_m}{N_l} + C \tag{3.14}$$

and

$$\Delta G^{(BAR)}_{AB} = \sum_{A=l}^{n-1} \Delta G^{(BAR)}_{l+1,m} \tag{3.15}$$

where N_l and N_m are the the number of coordinates for the lambda parameter λ_l and λ_m, respectively. The accuracy of this processes can be insured only if there is a sufficient overlap between the energy integrals. This overlap can be monitored with the overlap integral.

3.3 Experimental data for the amino acids

The affinity of a molecule for an aqueous environment can be calculated with a vapor pressure calculation experiment. Experimental free energies solutes can be determinate calculating the concentration on two phase systems: the vapor phase with a partial pressure $P^{(\mathrm{Vap})}_m$ of some solute molecule of type m and the aqueous solution with a concentration of solute as $\rho^{(\mathrm{Aq})}_m$. When these two phase are in equilibrium with respect to the transfer of molecules of type between the phases, the solvation free energy is given by

$$\Delta G_{\mathrm{Solv}} = k_B T \ln \left(\frac{P^{(\mathrm{Vap})}_m}{\rho^{(\mathrm{Aq})}_m k_B T} \right) \tag{3.16}$$

when to state are in equilibrium the equilibrium constant K is given by

$$K = \frac{[\mathrm{state}^{(m)}_k]}{[\mathrm{state}^{(m)}_j]} \tag{3.17}$$

Using moles instead of the number of molecules, the k_B needs to be replaced with R. In the same way the use of concentration $\rho^{(\mathrm{Aq})}_m$ produce an extra $k_B T$ and the logarithm term.[1] where the equilibrium constant K is related to the free energy of solvation as follows:

$$\Delta G = -RT \ln(K) = -RT \ln \left(\frac{\mathrm{state}^{(m)}_k}{\mathrm{state}^{(m)}_j} \right) = RT \ln \left(\frac{\mathrm{state}^{(m)}_j}{\mathrm{state}^{(m)}_k} \right) \tag{3.18}$$

[1] The conversion from pressure to density, from number of molecules to moles, using $\frac{n}{V} = \frac{P}{RT}$,needs the use of k_B contrariwise the R .

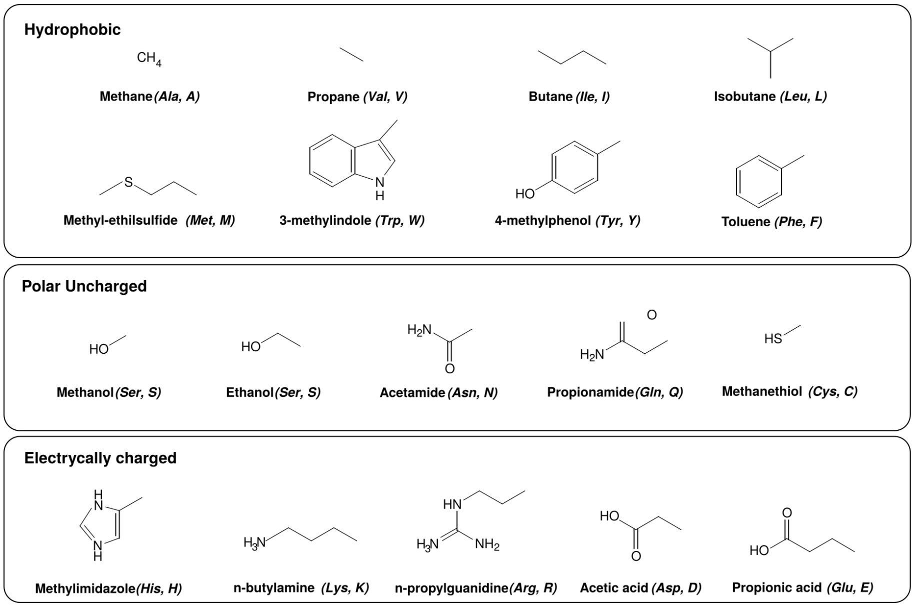


Figure 3.2: Amino acid side chain neutral analogs.

Measuring the aqueous concentration the vapor pressure can be deducted indirectly:

$$\Delta G = RT \ln \left(\frac{[m^{(\mathrm{aq})}]_i - [m^{(\mathrm{aq})}]_f}{[m^{(\mathrm{aq})}]_f} \right) = RT \ln \left(\frac{[m^{(\mathrm{aq})}]_i}{[m^{(\mathrm{aq})}]_f} + 1 \right) \tag{3.19}$$

where $[m^{(\mathrm{aq})}]_f$ and $[m^{(\mathrm{aq})}]_i$ are the final concentration, at equilibrium, and initial concentration, respectively.

3.4 Free energy calculation of single amino acid

The hydratation free energy has been calculated using thermodynamic integration as described in chapter (3.1). For each molecule the free energy calculation has been derived using n-different calculation as many λ points from $\lambda_A = 0$ (solute) to $\lambda_B = 1$ (dummy atoms). Compound investigated are the neutral analog of the amino acid side chain (Fig. 3.2). Each analog correspond to the side chain starting from the β-carbon. The truncated bond as been substituted with an hydrogen atom. All the compound investigated are in neutral form, so electrically charged amino acid as been neutralized adding or removing a proton.

3.4.1 Free energy calculation routine

The simulation were performed at constant temperature and pressure in cubic box. The force field used was gromos54a7 modified accordingly (from here onwards GROMOS(ATB)). This revised force field has been customized in bonded and non bonded interaction using building blocks and interaction parameter from Automated Topology Builder (ATB) [47–50]. The neighbor searching distance was defined as Verlet with a searching distance of 1.2 nm. Solvation has been set to build a shell of 1 nm of SPC water around the molecule shaping a ≈3 nm box with ≈900 water molecules. The non bonded interaction were set as 1.2 nm for the electrostatic parameter,the van der Waals set as Particle-Mesh Ewald (PME) and cutoff, respectively. The isothermal-isobaric ensemble at 300 K,a leap-frog stochastic dynamics integrator (sd) has been used to integrate the equations of motion. Pressure has been maintained constant using the Parrinello-Rahman barostat. Pressure has been set as 1 atm. The inverse friction constant as been set as 1.0 ps Water compressibility was set as $4.5 \times 10^{-5}\,\mathrm{bar}^{-1}$. Free energy parameters were evaluated starting with 0.05 increment in λ parameters (from 0.00 to 1.00) for a total of 20 λ. Thermodynamic integration has been calculated transforming Coulomb ans Van der Waals interaction with the same rate. Simulation was recalculated, increasing the value of λ steps, if Bennett Acceptance Ratio was not achieved. Decoupling parameter for soft-core α, σ and p parameter were set as 0.5, 0.3 and 1. The angles and the bond distances were constrained using LINCS algorithm with the highest order in the expansion of the constraint coupling matrix set as 12. Time steps in each simulation was 0.002 ps and 50000 steps for a total of 100 ps for NVT and NPT routines. Production run has been simulated with time steps of 0.002 ps and 500000 steps for a total of 1 ns

3.5 Results

Free energy calculation were performed from a fully interacting neutral analog to a non interacting molecule. Assuming the reversibility of the processes, so starting with a complete uncoupled compound in water and gradually turn off the interaction, the energy value change in sign. This transformations required a minimum of 20 λ points and a maximum of 34 λ. The calculated hydratation free energy statistical error is in the range $0.11\,\mathrm{kJ\,mol^{-1}}$ to $0.49\,\mathrm{kJ\,mol^{-1}}$. The process produced during the free energy calculation, compound interaction with solvent are gradually turned off until the dummy state. Simulation has been done at 300 K, on the other hand some experimental and calculated simulation has been done with different reference temperature (Villa et al. ($T = 293\,\mathrm{K}$); Wolfenden et al. ($T = (298.15\,\mathrm{K})$). Change

MOLECULE	Preset work	ATB	Villa et al.	Abraham et al.	Wolfenden et al.	Rizzo et al.
Methylimidazole (His)	$-28.48 \pm (0.11)$	-29.6 ± 0.9	$-27.4 \pm (1.2)$	-	-42.1	$-35.19 \pm (2.51)$
n-butylamine (Lys)	$-13,27 \pm (0,27)$	-	$-15,5 \pm (2,2)$	-	-18	-
n-propylguanidine (Arg)	$-44.6 \pm (0.49)$	-	$-30.1 \pm (2.4)$	-	-44.8	-
Acetic acid (Asp)	$-31.27 \pm (0.33)$	$31.4 \pm (1.4)$	$-18.2 \pm (1.1)$	-	-27.5	$-27.99 \pm (2.51)$
Propionic Acid (Glu)	$-34.66 \pm (0.33)$	$-34.8 \pm (1.5)$	$-16.2 \pm (1.1)$	-	-26.6	$-27.03 \pm (2.51)$

Table 3.1: Calculated and experimental free energy (kJ mol^{-1}) at 300 K for electrically charged amino acid side chain neutral analogs in water.

MOLECULE	Preset work	ATB	Villa et al.	Abraham et al.	Wolfenden et al.	Rizzo et al.
Methane (Ala)	$8.54 \pm (0.12)$	4.27 ± 1.0	$9.2 \pm (0.6)$	6.27 ± 0.6	8.1	-
Propane (Val)	$7.93 \pm (0.14)$	6.6 ± 1.0	$10.7 \pm (1.2)$	8.37 ± 0.8	8.2	-
Butane (Ile)	$8.36 \pm (0.17)$	7.7 ± 1.2	$10.7 \pm (1.0)$	8.79 ± 0.84	8.8	-
Isobutane (Leu)	$8.72 \pm (0.17)$	$6.7 \pm (1.0)$	$10.4 \pm (1.1)$	9.6 ± 0.8	9.4	-
Methyl-ethylsulfide (Met)	$-0.58 \pm (0.28)$	$-9 \pm (0.9)$	$-5.5 \pm (1.0)$	-	-6.1	-
4-methylphenol (Tyr)	-	$-25.5 \pm (0.5)$	$-22.4 \pm (2.2)$	-	-25.2	-
Toluene (Phe)	$1.11 \pm (0.23)$	$1.1 \pm (0.8)$	$3.4 \pm (1.3)$	$-3.8 \pm (0.84)$	-3.1	-

Table 3.2: Calculated and experimental free energy (kJ mol^{-1}) at 300 K for hydrophobic amino acid side chain neutral analogs in water.

in temperature increase the energy of approximately 0.006 kJ mol^{-1} and has been defined as irrelevant.

Tables 3.1, 3.2, and 3.5 shows that in general calculated value using GROMOS(ATB) have comparable value with the calculated [47, 51] and experimental value [52–54].

Methyl-ethylsulfide (Met) and 4-methylphenol (Tyr) simulations showed more difficult and we have not be able to give a valid result. In general, compounds with a sp^2 hybridization like toluene (Phe), 4-methylphenol (Tyr), and 3-methylindole (Trp) have different value respect of the experimental value but, on the other hand, are generally in agreement with the calculated reference. Electrically charged amino acid side chain neutral analogs in water, without taking into account the amino acids with a sp^2, seems to be slightly over estimated in negative way given by a

MOLECULE	Preset work	ATB	Villa et al.	Abraham et al.	Wolfenden et al.	Rizzo et al.
Methanol (Ser)	$-27.75 \pm (0.19)$	-22.2 ± 1.4	$-14.1 \pm (0.9)$	6.27 ± 0.6	-20.8	-20.8 ± 2.51
Ethanol (Thr)	$-21.37 \pm (0.2)$	-19.1 ± 0.8	$-13.7 \pm (1.1)$	8.37 ± 0.8	-21.1	-20.92 ± 2.51
Acetamide (Asn)	$-41.38 \pm (0.33)$	-41.8 ± 0.6	$-18.8 \pm (1.7)$	8.79 ± 0.84	-39.9	-40.23 ± 2.51
Propionamide (Gln)	$-44.66 \pm (0.09)$	$-40.3 \pm (0.9)$	$-18.7 \pm (2.0)$	9.6 ± 0.8	-38.7	-
Methanethiol (Cys)	$-8.38 \pm (0.11)$	-	$5.5 \pm (1.0)$	-5.02 ± 0.84	-5.1	-
3-methylindole (Trp)	$-19.66 \pm (0.14)$	-	$-12.3 \pm (1.9)$	-	-24.3	$-25.65 \pm (2.51)$

Table 3.3: Calculated and experimental free energy (kJ mol^{-1}) at 300 K for polar uncharged amino acid side chain neutral analogs in water.

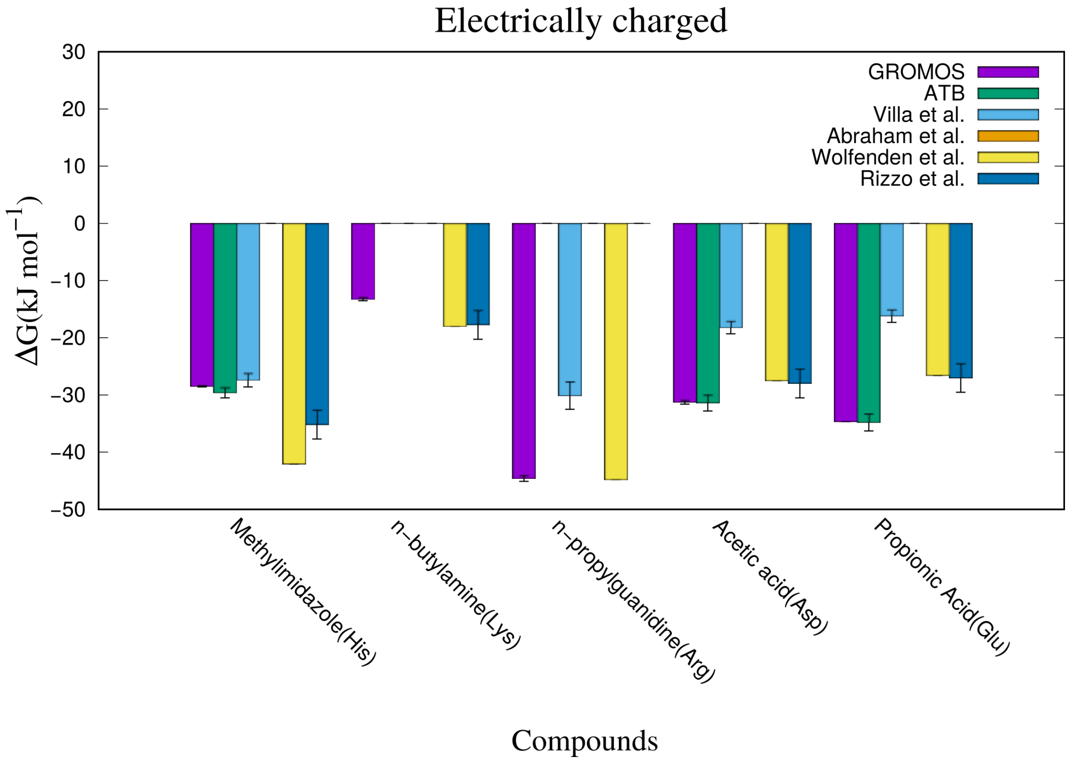


Figure 3.3: Calculated and experimental free energy (kJ mol^{-1}) at 300 K for charged amino acid side chain neutral analogs in water.

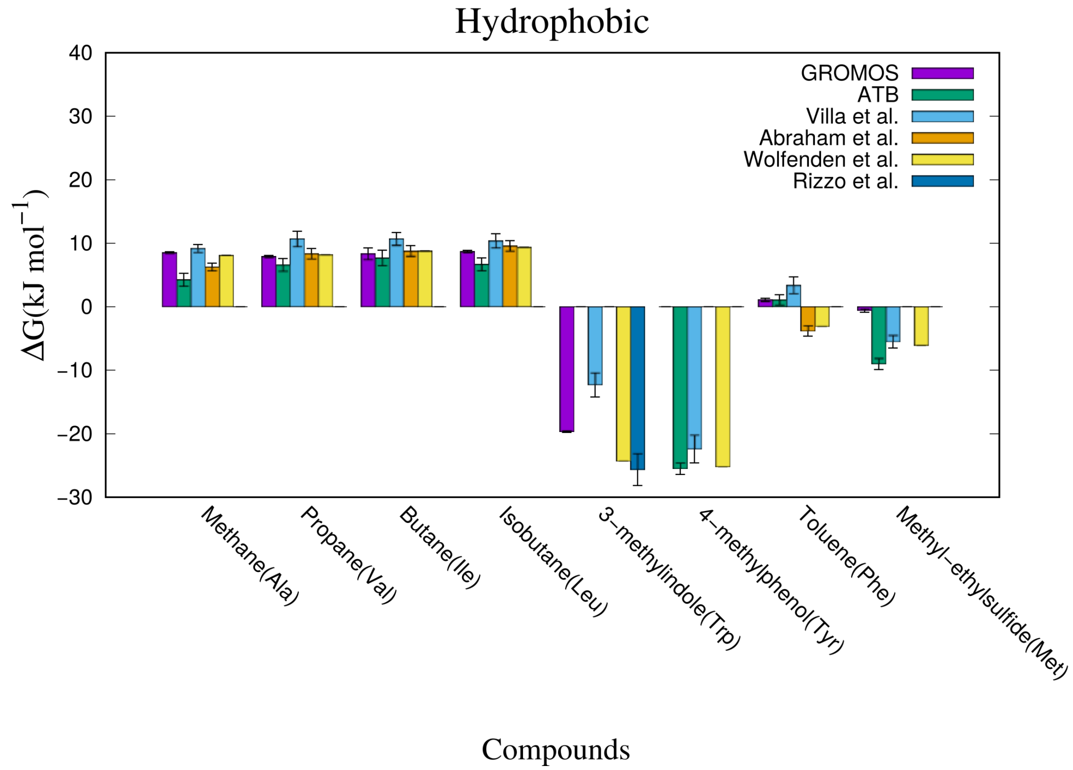


Figure 3.4: Calculated and experimental free energy (kJ mol^{-1}) at 300 K for hydrophobic charged amino acid side chain neutral analogs in water.

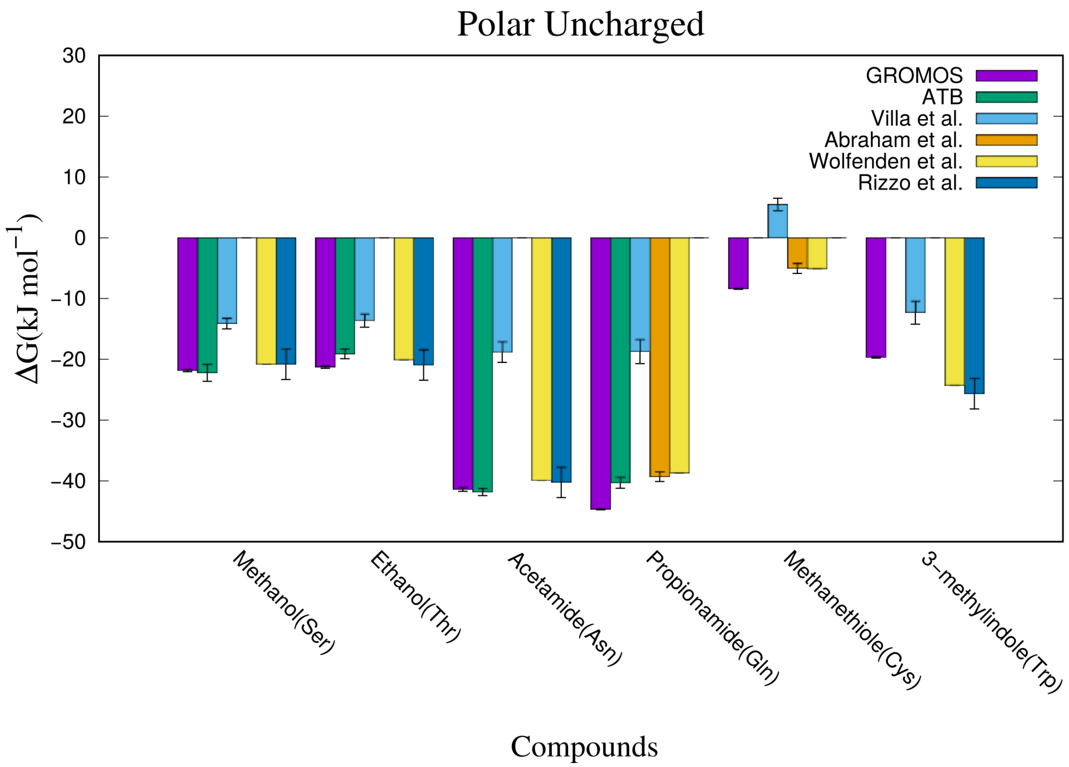


Figure 3.5: Calculated and experimental free energy (kJ mol^{-1}) at 300 K for polar uncharged amino acid side chain neutral analogs in water.

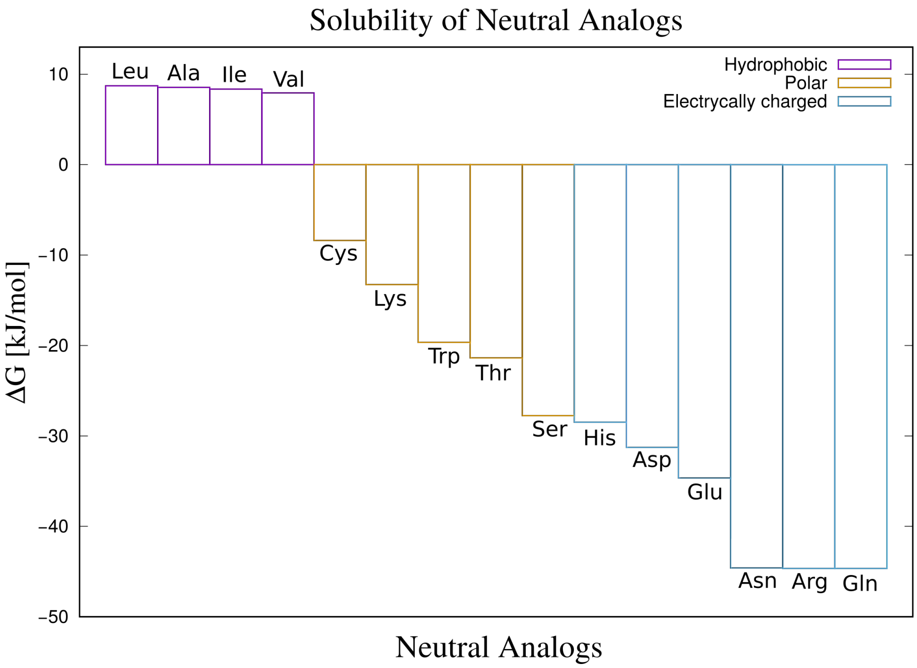


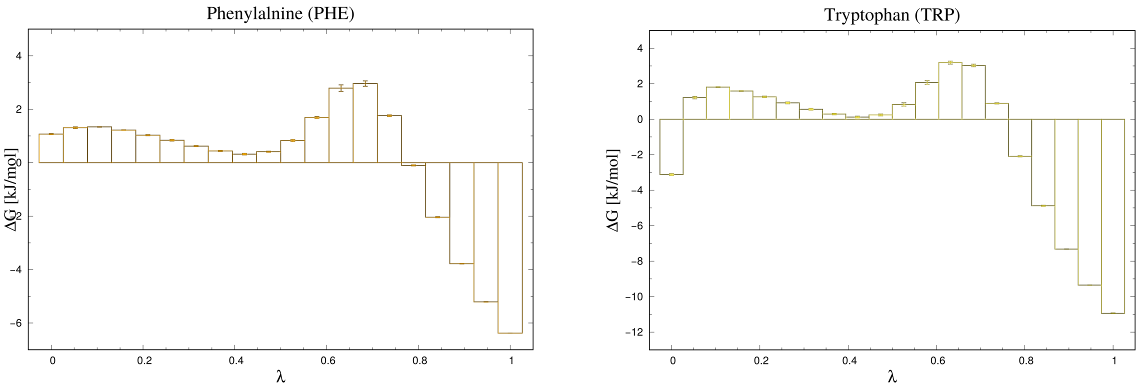
Figure 3.6: Calculated free energy of solvation using GROMOS(ATB) sorted in solubility order.

force field distortion that induces the analogs to by more hydrophilic then expected. In contrast, polar uncharged amino acids, are in good agreement with the experimental value. Methyl-ethylsulfide (Met), as methanethiol (Cys), show an anomalous behavior probably due to the tioeter group R−S−R.

Figs. from 3.8 to 3.10 show the cumulative ΔG as function of λ. The value of cumulative value in ΔG is obtained by the sum of two consecutive λ value. If, for instance, $\lambda = 0$ is $0.3\,\mathrm{kJ\,mol^{-1}}$ and in $\lambda = 0.02$ is $0.5\,\mathrm{kJ\,mol^{-1}}$ the cumulative value is $0.8\,\mathrm{kJ\,mol^{-1}}$. Formally,it is the integral of all the ΔG during the uncoupling process:

$$\int_0^{\lambda} \left(\frac{dG}{d\lambda'} \right) d\lambda' \tag{3.20}$$

Bennet acceptance ratio algorithm generate useful information to understand the interaction during the decoupling. From Fig. 3.8 toluene (Phe) and 3-methylindole (Trp) share the same pattern as highlighted in Fig. 3.7. These structures share a six-membered benzene ring, but tryptophan analog $\lambda = 0$ is negative ($-3.12\,\mathrm{kJ\,mol^{-1}}$). This affinity could be generated by the secondary amine in the pyrrolic ring able to form a H-bond with water. Decoupling process change this value in one step leading the 3-methylindole (Trp) to ensue the patter of a the hydrophobic Toluene (Phe).



(a)

(b)

Figure 3.7: (a) $dG/d\lambda$ for toluene (Phe) with $\lambda = 20$; (b) $dG/d\lambda$ for 3-methylindole (Thr) with $\lambda = 24$. This compounds have, qualitatively, the similar decoupling process.

Methane (Ala), propane (Val), butane (Ile), isobutane (Leu), and methyl-ethylsulfide (Met) in Fig. 3.8 share same cumulative ΔG pattern as in the ΔG function of λ value (Figs. 3.11(a) to 3.11(e)). These compounds are generally hydrophobic and not so

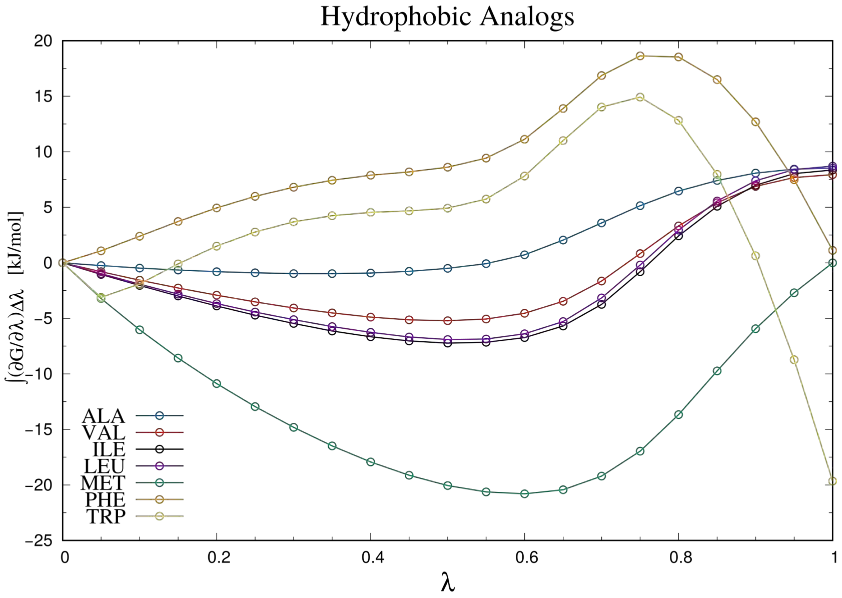


Figure 3.8: $\int_0^\lambda \left(\frac{dG}{d\lambda'}\right) d\lambda'$ for the Hydrophobic analogs: methane (Ala), propane (Val), isoleucine (Ile), isobutane (Leu), methyl-ethylsulfide (Met), toluene (Phe), and 4-methylphenol (Trp).

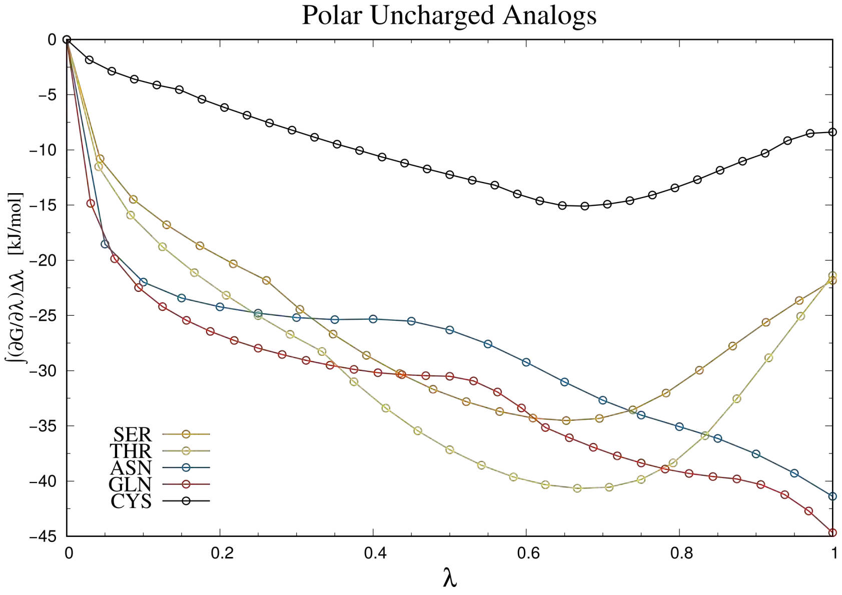


Figure 3.9: $\int_0^\lambda \left(\frac{dG}{d\lambda'}\right) d\lambda'$ for the Polar uncharged analogs: methanol (Ser), ethanol (Thr), acetamide (Asn), propionamide (Gln), and methanethiol (Cys).

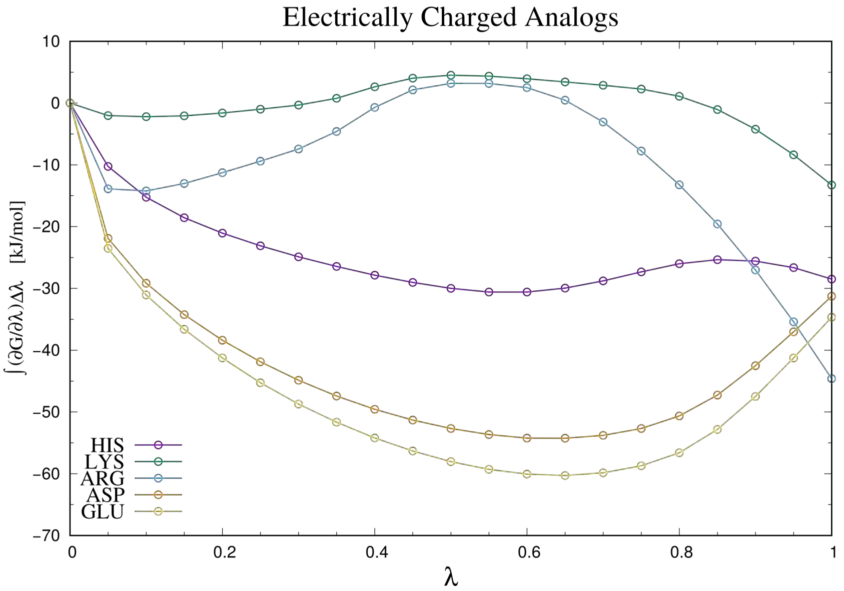


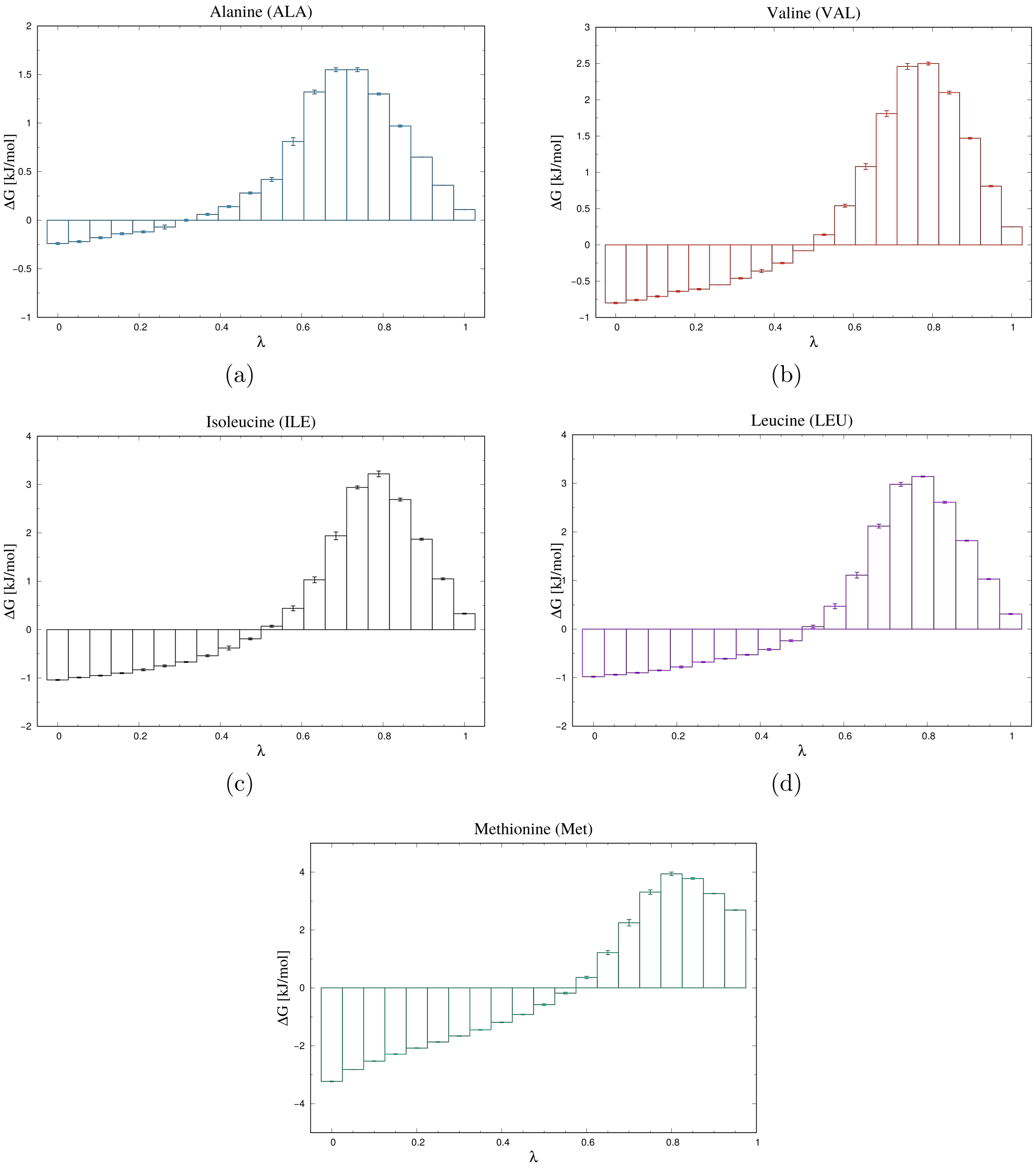
Figure 3.10: $\int_0^{\lambda}\left(\frac{dG}{d\lambda'}\right)d\lambda'$ for the Electrically charged analogs: methylimidazole (His), n-butylamine (Lys), n-propylguanidine (Arg), acetic acid (Asp), and propionc acid (Glu).

well solubilized in water. Methyl-ethylsulfide, as Wolfenden et al. report, slightly soluble in water ($\Delta G = -6.1\,\mathrm{kJ\,mol^{-1}}$). This value is different from the calculated one $\Delta G = -0.58\,\mathrm{kJ\,mol^{-1}}$, but the similar pattern could suggest that this arise from an overstatement of the tioeter group R−S−R by the force field.

Polar amino acid neutral compounds cumulative solvation free energies are reported in Fig. 3.9. Acetamide (Asn) and propionamide (Gln) have similar pattern as in the Figs. 3.12(a) and 3.12(b) These molecules are miscible in water and as for the 3-methylindol (Trp) the first λ abruptly change the interaction with water. Methanethiol (Cys) profile in Fig 3.9 is vaguely related with acetamide and propionamide one. The cumulative free energy (Fig. 3.12(c)) is characterized with two drops ($7/34\lambda$ value $\Delta\lambda = 0.206$ and $20/34\lambda$ value $\Delta\lambda = 0.588$). Even in this case, as precedently supposed, the thiol group −SH an overstate free energy of solvation.

Methanol (Ser) and Ethanol (Thr) in Figs. 3.13(a) and 3.13(b) share a peculiar drop in at one-third of the decoupling (methanol: $7/20\lambda$ value, $\lambda = 0.35$; Ethanol: $8/24\lambda$ value, $\lambda = 0.33$). These compound are characterized by an hydroxyl group −OH and the drop can be associated with this feature.

From Fig. 3.10 methylimidazole (His), acetic acid (Asp), and propionic acid (Glu) can be correlated in one trend, as n-butylamine (Lys), n-propylguanidine



(a)

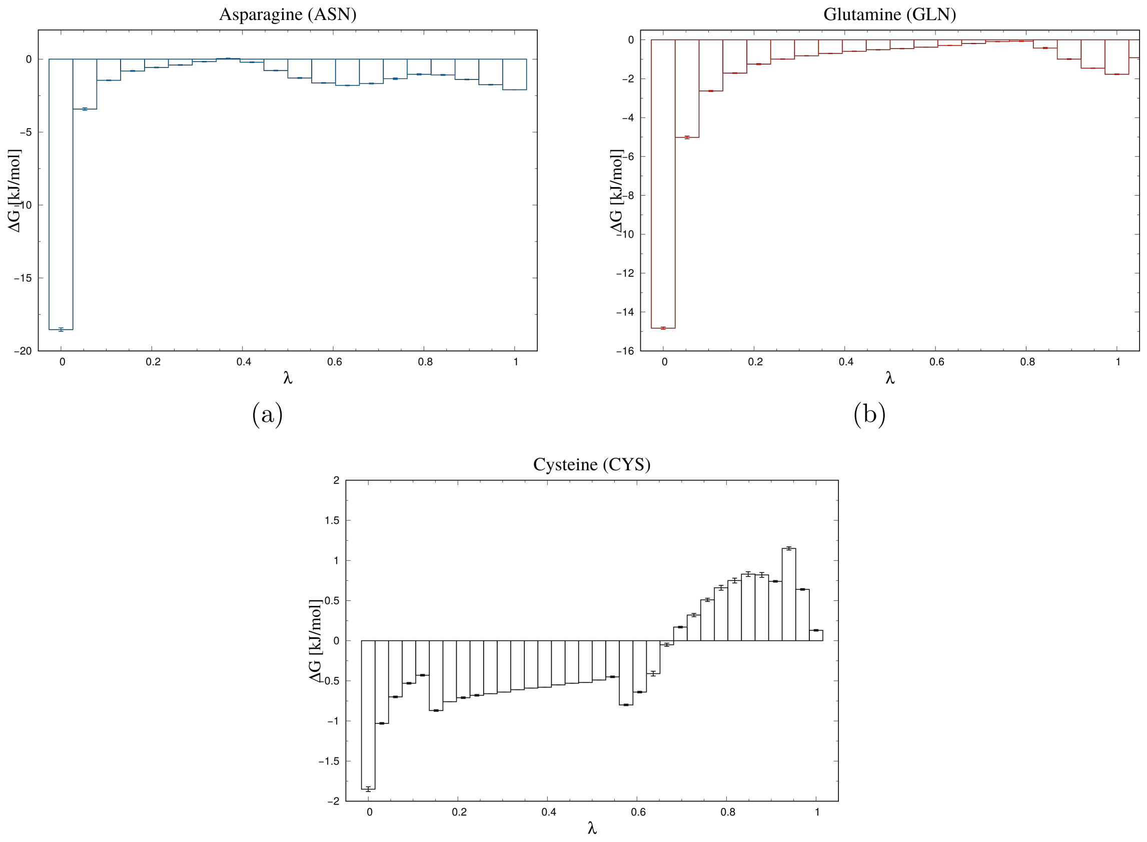
(b)

(c)

(d)

(e)

Figure 3.11: (a) $dG/d\lambda$ for Methane (Ala) with $\lambda = 20$; (b) $dG/d\lambda$ for Propane (Val) with $\lambda = 20$; (c) $dG/d\lambda$ for Butane (Ile) with $\lambda = 20$; (d) $dG/d\lambda$ for Isobutane (Leu) with $\lambda = 20$; (e) $dG/d\lambda$ for Methyl-ethylsulfide (Met) with $\lambda = 20$. This compounds have, qualitatively, the similar decoupling process.



(a)

(b)

(c)

Figure 3.12: (a) $dG/d\lambda$ for acetamide (Asn) with $\lambda = 20$; (b) $dG/d\lambda$ for propionamide (Gln) with $\lambda = 32$; (c) (b) $dG/d\lambda$ for methanethiol (Cys) with $\lambda = 3$. This compounds have, qualitatively, the similar decoupling process.

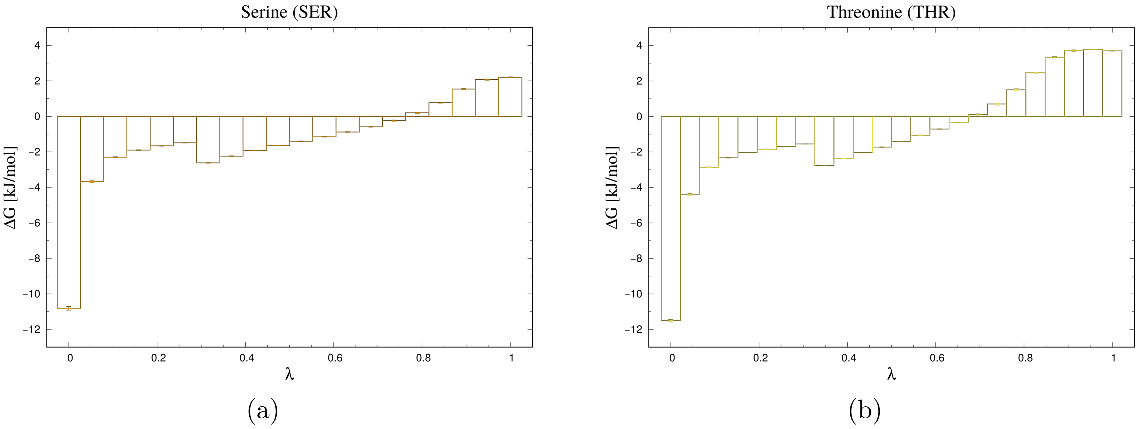
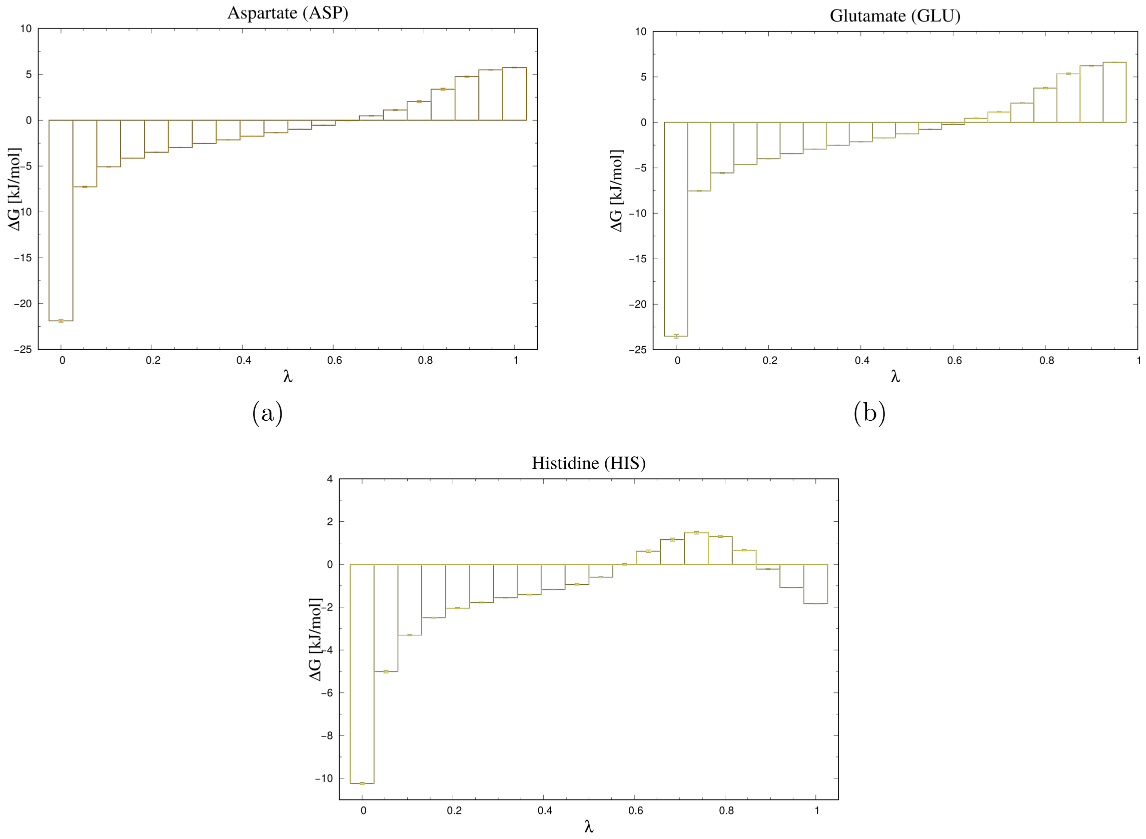


Figure 3.13: (a) $dG/d\lambda$ for methanol(Ser) with $\lambda = 23$; (b) $dG/d\lambda$ for ethanol(Thr) with $\lambda = 24$. This compounds have, qualitatively, the similar decoupling process.

(Arg), acetic acid (Asn), and propionic acid (Gln) in a second one. Acetic acid and propionic acid (Figs. 3.14(a) and 3.14(b)) patterns look more alike than the methylimidazole one (Fig. 3.14(c)) since these have a positive ΔG as function of λ in the final part, controversially, methylimidazole ΔG as function of λ is negative. Furthermore, acetic acid and propionic acid are correlated by the presence of a carboxylic group $-$COOH. The n-butylamine, n-propylguanidine, acetic acid, and propionic acid group is characterized by side chain with an amidic group in the side chain. Moreover, n-butylamine and n-propylguanidine, characterized by a primary amine group $-NH_2$ and guanidino group $-HN-C-N_2H_3$ respectively, have similar pattern (Figs. 3.15(a) and 3.15(b)) therefore, asparagine (Asn) and glutamine (Gln) due the presence of the amide group $-CONH_2$.

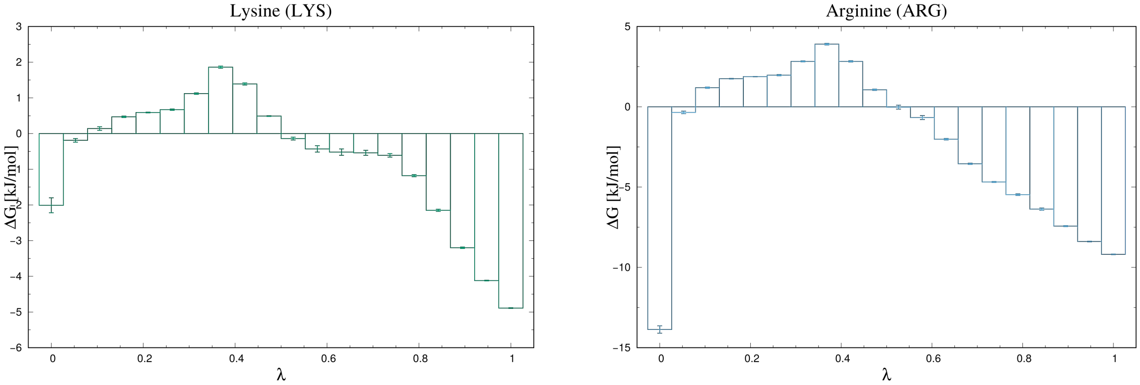


(a)

(b)

(c)

Figure 3.14: (a) $dG/d\lambda$ for acetic acid (Asp) with $\Delta\lambda = 20$; (b) $dG/d\lambda$ for propionic Acid (Glu) with $\Delta\lambda = 20$; (c) $dG/d\lambda$ for methylimidazole (His) with $\Delta\lambda = 20$ This compounds have, qualitatively, the similar decoupling process.



(a)

(b)

Figure 3.15: (a) $dG/d\lambda$ for n-butylamine (Lys) with $\lambda = 20$; (b) $dG/d\lambda$ for n-propylguanidine (Arg) with $\lambda = 20$. This compounds have, qualitatively, the similar decoupling process.

Chapter 4

Protein folding in non water solvent

In Chapter 3 we discussed how to compute the solvation free energy of a solute in water, as well as in other solvents, using thermodynamic integration. This calculation has been used in the case of single amino acid side chains. Comparison with previous, both experimental and numerical works, showed consistent results, indicating the soundness of this approach. Scaling up the system to the simulation of a full protein, however, proves a very challenging task. In this chapter we present an alternative approach, hinging upon an approximated method called *morphometric approach*, to evaluate the solvent free energy. This approach was devised by a group at the University of Kyoto whom we have been collaborating with. In protein simulation a decoys is used generally to compare a native protein with alternative geometrical structure in the way of test algorithm able to identify a native state [55]. Park and Levitt [56] used decoys conformations to investigate the ability of numerical methods to distinguish, contact and surface area, and distance-dependent energy between native configuration and decoys set. Decoys are even used to generate large set of configuration as prediction to *ab initio methods*, that require vast computational resources, and select the most fitting ones based on a scoring or free energy function [57]. The several strategies able to generate protein decoys are correlated by that structures are slightly different from the native structure generally in orded of 5 Å to 20 Å [57,69]. This feature is a direct consequence of the former application. Within the morphometric method, however, this thesis contributes to develop a bioinformatic tool capable to build several alternative decoys with a huge variety of topology and prescribed secondary structure content. Structure like the just mentioned will be used to calculate the difference in stability between the native state and the several decoys generate both in water than in other solvents. Using as framework a precedent work, the general procedure to generate decoys has been kept, we looked for significantly improve the efficiency of structures generated and structures information useful for population studies.

The protein folding process leads protein to a well defined three-dimensional structure called ”native structure” in physiological environment. The commonly accepted protein folding mechanism hinges upon the idea that all the information required for a protein to reach the native conformation are encoded in its primary sequence. The native state is consequence of the complex interplay of physical factors. Protein folding can be split in a three steps thermodynamic cycle: protein desolvatation, the relaxation of the structure and the solvation of the completely unfolded structure. Change in free energy by the folding process is defined as:

$$\Delta G^{(\text{Folding})} = \Delta H^{(\text{Folding})} - T\Delta S^{(\text{Folding})} = -\Delta G^{(\text{Unfolding})} \tag{4.1}$$

where ΔH e ΔS are the change in enthalpy and entropy in the system. The change in free energy can be calculated using thermodynamic integration as

$$\Delta G^{(\text{Unfolding})}_{(\lambda=0)} = \Delta G^{(\text{Native})}_{(0\to 1)} + \Delta G^{(\text{Unfolded})}_{(\lambda=1)} - \Delta G^{(\text{Unfolded})}_{(0\to 1)}. \tag{4.2}$$

where $(\lambda = 0)$ is the total interaction of the protein with water and $(\lambda = 1)$ describes the system completely decoupled. $\Delta G^{\text{Native}}_{(0\to 1)}$ are the $\Delta G^{\text{Unfolded}}_{(0\to 1)}$ change in free energy due of the path done from the complete coupled state to the uncoupled one. Using MD simulation as proved in Chapter 3 is able to calculate free energy of solvation for a single amino acids. Proteins solvation free energy

Protein folding can be regarded as sum of different contributions. The hydrophobic effect, decrease in protein intramolecular electrostatic interaction energy, decrease in protein intramolecular van der Waals interaction energy, the lose of conformational-entropy of protein, and the increase in protein-water electrostatic interaction energy.

The increase in protein-water electrostatic interaction energy entail a decease in water-water electrostatic interaction energy. During protein folding process the amount of solvatable surface in the protein decrease forcing water molecules to reorganize the solvent configuration. The contrasting contribute of these outlined processes sums implying a partial cancel-out of the energy increase in protein-water electrostatic interaction energy. Increase in protein-water Van der Waals interaction energy leads a change in water-water van der Waals interaction energy as the water reorganization.

The canonical concept of hydrophobic effect by the idea that a protein define an excluded volume which solvent cannot access. [58–61]. The same concept can by applied for the water molecules itself. Water occupy a volume that cannot be accessed by the other water molecules in the solvent consequently molecules in the solvent are etropically correlated create an effect defined as *water crowding* [59–61]. Folding leads to an increase in volume accessible for the solvent molecules which are more capable to move. This entropy gain, that is an immediate consequence to the reduction of water crowding, has been named as *entropic excluded-volume*.

For this studies has been choose two protein. First protein is a CREB protein (Fig. 4.1 (a)), the 4ouf [65] protein with a total structure weight of 28 050.16 Da, 238 residues, homo 2-mero stoichiometry. This protein as focal role in the transcritpional coactivation of different transcription factors, regulating the growth control and the embryonic development. This family share the histone acetyltransferase activity by which catalyze the lysine amino acids on histone proteins. Histones are a proteins that package the DNA in nucleosome. The lysine acetylation permit the dissociation of histones from DNA promoting the gene expression. The chain A of this protein as been choose to depict the all-α protein category since this 114 residues protein have 62% of helical structure (two 3_{10} helices, four α-helices, 73/114 residues participants) as reported in Fig. 4.2(a).

The second proteins (Fig. 4.1 (b)), 2pcy [66] (10 493.61 Da, 99 residues and 738 atoms), is an apoplastocynin protein, electron transport protein located in the apoplast, a continuum between two plant cell walls, that have as principal role the water and solute transport. 2pcy has been choose as the all-β reference structure because of only the 6% of α-helices (two 3_{10} helices, 6/99 residues participants) and 40% β-sheet (ten strands, 40/99 residues participants) as reported in Fig. 4.2(b).

4.1 Model

This section explain which models has been take into account to explore the physical property that permit the stability of a folded protein in a defined solvent. Two order of effects has been selected: the water crowding based on the entropic excluded-volume concept, and the H-bonds acceptor-donor interaction. In a solvated protein can occur three type hydrogen bonds: protein intramolecular (P-P), protein-water (P-W) and water-water (W-W) . The hydrophobic effect as entropic excluded-volume hinges on the physico-chemical characteristics of the solvent. With the idea of explore the hierarchy of interaction in the protein-solvent system eight models have been built up. Proteins are taken into account at atomistic level using neutral hard-sphere centered in the (x,y,z) coordinates of the atoms position with diameter set as the corresponding value of Lennard-Jones parameter, σ to substitute the atom position. Models are defined by the combination of some common features: *HS* (hard spheres)

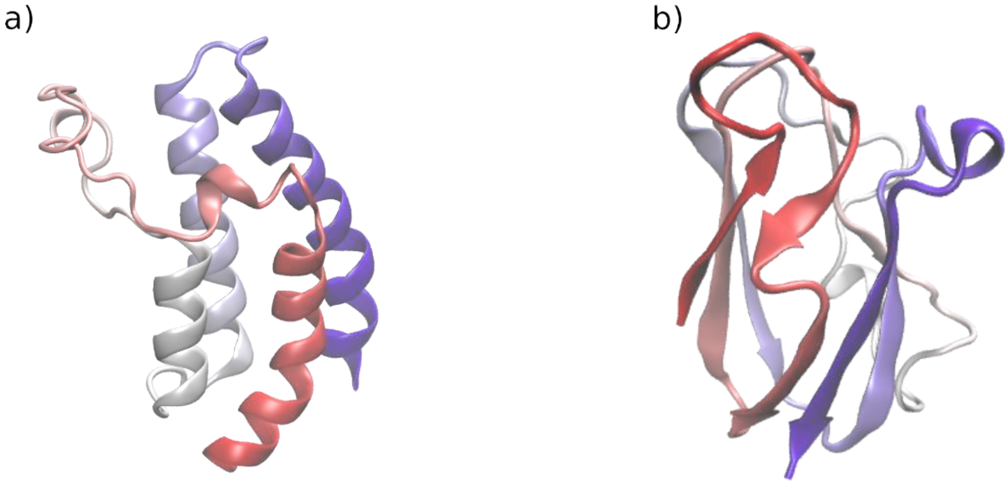


Figure 4.1: a) CREB protein 4ouf; b) apoplastocynin protein 2pcy

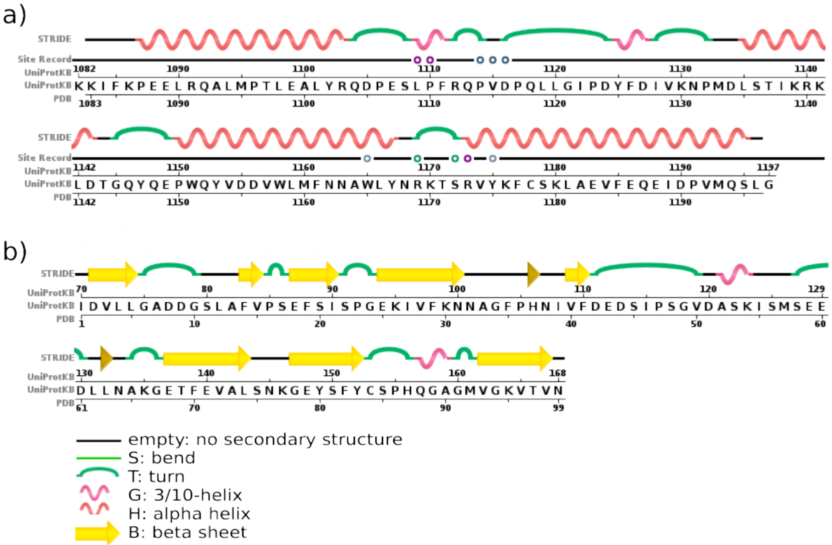

a)
STRIDE
Site Record
UniProtKB 1082 1090 1100 1110 1120 1130 1140
UniProtKB K K I F K P E E L R Q A L M P T L E A L Y R Q D P E S L P F R Q P V D P Q L L G I P D Y F D I V K N P M D L S T I K R K
PDB 1083 1090 1100 1110 1120 1130 1140
STRIDE
Site Record
UniProtKB 1142 1150 1160 1170 1180 1190 1197
UniProtKB L D T G Q Y Q E P W Q Y V D D V W L M F N N A W L Y N R K T S R V Y K F C S K L A E V F E Q E I D P V M Q S L G
PDB 1142 1150 1160 1170 1180 1190
b)
STRIDE
UniProtKB 70 80 90 100 110 120 129
UniProtKB I D V L L G A D D G S L A F V P S E F S I S P G E K I V F K N N A G F P H N I V F D E D S I P S G V D A S K I S M S E E
PDB 1 10 20 30 40 50 60
STRIDE
UniProtKB 130 140 150 160 168
UniProtKB D L L N A K G E T F E V A L S N K G E Y S F Y C S P H Q G A G M V G K V T V N
PDB 61 70 80 90 99
empty: no secondary structure
S: bend
T: turn
G: 3/10-helix
H: alpha helix
B: beta sheet

Figure 4.2: a) CREB protein 4ouf structure comprehend: two 3_{10} helices, four α-helices, 73/114 residues participants; b) apoplastocynin protein 2pcy embody two $3_{1}0$ helices with 6/99 residues participants and ten strands that counts 40/99 residues participants)

denotes non polar solvent of diameter and density set as water ($298K$, 1atm), WT stands for water, HB (hydrogen bond) means that model protein-protein or protein-water interaction. If neither of these is used HB features is neglected. SC stands that side chain is used in the model, otherwise every residues is replaced with a glycine. The eight models (Fig. 4.3) are the following:

Model HS Protein backbone is built up with hard sphere and soluted in hard-sphere solvent.

Model WT Protein backbone is built up with hard sphere and soluted in water. W-W H-bond can be formed.

Model HS-SC Protein is modeled up with hard sphere and and soluted in hard-sphere solvent. This model can be compared with HS model to understand the side chain contribution.

Model WT-SC Protein is modeled up with hard sphere and and soluted in water. W-W H-bond can be formed. This model can be compared with HS model to understand the side chain contribution.

Model HS-HB Protein backbone is built up with hard sphere and soluted hard-sphere solvent. P-P H-bonds can be formed.

Model WT-HB Protein backbone is built up with hard sphere and soluted in water. W-W, P-P and P-W H-bond can be formed.

Model HS-SC-HB Protein is modeled up with hard sphere and and soluted in hard-sphere solvent. P-P H-bonds can be formed.

Model WT-SC-HB Protein is modeled up with hard sphere and and soluted in water. W-W, P-P and P-W H-bond can be formed.

4.2 Entropic excluded-volume effect

The accomplishment of secondary structure as α-helix and β-sheet is driven by a increment in solvent entropy and a reduction in entropic excluded-volume effect. Amino acids side chain presence leads to a reduction of the total excluded-volume and solvent entropy gain.

$$\frac{F}{(k_BT)} = \frac{(\Lambda - TS)}{(k_BT)} = \left[\frac{\Lambda}{(k_BT)} - \frac{S}{k_B}\right] \tag{4.3}$$

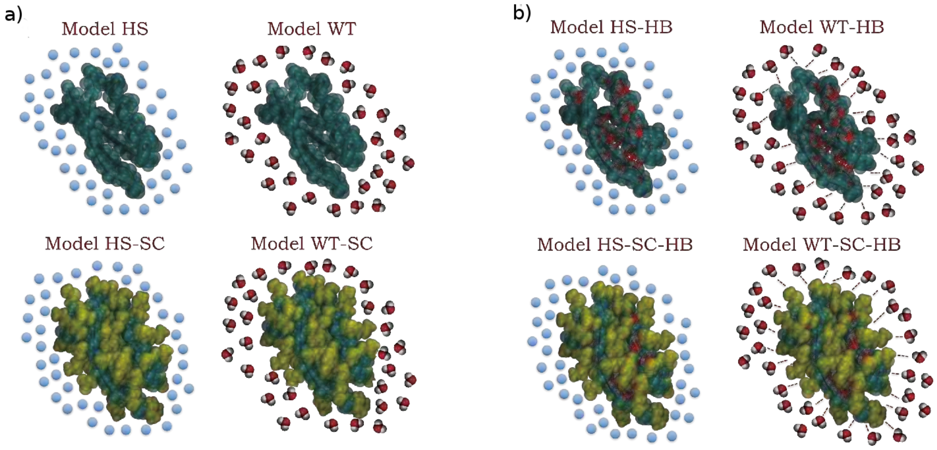


Figure 4.3: Hierarchy of models with increasing level of detail. (a) Models WT and WT-SC, only water-water (W-W) are take into account. In models HS and HS-SC, no H-bonds are incorporated. (b) In models HS-HB and HS-SC-HB, only P-P H-bonds are incorporated. In models WT-HB and WT-SC-HB, all interaction incorporated.

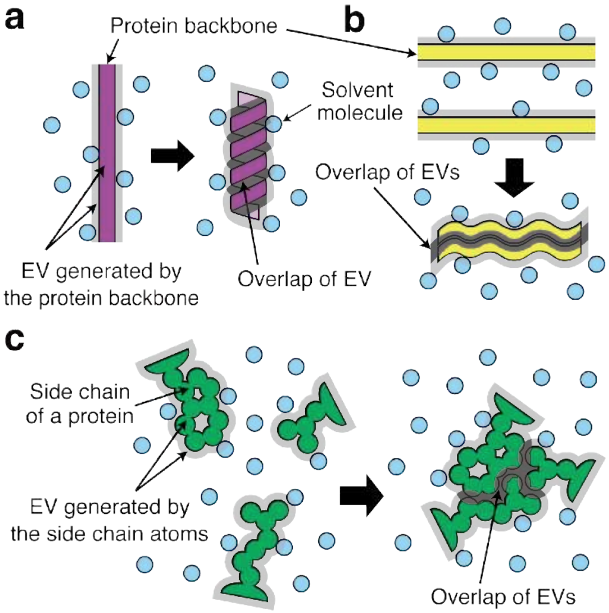


Figure 4.4: (a) α-helix by a portion of the backbone; (b) β-sheet by portions of the backbone. (c) Close packing of side chains.

where Λ is the sum of solvation intramolecular energy and S is the solvating entropy calculated under isochoric condition. When free energy function F is used to calculate the energy of well compactify structure, the protein conformational entropy does not need to be incorporated in F. Free energy function F is defined through the model employed as the the energetic and entropic components change. The stability of the real native state in different solvent is compared with the stability a number of predefined decoys. A decoy is defined as an alternative artificial native state for the protein that share the same primary structure but a different three-dimensional geometry. Stability depends upon the chosen model as well as the chosen solvent. To compare the stability between real native states and decoys has been adopted a stability parameter ΔZ is defined as:

$$\Delta Z = Z_{\text{decoy}} - Z_{\text{native}} \tag{4.4}$$

where a negative value of ΔZ means that the decoy configuration is more stable that, with a defined solvent and model, than the native configuration.

The solvation entropy S is calculated using two method the integral equation theory (IET) and the morphometric approach (MA). The morphometric method is based on differential geometry involving a class of functional called *Minkowski functionals*. Theorem states that the solvating free energy of hard core body has the form

$$F_{\text{solv}} = pV_{\text{ex}} + \sigma A + \kappa X + \bar{\kappa} Y \tag{4.5}$$

where V_{Ex}, A, X, Y are the four Minkowski functional. V_{Ex} and A represents the excluded volume and the exposed area, respectively. p and σ correspond to the intensive variables the pressure and the surface tension in the solvent. X and Y a the integrated mean surface area and the Gaussian curvature, respectively, and κ and $\bar{\kappa}$ are the corresponding intensive variables, the bending rigidities. The integrated mean over the surface area X and the Gaussian curvature of the accessible area Y are defined as

$$C = \int_{\partial V} dAH \qquad \text{where} \qquad H = \frac{1/R' + 1/R''}{2} \tag{4.6}$$

$$K = \int_{\partial V} dAK \qquad \text{where} \qquad K = \frac{1}{R'R''} \tag{4.7}$$

R' and R'' are the principal radii of curvature. The integrated mean over the surface area X needs to follow the Euler characteristics. The Euler characteristic χ can be defined for complex surface as in the case of a protein, so:

$$\chi = 4\pi N \qquad \text{where} \qquad N = 0, \pm 1, \pm 2, \ldots \tag{4.8}$$

The accessible surface area A defined by Lee and Richards as the area on the surface sampled by a sphere of solvent of radius R, in order that the solvent sphere can be reach out each point on the surface of a j-th atom without penetrate the $j - N$ nearby [67]. This surface is calculated with a so-called rolling ball algorithm where the sphere "roll" on the surface. The accessibility P_{access} is defined as:

$$P_{\text{access}} = \frac{A}{4\pi R^2} \tag{4.9}$$

In approximated way the accessible surface area A is:

$$A = \sum \left(\frac{R}{\sqrt{R^2 - Z_i}} \right) DL_i \qquad \text{where} \qquad D = \frac{\Delta Z}{2 + \Delta' Z} \tag{4.10}$$

where L_i is the length of the arc of drawn on a given section i, Z_i is the normal distance between the center of the sphere and the to the section i, ΔZ is the space between the section and $\Delta' Z$ is $\Delta Z/2$. This model cannot take in to account the cavities presence if the solvent molecules is to large to sample this convex surface.

$$P_{\text{access}} = \frac{100A}{4\pi R^2} \tag{4.11}$$

V_{ex} is the volume that is enclosed in this surface a accessible surface area A. Volume and surface a accessible surface area A are connected to:

$$A = \partial V_{\text{ex}} = \lim_{\varepsilon \to 0} \frac{V_\varepsilon - V}{\varepsilon} \tag{4.12}$$

V_ε is the volume excluded changing the radius sphere $\varepsilon \to 0$.

The determination of p, σ, κ and $\bar{\kappa}$ values is carried out using spherical solutes and the radial-symmetric integral equation (IET) approach. The solvent-solvent and the protein-solvent interactions are defined only by the solvent species as the p, σ,κ and $\bar{\kappa}$ values. In water models coefficients are different from the hard-sphere model due the difference in translational component due by the water-water H-bonds. The calculation require this four steps:

1. S is calculated for an isolate hard-sphere of solute S_{IHSS}. The diameter d_u of this sphere is $0.6 \leq d_u/d_s \leq 10$ to collect data for S using IET.

2. The p, σ, κ and $\bar{\kappa}$ are evaluated using least-square method giving the morphometric equation for a isolate sphere

$$\frac{S_{\text{IHSS}}}{k_B} = p \left(\frac{4\pi R^3}{3} \right) + \sigma(4\pi R^2) + \kappa(4\pi R) + \bar{\kappa}(4\pi R) = \frac{(d_U + d_s)}{2} \tag{4.13}$$

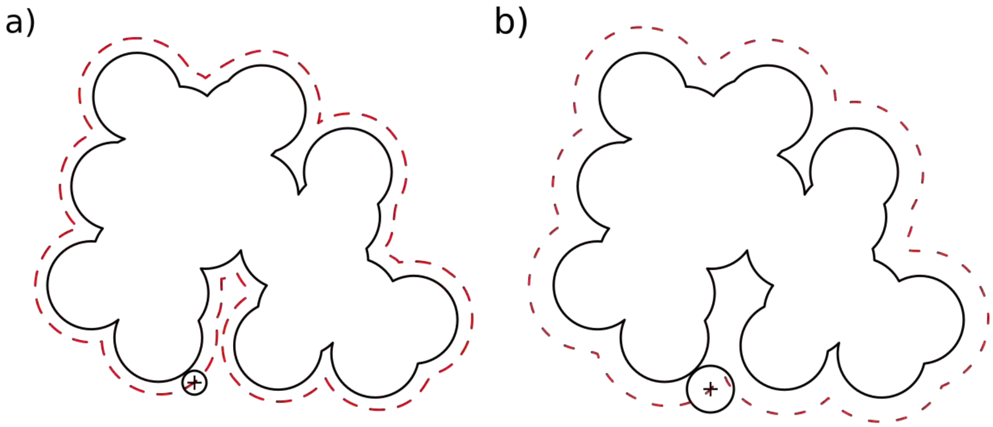


Figure 4.5: Lee-Richards molecular surface calculated with a so-called rolling ball algorithm where the sphere "roll" on the surface. Different size in radius of the sphere generate surface with different resolution.

3. Calculate V_{ex}, A, X and Y as precedently described. The (x,y,z) coordinates and the diameter D of each proteins atom are used. The atoms diameter D is set as the value of Lennard-Jones parameter σ.

4. Calculate the equation 4.5

4.3 The energetic component

Protein structure is able to form an H-bond with the secondary amine group $-\mathrm{R_2NH}$ and the carbonyl group $-\mathrm{C{=}O}$ taking part in the peptide bond. This groups can interact with solvent, if situated in the accessible surface area, or be buried inside the globular structure of the protein. Calculation of Λ can be reach setting as reference the fully extended structure, completely unfolded, in which the protein have the maximum number of P$-$W H-bonds, where $\Lambda = 0$. The folding process can give rise to the change of a P$-$W H-bond to a P$-$P H-bond, for example $\mathrm{CO} \cdots \mathrm{W} + \mathrm{NH} \cdots \mathrm{W} \rightarrow \mathrm{CO} \cdots \mathrm{NH} + \mathrm{W} \cdots \mathrm{W}$. Assuming that all the four H-bonds type have the same energy the net energy in the system when a H-bond change from P$-$W to P$-$P remains unchanged. Otherwise, when a donor or an acceptor is inside the globular structure of the protein but not P$-$P is formed the energy in the system increase of E. From the thermodynamic cycle in Fig. 4.6(a) the formation of a P$-$P H-bonds leads to an energy decrease of $-2E$.

The value of E has been calculated using quantum chemistry for the formation of

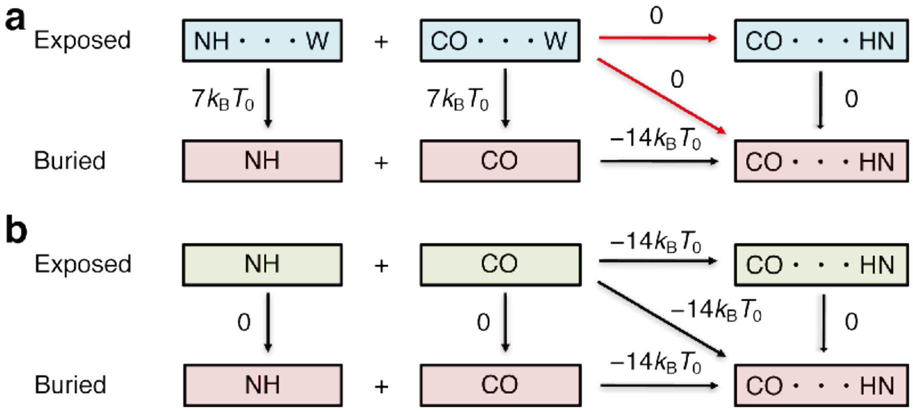


Figure 4.6: (a) Thermodynamic cycle of H-bonds water.(b) Thermodynamic cycle in the hard-sphere solvent.

an H-bond in gas phase and correspond to $-10k_BT_0$, so, in vacuuman donor-acceptor can be defined as $10k_BT_0$. The free energy decrease carried by the H-bond formation in two molecule of formamide in non-polar solvent was calculated to be $-14k_BT_0$. This interaction can be considered as the formation of a P−P H-bond in the non solvatable environment inside the globular structure of the protein. Looking at this result Kinoshita et al. sat $2E$ as $-14k_BT_0$. The water-accessible surface is calculated using the Connolly's algorithm [68]. The hard-sphere solvent is not able to form H-bonds with the protein and the use of this kind of model imply that formation of a P−P inevitably define an energy decrease of $-2E$ (Fig. 4.6(b)). Λ can be easily calculated by the number of P−P H-bonds.

4.4 Decoys preparation

To generate protein with a different three dimensional configuration has been adopted two strategies. The first one is the *3Drobot* methodology [69] which generate an arbitrary number of compact alternative structure of a defined protein without break the P−P bonds in the protein and change the overall compactness of the structure. This procedure provide a low variation(< 1Å) in RMSD in therms of the α-carbon atoms

The second one, which has been developed in this thesis, grants a wider distribution in RMSD and variety in α-helices and β-sheets. This approach generate protein decoys starting from real protein geometry to build alternative geometrical conformation of the studied proteins. The alternative configuration generated should be defined not an alternative native state, but, in contrast, geometrical native state. This definition is a direct consequence that before the free energy calculation, there are no relevant information about the solubility and the protein-protein and the protein-solvent interaction. Complex interaction pattern reviews in the cannot be transposed in this new structure since the completely change of amino acid structure.

Site directed mutagenesis (SDM) developed by Kunkel in 1985 has been wider used to understand the single site change in amino acid sequence on protein stability and solubility [70]. As example Serrano et al. states that a mutations on glycine to alanine depends on the position of the mutated amino acid in the α-helix position [71]. Glycine is preferred to the amino or the carboxylic terminal part. A single mutation can change irremediably the protein stability. The proposed strategy for the production of alternating configuration have as main purpose to probe the configuration landscape and not understand the precise role of amino acid change as in the site directed mutagenesis. This exploration use as tool the morphometric rationale.

In the aim to build a data bank of misfolded protein, each protein in the *Top500* database has been sampled [72]. Top 500 database collect proteins with a high quality in resolution, no clashcore presences, no unusual amino acids, no free-atoms refinement. These proteins structure, on account of this aspects, are used for the Ramachandran-plot distributions.

General algorithm can be divided in sub-routines (Fig. 4.7):

Sampling Protein from a database are sampled to generate fragment of N residues. In this step only the α carbon atoms are take into account ().

Compactification Fragment generated are compactificated with Metropolis Monte Carlo procedure.

Checking compactification and reconstruction From the compactification trajectory we need the to find the most compactificated protein-like structure fragment. To check this feature, the atomistic detail is rebuilt using the reference protein.

Analysis Calculate secondary structure and sort the fragments into categories discriminated by secondary structure content.

Energy minimization The chosen fragment is submitted to a quick energy minimization.

Final analysis and rendering Calculate secondary structure of the minimized structure, analyze, and render protein.

A more detailed scheme is reported in Fig. B.1

4.5 Sampling

Sampling method was performed sequentially for all database proteins (from here onwards each protein will referred to as Top500protein) to generate a number of fragments (from here onwards fragments) of N amino acid residues where N is equal to the the number of residue of the interest protein N_{protein} (4ouf, 2pcy). Every iteration a Top500protein file m is copied in a the work directory. This folder contains the Top500protein.pdb copied from Top500 proteins database.

Before sampling practice we need to remove unnecessary information from Top500 protein file and create a new Top500 protein file composed only by the coordinates of α-carbon atoms. Atoms in PDB file can be can be defined under two categories, ATOMS and HETATM (heteroatoms) [73]. ATOM keyword is reserved for atoms in standard residues of protein, instead, HETATM is applied to non-standard atoms of protein, like other kinds of groups, such as ligands, solvent, and metal ions. Only rarely in Top500 protein files, despite the high quality of the X-ray protein crystal informations, can be funded uncanonical species of amino acids, different kind of reagents that intercalate between chains or amino acids vacancy during the crystallization process. In .pdb file all this information are defined after the TER keyword. Removing information after this point, and selecting only strings with ATOMS and CA or HETATM and CA is able to prepare a good input file for the sampling algorithm. Intrachain HETATM are re-definite as ATOM in the way that compactification algorithm is able to read it. Only α-carbon coordinates file is named as Top500 protein_CA. Top500 protein_CA file are checked by the number of α-carbon atoms, if smaller than number of residues in candidate N algorithm stops and proceed to the next Top500 protein in the database. In the other hand, if is bigger, sampling algorithm select N residues from the first residue j from the amino terminal of the Top500 protein generating the first fragment k. This fragment k will be handled and analyzed. Regardless of the real accomplishment of this analysis, the next fragment $k + 1$ will be generated selecting the newest first residue adding a shifting value ΔS.

Shifting value ΔS has been choose as 20 residues because this is the typical length of α-helix structures. Sampling algorithm stops and restart from a $m + 1$ Top500 protein from the database when the number of remaining residue r in the Top500 protein in the working directory is less to the shifting value ΔS.

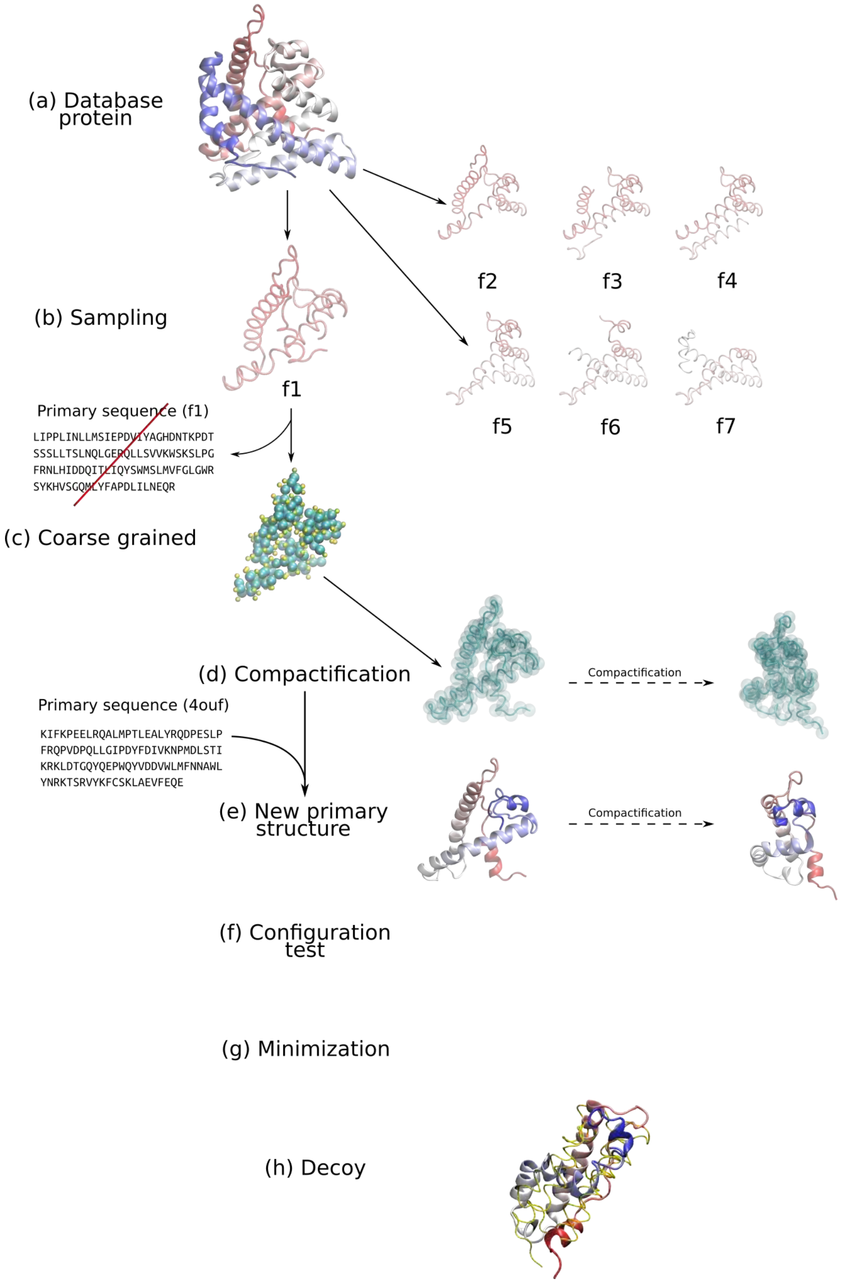
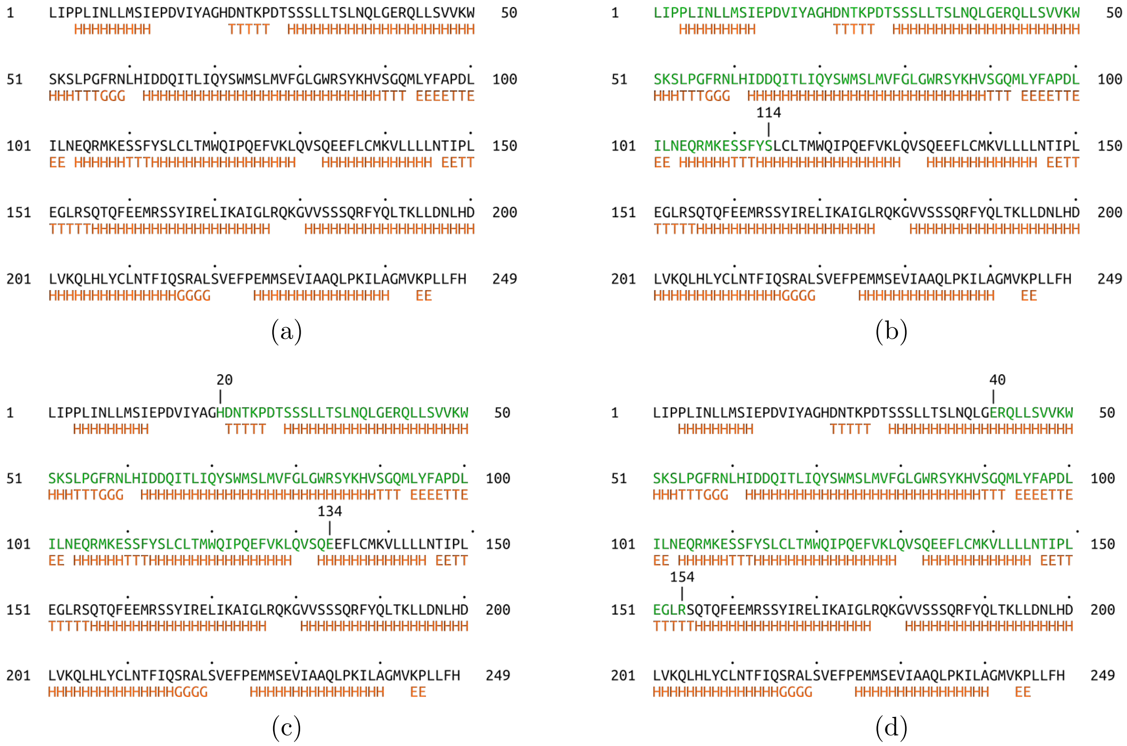

(a) Database protein
f2
f3
f4
(b) Sampling
f1
f5
f6
f7
Primary sequence (f1)
~~LIPPLINLLMSIEPDVIYAGHDNTKPDT~~
~~SSSLLTSLNQLGERQLLSVVKWSKSLPG~~
~~FRNLHIDDQITLIQYSWMSLMVFGLGWR~~
~~SYKHVSGQMLYFAPDLILNEQR~~
(c) Coarse grained
(d) Compactification
Compactification
Primary sequence (4ouf)
KIFKPEELRQALMPTLEALYRQDPESLP
FRQPVDPQLLGIPDYFDIVKNPMDLSTI
KRKLDTGQYQEPWQYVDDVWLMFNNAWL
YNRKTSRVYKFCSKLAEVFEQE
(e) New primary structure
Compactification
(f) Configuration test
(g) Minimization
(h) Decoy

Figure 4.7: Representation of the general procedure for the decoy production. (a) Protein from Top500 database (1a28BH); (b) Sampling process produce seven fragments (f1-f7); (c) Coarse grained model is generated taking into account the α-carbon atoms position. The atomistic detail of the primary sequence is neglected; (d) This coarse grained model is compactificated with a Monte Carlo procedure; (e) Primary sequence of a the studied protein, i.e. 4ouf, is imposed on all the accepted state in the Monte Carlo procedure; (f) Test routine check among the accepted states the most compactificated protein-like structure; (g) Quick energy minimization; (h) Decoy representation: 4ouf protein is pictures in yellow. Fragment generated is represented in Cartoon style.



```
1    LIPPLINLLMSIEPDVIYAGHDNTKPDTSSSLLTSLNQLGERQLLSVVKW   50
        HHHHHHHHH         TTTTT  HHHHHHHHHHHHHHHHHHHHHH

              .         .         .         .         .
51   SKSLPGFRNLHIDDQITLIQYSWMSLMVFGLGWRSYKHVSGQMLYFAPDL  100
     HHHTTTGGG  HHHHHHHHHHHHHHHHHHHHHHHHHHHHTTT EEEETTE

              .         .         .         .         .
101  ILNEQRMKESSFYSLCLTMWQIPQEFVKLQVSQEEFLCMKVLLLLNTIPL  150
     EE HHHHHTTTHHHHHHHHHHHHHHHHHH   HHHHHHHHHHHHH EETT

              .         .         .         .         .
151  EGLRSQTQFEEMRSSYIRELIKAIGLRQKGVVSSSQRFYQLTKLLDNLHD  200
     TTTTTHHHHHHHHHHHHHHHHHHHHH    HHHHHHHHHHHHHHHHHHHH

              .         .         .         .
201  LVKQLHLYCLNTFIQSRALSVEFPEMMSEVIAAQLPKILAGMVKPLLFH   249
     HHHHHHHHHHHHHHHGGGG     HHHHHHHHHHHHHHHH   EE
```

(a)

```
1    LIPPLINLLMSIEPDVIYAGHDNTKPDTSSSLLTSLNQLGERQLLSVVKW   50
        HHHHHHHHH         TTTTT  HHHHHHHHHHHHHHHHHHHHHH

              .         .         .         .         .
51   SKSLPGFRNLHIDDQITLIQYSWMSLMVFGLGWRSYKHVSGQMLYFAPDL  100
     HHHTTTGGG  HHHHHHHHHHHHHHHHHHHHHHHHHHHHTTT EEEETTE
                 114
              .   |     .         .         .         .
101  ILNEQRMKESSFYSLCLTMWQIPQEFVKLQVSQEEFLCMKVLLLLNTIPL  150
     EE HHHHHTTTHHHHHHHHHHHHHHHHHH   HHHHHHHHHHHHH EETT

              .         .         .         .         .
151  EGLRSQTQFEEMRSSYIRELIKAIGLRQKGVVSSSQRFYQLTKLLDNLHD  200
     TTTTTHHHHHHHHHHHHHHHHHHHHH    HHHHHHHHHHHHHHHHHHHH

              .         .         .         .
201  LVKQLHLYCLNTFIQSRALSVEFPEMMSEVIAAQLPKILAGMVKPLLFH   249
     HHHHHHHHHHHHHHHGGGG     HHHHHHHHHHHHHHHH   EE
```

(b)

```
                        20
                         |
1    LIPPLINLLMSIEPDVIYAGHDNTKPDTSSSLLTSLNQLGERQLLSVVKW   50
        HHHHHHHHH         TTTTT  HHHHHHHHHHHHHHHHHHHHHH

              .         .         .         .         .
51   SKSLPGFRNLHIDDQITLIQYSWMSLMVFGLGWRSYKHVSGQMLYFAPDL  100
     HHHTTTGGG  HHHHHHHHHHHHHHHHHHHHHHHHHHHHTTT EEEETTE
                                     134
              .         .         .   |     .         .
101  ILNEQRMKESSFYSLCLTMWQIPQEFVKLQVSQEEFLCMKVLLLLNTIPL  150
     EE HHHHHTTTHHHHHHHHHHHHHHHHHH   HHHHHHHHHHHHH EETT

              .         .         .         .         .
151  EGLRSQTQFEEMRSSYIRELIKAIGLRQKGVVSSSQRFYQLTKLLDNLHD  200
     TTTTTHHHHHHHHHHHHHHHHHHHHH    HHHHHHHHHHHHHHHHHHHH

              .         .         .         .
201  LVKQLHLYCLNTFIQSRALSVEFPEMMSEVIAAQLPKILAGMVKPLLFH   249
     HHHHHHHHHHHHHHHGGGG     HHHHHHHHHHHHHHHH   EE
```

(c)

```
                                            40
                                             |
1    LIPPLINLLMSIEPDVIYAGHDNTKPDTSSSLLTSLNQLGERQLLSVVKW   50
        HHHHHHHHH         TTTTT  HHHHHHHHHHHHHHHHHHHHHH

              .         .         .         .         .
51   SKSLPGFRNLHIDDQITLIQYSWMSLMVFGLGWRSYKHVSGQMLYFAPDL  100
     HHHTTTGGG  HHHHHHHHHHHHHHHHHHHHHHHHHHHHTTT EEEETTE

              .         .         .         .         .
101  ILNEQRMKESSFYSLCLTMWQIPQEFVKLQVSQEEFLCMKVLLLLNTIPL  150
     EE HHHHHTTTHHHHHHHHHHHHHHHHHH   HHHHHHHHHHHHH EETT
       154
        |     .         .         .         .         .
151  EGLRSQTQFEEMRSSYIRELIKAIGLRQKGVVSSSQRFYQLTKLLDNLHD  200
     TTTTTHHHHHHHHHHHHHHHHHHHHH    HHHHHHHHHHHHHHHHHHHH

              .         .         .         .
201  LVKQLHLYCLNTFIQSRALSVEFPEMMSEVIAAQLPKILAGMVKPLLFH   249
     HHHHHHHHHHHHHHHGGGG     HHHHHHHHHHHHHHHH   EE
```

(d)

Figure 4.8: Protein sampling routine. In black primary structure in one letter code: alpha helix H, extended conformation G, isolated bridge B or b, turn T and coil C. In green the fragment selected. Numbers correspond to number of residues. (a) Protein taken from Top500 database (1a28BH); (b),(c),(d) First, second, and third fragment sampling, respectively.

4.6 Compactification

The next step is prepare all the preliminary file needed for the compactification. Monte Carlo algorithm for the compactification use as input three files: Top500 configuration, primary sequence Top500, primary sequence studied protein. File Top500 conf is merely a copy of fragment k, primary sequence Top500 is the primary sequence of the fragment k, and primary sequence studied protein is the primary sequence of the interest protein which is stored in the script folder.

The model for the Metropolis Monte Carlo [74] is constituted as a sequences of N beads with position defined by the α-carbon position (x,y,z) in pdb file, with diameter σ. For each position is associated a vector $(\mathbf{r_1}, \dots, \mathbf{r_N})$. The consecutive bonds, as n and $n+1$, are connected by a fixed bond length b, instead the non-consecutive interact with a square-well potential.

$$V^{(\mathrm{SW})} = \begin{cases} +\infty & \text{for} \quad r_{ij} < \sigma \\ -\varepsilon & \text{for} \quad \sigma < r_{ij} < R_c \\ 0 & \text{for} \quad r_1 > \lambda\sigma \end{cases} \tag{4.14}$$

where $r_{ij} = |\mathbf{r_{ij}}| = |\mathbf{r_j} - \mathbf{r_i}|$, the well width $\lambda - 1$ in σ units. This value define the range of interaction $R_c = \lambda\sigma$. The ε defines the well depth. Side chain plays a relevant role in protein folding. In this model we take in to account the presence of the side chain with N-2 hard sphere of diameter σ_s. First and last side chain are excluded in the model. For define the position of the side chains sphere, one define the tangent $\mathbf{T}_i$ and a normal vector $\mathbf{N}_i$ as

$$\mathbf{T}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_{i-1}}{|\mathbf{r}_{i+1} - \mathbf{r}_{i-1}|} \tag{4.15}$$

$$\mathbf{N}_i = \frac{\mathbf{r}_{i+1} - 2\mathbf{r}_i + \mathbf{r}_{i-1}}{|\mathbf{r}_{i+1} - 2\mathbf{r}_i + \mathbf{r}_{i-1}|} \tag{4.16}$$

and the binormal vector $\mathbf{B}_i$

$$\mathbf{B}_i = \mathbf{T}_i \times \mathbf{N}_i \tag{4.17}$$

The side chain sphere is positioned in anti-normal direction with position

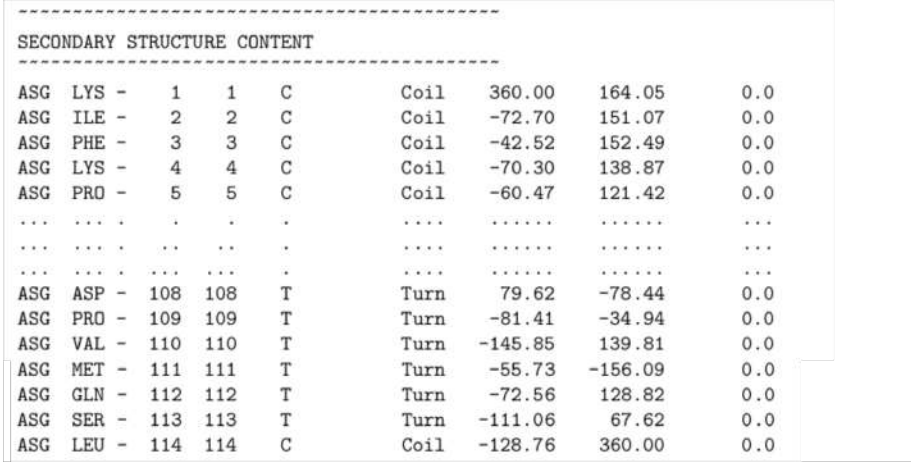
$$\mathbf{r}_i^{(\mathrm{sc})} = \mathbf{r}_i - \mathbf{N}_i \left[\frac{(\sigma + \sigma_s)}{2} \right] \tag{4.18}$$

4.7 Checking compactification and reconstruction

Compactification tool should produce a trajectory file where are reported all the accepted configurations during the Monte Carlo procedure. If compactification had some compromising unsuspected error will not return this file and just in this case algorithm stops and restart with another fragment $m+1$. During the reparation process of the structure, due to undefined coordinate, some frame in the compactification trajectory can report the NaN (Not a Number) information instead of coordinates. In this case this frame will be not take in account. Trajectory file has been divided in several files, one for each frame, numbered with pad number. Reconstruction of atomistic content has been done with PULCHRA [75], with options: -q optimizes backbone hydrogen bonds pattern, usually gaining a slightly better RMSD, but a little bit more time consuming (default: off),-r starts optimization from a random alpha carbon chain rather than from initial coordinates (default: off). Before that test that define which configuration is considerable as good candidate, first frame of the trajectory file has been reconstructed to analyses original fragment from protein. We choose to rebuild the atomistic structure of the fragment instead of using the original one to compare structure re generated with the same tool, instead of use an the X-ray data from pdb and the PULCHRA generated. First frame in fragment was reconstructed with the primary sequence Top500, and analyzed with STRIDE [76]. Stride (STRuctural IDEntification) is an algorithm for the assignment of protein secondary structure elements given the atomic coordinates of the protein. This tool is implemented inside VMD [77] to elaborate the New Cartoon visualization but can be used as standalone. All the information about these steps have been saved a file log as listed below. FASTA format is a text-based format representing peptide sequences using single-letter codes. Information about the detailed secondary structure assignment are reported in as in a log file (Fig. 4.9 - 4.9).

The core part of the algorithm is how to choose and which condition operate to select a good candidate in a list of $\approx 10^3$ possible configurations written in the trajectory of compactification. Test protocol is supported by the VMD analysis that print a short description of the structures visualized. Two informations were used to define if a structure can be able to pass the test (optimal visualization by VMD) or not.

Residue numbers Number of residue is not read from the input file but is calculated by the relative atom position in PDB file. PDB do not have any informations about bonds between atoms. VMD compute bonds location using typical length between two atom, consequently, does not provide en excellent calculation of coordination number of a specific element. When VMD is forced to guest the connectivity, it considers a bond to be formed whenever two atoms are within $R_1 \times R_2 \times 0.6$ of each other, where R_1 and R_2 are the respective radii of candidate atoms.



```
--------------------------------------------
SECONDARY STRUCTURE CONTENT
--------------------------------------------
ASG  LYS -    1    1    C         Coil    360.00    164.05       0.0
ASG  ILE -    2    2    C         Coil    -72.70    151.07       0.0
ASG  PHE -    3    3    C         Coil    -42.52    152.49       0.0
ASG  LYS -    4    4    C         Coil    -70.30    138.87       0.0
ASG  PRO -    5    5    C         Coil    -60.47    121.42       0.0
...  ... .    .    .    .         ....    ......    ......       ...
...  ... .   ..   ..    .         ....    ......    ......       ...
...  ... .  ...  ...    .         ....    ......    ......       ...
ASG  ASP -  108  108    T         Turn     79.62    -78.44       0.0
ASG  PRO -  109  109    T         Turn    -81.41    -34.94       0.0
ASG  VAL -  110  110    T         Turn   -145.85    139.81       0.0
ASG  MET -  111  111    T         Turn    -55.73   -156.09       0.0
ASG  GLN -  112  112    T         Turn    -72.56    128.82       0.0
ASG  SER -  113  113    T         Turn   -111.06     67.62       0.0
ASG  LEU -  114  114    C         Coil   -128.76    360.00       0.0
```

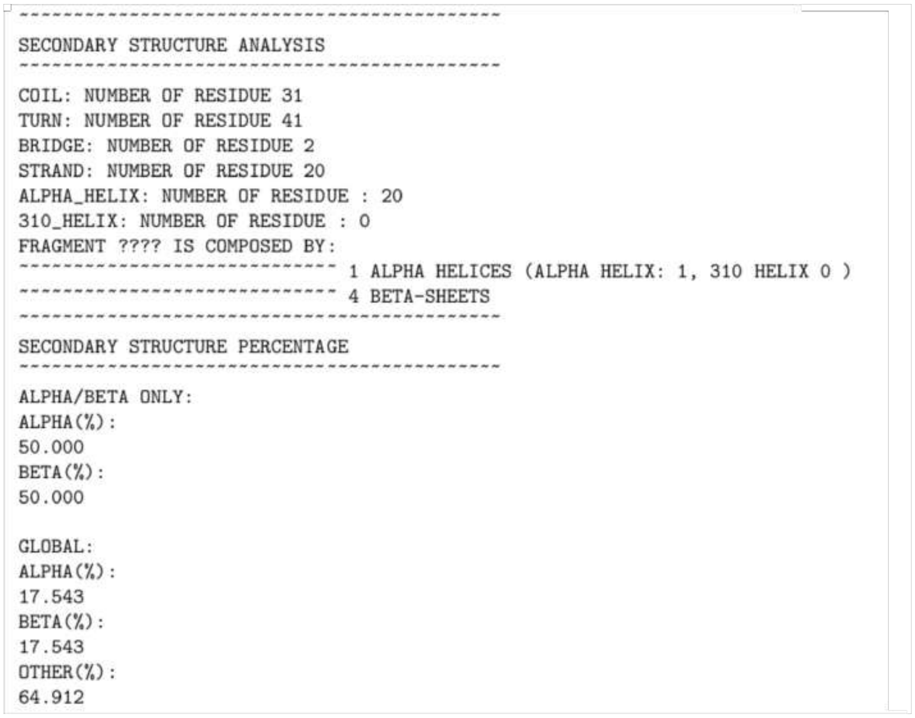
Figure 4.9: The secondary structure assignment from left to right: ASG (Detailed secondary structure assignment) identifier, residue name, protein chain identifier (none in this case), PDB residue number, ordinal residue number, one letter secondary structure code, full secondary structure name, ϕ angle, ψ angle, and the residue solvent accessible area.

Unusual bonds It is quite possible for a protein to be connected to a nucleic acid or some other non-protein. When this occurs, a warning message is printed. These warnings are known to occur with terminal amino acids, zinc fingers, myristolated residues, and poorly defined structures.

Test script reads this lines and determinate with a two steps control if a structure is protein-like (P) or not (G). This code were defined arbitrarily. The algorithm double check the structure, firstly, verifying if the number of residue fits with number of $\alpha-$carbon atoms in the studied protein. If true, structure is labeled as P only if even the second check pass and so no unusual bonds warnings are present. In other case, if number of residues is different from number of $\alpha-$carbon atoms in the studied protein or unusual bond warnings presence, fragment will be defined as G.

The P configuration search has been led by that significant structures have to be the most compactificated and with no test errors.

Looking over gyration radius change in the trajectory of the compactification is explicit that the last frames are smaller than the starting part of the trajectory, so a good candidate is located in the last portion of trajectories. We adopt the *binary search*, also known as *half-interval search* like strategy to find a candidate. This search algorithm works on the principle of divide and check and is generally used to find a particular number in a set. Binary search looks for a particular item by



```
--------------------------------------------
SECONDARY STRUCTURE ANALYSIS
--------------------------------------------
COIL: NUMBER OF RESIDUE 31
TURN: NUMBER OF RESIDUE 41
BRIDGE: NUMBER OF RESIDUE 2
STRAND: NUMBER OF RESIDUE 20
ALPHA_HELIX: NUMBER OF RESIDUE : 20
310_HELIX: NUMBER OF RESIDUE : 0
FRAGMENT ???? IS COMPOSED BY:
----------------------------- 1 ALPHA HELICES (ALPHA HELIX: 1, 310 HELIX 0 )
----------------------------- 4 BETA-SHEETS
--------------------------------------------
SECONDARY STRUCTURE PERCENTAGE
--------------------------------------------
ALPHA/BETA ONLY:
ALPHA(%):
50.000
BETA(%):
50.000

GLOBAL:
ALPHA(%):
17.543
BETA(%):
17.543
OTHER(%):
64.912
```

Figure 4.10: Secondary structure analysis checks how many residues take part in each secondary structure and which and how many structures are formed. After that secondary composition has been calculated in two ways, taking in to account alpha and beta only and using alpha, beta and other, composed by the sum of COIL, TURN and BRIDGE.

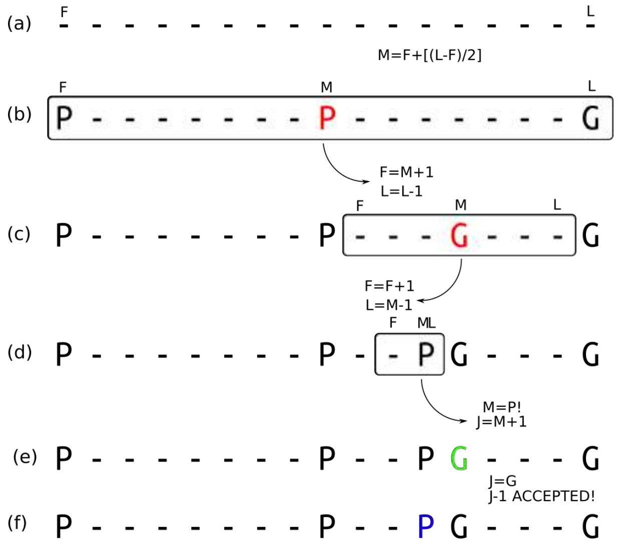

(a)
F
L
M=F+[(L-F)/2]
(b)
F
M
L
P - - - - - - - - P - - - - - - - G
F=M+1
L=L-1
(c)
F
M
L
P - - - - - - - - P - - - G - - - G
F=F+1
L=M-1
(d)
F
ML
P - - - - - - - - P - - P G - - - G
M=P!
J=M+1
(e)
P - - - - - - - - P - - P G - - - G
J=G
J-1 ACCEPTED!
(f)
P - - - - - - - - P - - P G - - - G

Figure 4.11: Schematics test example. (a) Unknown trajectory configuration. First frame F last frame L are defined by the number of total configuration; (b) Preparative step. Calculate M, first test define F=P, M=P, and L=G. Mid frame is a protein like configuration. Algorithm checks forward. F and L are recalculated with the following rules F=M+1 and L=L-1; (c) Starting the real procedure. We define only mid frame. Test define M=G. Algorithm checks backward. F and L are recalculated with the following rules F=F+1 and L=M-1; (d) Test define M=P. Starting jump routine to verify if this P configuration is the larger local P; (e) Jump one step forward until a G state is found; (f) Jump routine J=G. J-1 configuration is straightforwardly the larger local P.

comparing the middle item of the collection, and, if a match occurs, then the position of the item is returned. If the middle item is greater than the item, then the item is searched in the sub-array to the left of the middle item. Otherwise, the item is searched for in the subarray to the right of the middle item. This process continues on the sub-array as well until the size of the subarray reduces to zero. this algorithm to work properly, the data collection should be in the *sorted* form. Precisely for this reason we used the idea binary search with the adaptation of case. In trajectory file, obviously, we have not information about the test state before the test run on a defined structure, consequently the sorted form is not present. Rather, binary search logic has been employ to sample trajectory of compactification in the aim to understand which test state is a defined frame and, depending on the test output, algorithm changes the half-interval position.

Algorithm to find a possible candidate works in this way:

1. Frame number in the trajectory define the *last frame number* variable $N^{(\mathrm{Last})}$, meanwhile, *first frame number* variable $N^{(\mathrm{First})}$ has been set as one.

2. *Mid frame number* $N^{(\mathrm{Mid})}$ has been defined as:

$$N^{(\mathrm{Mid})} = N^{(\mathrm{First})} + \left(\frac{N^{(\mathrm{Last})} - N^{(\mathrm{First})}}{2} \right) \tag{4.19}$$

3. Run a first test in the aim to globally check the trajectory the status of first frame number variable $N^{(\mathrm{First})}$, mid frame number $N^{(\mathrm{Mid})}$, last frame number $N^{(\mathrm{Last})}$ the result, as P or G, is written in these variables ,$N_{\mathrm{FT}}^{(\mathrm{First})}$, $N_{\mathrm{FT}}^{(\mathrm{Last})}$ and $N_{\mathrm{FT}}^{(\mathrm{Mid})}$ where *FT* states for first test. Each test results is printed in the log file and define a different path in the searching algorithm:

Last frame = P Last frame is P ($N_{\mathrm{FT}}^{(\mathrm{First})} = \mathrm{P}$) configuration means that the most compactificated structure can be read without problem from VMD.

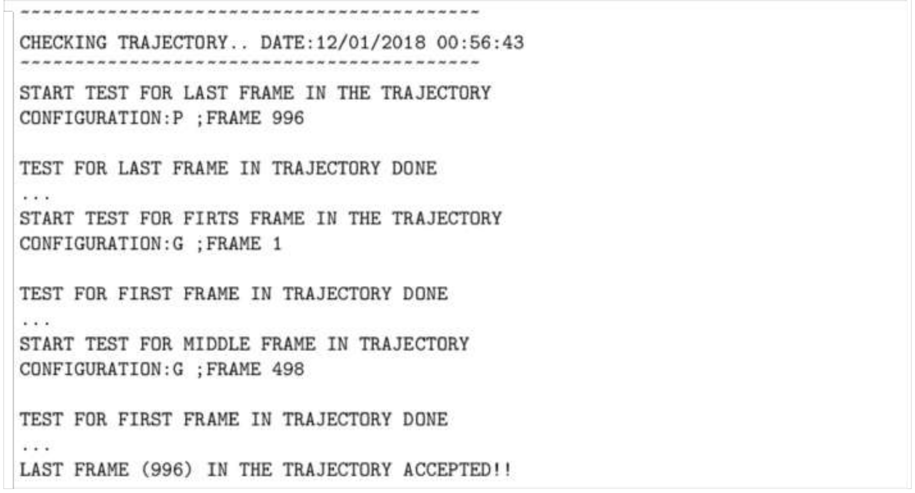
Last frame = G Last frame in G ($N_{\mathrm{FT}}^{(\mathrm{First})} = \mathrm{G}$)configuration means that something wrong happen. Check mid frame number $N^{(\mathrm{Mid})}$.

Mid frame = G when the mid frame is G ($N_{\mathrm{FT}}^{(\mathrm{Mid})} = \mathrm{G}$) algorithm start to search in the first-half (from $N^{(\mathrm{First})}$ to $N^{(\mathrm{Mid})}$) , after a conclusive output, in the second one of the trajectory (from $N^{(\mathrm{Mid})}$ to $N^{(\mathrm{Last})}$). Only the biggest value is accepted. When ($N^{(\mathrm{Mid})} = \mathrm{G}$), this frame is re-analyzed by test function (results is merely the same G). First frame number $N^{(\mathrm{First})}$ and last frame number $N^{(\mathrm{Last})}$ are redefined and,consequently, mid frame number $N^{(\mathrm{Last})}$

$$N^{(\mathrm{First})} = N^{(\mathrm{First})} + 1 \tag{4.20}$$

$$N^{(\mathrm{Last})} = N^{(\mathrm{Mid})} - 1 \tag{4.21}$$

This test and variables change is looped until test result is P. This operation search in the first part of compactification trajectories where the r_g is certainly too large respect our goal. This logic is forced by the ($N_{\mathrm{FT}}^{(\mathrm{Mid})} = \mathrm{G}$) results and, more likely, the second half of the trajectory is not usable. Undoubtedly, this P configuration, does not correspond to the local larger r_g. To score the local P with larger r_g easily we check frame by frame moving forward until we find a G. The step before is the larger local P configuration for the first half. The second half the trajectory is



```
------------------------------------------
CHECKING TRAJECTORY.. DATE:12/01/2018 00:56:43
------------------------------------------
START TEST FOR LAST FRAME IN THE TRAJECTORY
CONFIGURATION:P ;FRAME 996

TEST FOR LAST FRAME IN TRAJECTORY DONE
...
START TEST FOR FIRTS FRAME IN THE TRAJECTORY
CONFIGURATION:G ;FRAME 1

TEST FOR FIRST FRAME IN TRAJECTORY DONE
...
START TEST FOR MIDDLE FRAME IN TRAJECTORY
CONFIGURATION:G ;FRAME 498

TEST FOR FIRST FRAME IN TRAJECTORY DONE
...
LAST FRAME (996) IN THE TRAJECTORY ACCEPTED!!
```

Figure 4.12: Log file showing TEST steps.

explored in the same way taking in to account some modification. First frame number $N^{(\text{First})}$ and last frame number $N^{(\text{Last})}$ are redefined as:

$$N^{(\text{First})} = N^{(\text{Mid})} + 1 \tag{4.22}$$

$$N^{(\text{Last})} = N^{(\text{Last})} - 1 \tag{4.23}$$

This test function and variables change is looped until test result is G. Local P with larger r_g is reached frame by frame moving backward until we find a P. This frame is the larger local P configuration for the second half. Between and second half configuration algorithm choose the most compactificated.

Mid frame = P when the mid frame is P algorithm start to search only in the second-half.

In Fig. 4.11 is schematized is visualized an test example.

4.8 Stride analysis

After that algorithm finds a candidate, we proceed to categorize fragments in one of five categories using STRIDE:

Type A All-α: protein with 85% of α-helices

Type B $70\%\alpha-30\%\beta$: protein with a percentage between 85% and 65% of α-helices

Type C $50\%\alpha-50\%\beta$ protein with a percentage between 65% and 45% of α-helices

Type D $30\%\alpha-70\%\beta$ protein with a percentage between 45% and 15% of α-helices

Type E All-β protein with less than 15% of α helices

Files are moved inside a specific percentage folder due the STRIDE analysis, the file data are collected in a new work folder where name highlight Top500protein of derivation, the fragment number and is the chosen frame number.

Before a detailed quantitative analysis algorithm attempt to minimize the energy ensuring that the system has no steric clashes or inappropriate geometry and improves the H-bonds pattern in decoys. The simulations were performed in periodic cubic box with a distance between the solute and the box of 1 nm (approximately 6 nm and 7500 water molecules). The integrator algorithm for energy minimization was set as steep descent. The minimization is converged when the max force is smaller than 100 kJ mol^{1}nm^1, maximum step size in nm 0.01 nm and the maximum number of steps was set as 50000. Frequencies to update the neighbor list (and the long-range forces, when using twin-range cut-off) were updated every step. Bonds and angles were not constrained.

Minimized structures were tested to check if the configuration contains unusual bonds or number of residues were different as arranged beforehand. In negative case algorithm analyze the structure even before the minimization. In the word directory can be written three possible folder, in function of the minimization routine. If the minimization fails or test has been settled as negative (G), non minimized fragment is goes to the rendering and analysis procedure.

Analysis routine calculate and compare: the native protein gyration radius r_g^{Native}, fragment gyration r_g^{Fragment}, RMSD between native protein and fragment, and the variation between gyration radii of native protein and fragment Δ. There are two atom selections needed to do an RMSD computation in VMD, the list of atoms to compare in both molecules. RMSD is define as:

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{4.24}$$

where δ_i is the distance between atom i and either a reference structure or the mean position of the N equivalent atoms. The first atom of the first selection is compared to the first atom of the second selection. To fit structures α-carbon atoms were

TYPE	$\langle r_g \rangle$ (Å)	r_g^{Min}-r_g^{Max} (Å)	$\langle$RMSD$\rangle$ (Å)	RMSD$^{\mathrm{Min}}$-RMSD$^{\mathrm{Max}}$ (Å)	$\langle \Delta r_g \rangle$ (%)	$\Delta r_g^{\mathrm{Min}}$-$\Delta r_g^{\mathrm{Max}}$ (%)
A	14.19	11.04-21.82	61.45	51.16-80.00	−1.10	−23.07-52.08
B	13.98	11.56-19.40	60.86	52.07-79.39	−2.54	−19.39-35.20
C	14.42	11.77-21.61	62.02	52.14-77.30	0.51	−17.97-50.62
D	14.18	12.18-19.36	61.45	54.59-73.62	−1.14	−15.09-34.95
E	13.94	11.89-22.88	61.44	54.18-92.47	−2.81	−17.11-59.47

Table 4.1: 4ouf protein statistics for types A-E.average gyration radius $\langle r_g \rangle$, minimum-maximum gyration radius (r_g^{Min} and r_g^{Max}), averages RMSD, minimum-maximum RMSD (RMSD$^{\mathrm{Min}}$ and RMSD$^{\mathrm{Max}}$), average gyration radius variation $\langle \Delta r_g \rangle$, minimum-maximum gyration radius variation (Δr_gMin and $\Delta r_g^{\mathrm{Max}}$).

choose as comparative coordinate. The actual order is identical to the order from the input PDB file. Returns the radius of gyration of atoms in selection using the given weight. The radius of gyration is computed as

$$r_g^2 = \frac{(\sum_{i=1}^{n} w_i (r_i - \bar{r})^2)}{(\sum_{i=1}^{n} w_i)} \tag{4.25}$$

where r_i is the position of the ith atom and $\bar{r}$ is the weighted center

4.9 Results

All decoys generated for 4ouf and 2pcy have been explored in meaning of gyration radius, RMSD, alpha-beta secondary structure percentage,and gyration radius variation respect of the studied protein ($r_g^{4\mathrm{ouf}} = 14.352\,$Å and $r_g^{2\mathrm{pcy}} = 12.342\,$Å).

For the 4ouf protein we obtained 1048 protein distributed as follows: for the type A 564 structures, type B 279 structures, type C 84 structures, type D 61 structures,and type E 60 structures. This proteins have been explored in meaning as

Algorithm generate 1294 fragments of which 717 for the type A protein decoys, 276 for type B protein decoys , 105 for protein decoys type C,103 for protein decoys type D, and 93 type for protein decoys E.

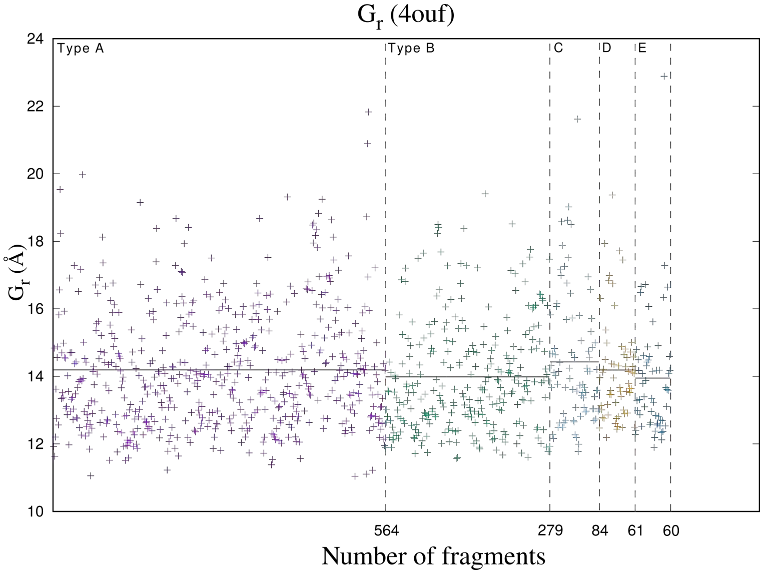


Figure 4.13: Gyration radius value for type A-E for 4ouf decoys. In solid black line the average gyration radius for each type.

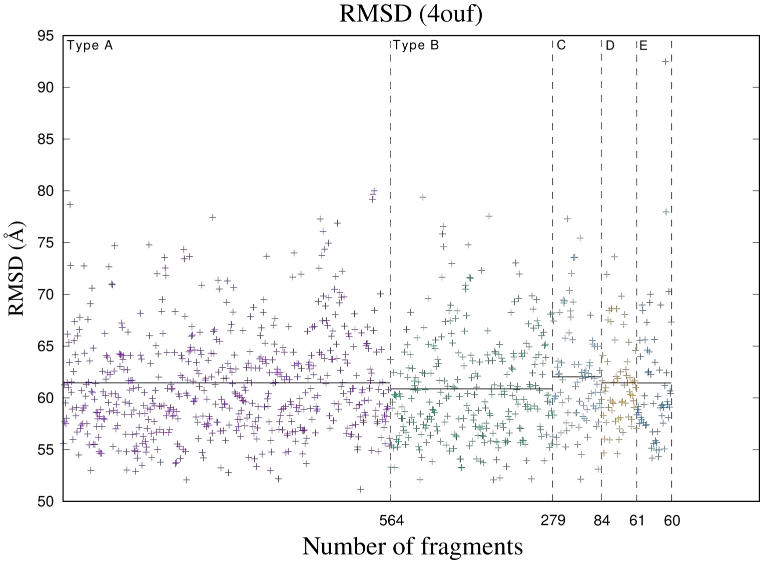


Figure 4.14: RMSD radius value for type A-E for 4ouf decoys. In solid black line the average gyration radius for each type.

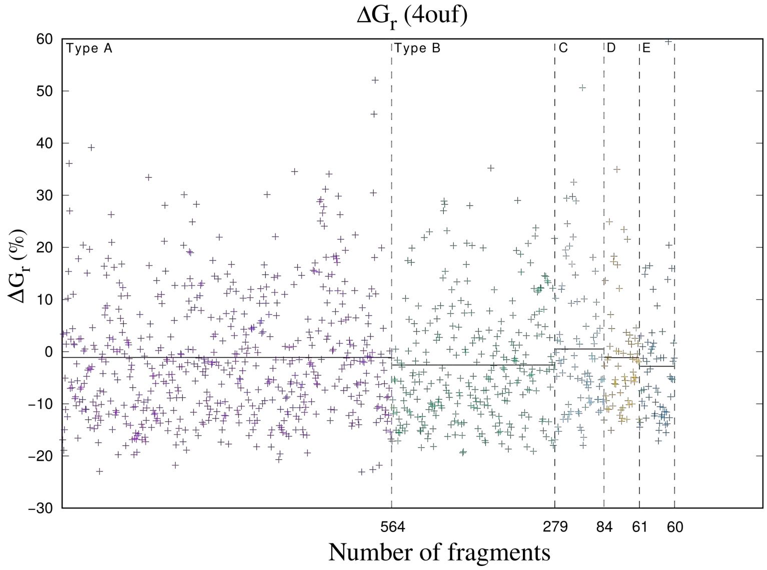


Figure 4.15: Variation in gyration radius value for type A-E for 4ouf decoys. In solid black line the average gyration radius for each type.

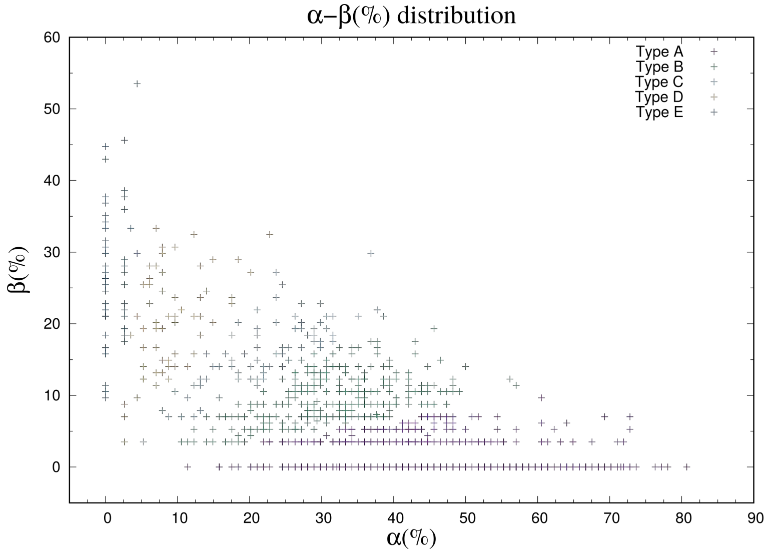


Figure 4.16: α-β percentage for 4ouf decoys generated.

TYPE	$\langle r_g \rangle$ (Å)	r_g^{Min}-r_g^{Max} (Å)	$\langle$RMSD$\rangle$ (Å)	RMSD$^{\mathrm{Min}}$-RMSD$^{\mathrm{Max}}$ (Å)	$\langle \Delta r_g \rangle$ (%)	$\Delta r_g^{\mathrm{Min}}$-$\Delta r_g^{\mathrm{Max}}$ (%)
A	14.19	10.16-20.83	38.69	35.88-42.64	8.12	−17.69-68.78
B	13.30	10.64-18.30	38.66	36.41-42.61	7.7	−13.78-48.28
C	14.03	10.91-22.59	38.93	36.19-42.10	13.69	−11.57-83.01
D	13.69	10.89-20.92	38.87	36.57-43.58	10.94	−11.74-69.56
E	13.12	11.30-20.27	38.68	36.89-40.48	6.29	−8.40-64.28

Table 4.2: 2pcy protein statistics for types A-E.average gyration radius $\langle r_g \rangle$, minimum-maximum gyration radius (r_g^{Min} and r_g^{Max}), averages RMSD, minimum-maximum RMSD (RMSD$^{\mathrm{Min}}$ and RMSD$^{\mathrm{Max}}$), average gyration radius variation $\langle \Delta r_g \rangle$, minimum-maximum gyration radius variation (Δr_gMin and $\Delta r_g^{\mathrm{Max}}$).

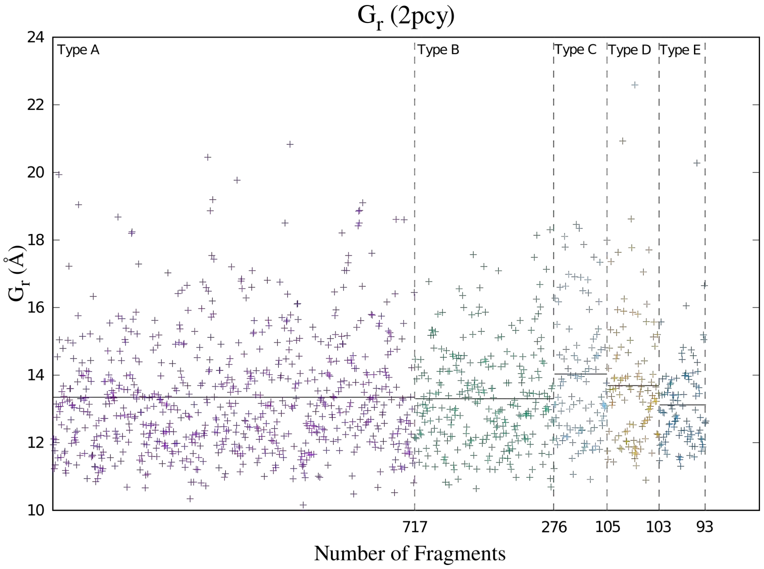


Figure 4.17: Gyration radius value for type A-E for 2pcy decoys. In solid black line the average gyration radius for each type.

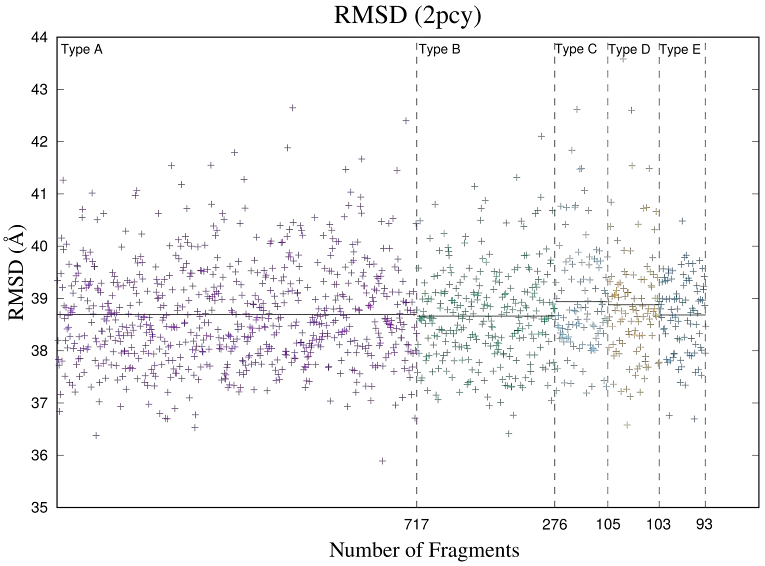


Figure 4.18: RMSD radius value for type A-E for 2pcy decoys. In solid black line the average gyration radius for each type.

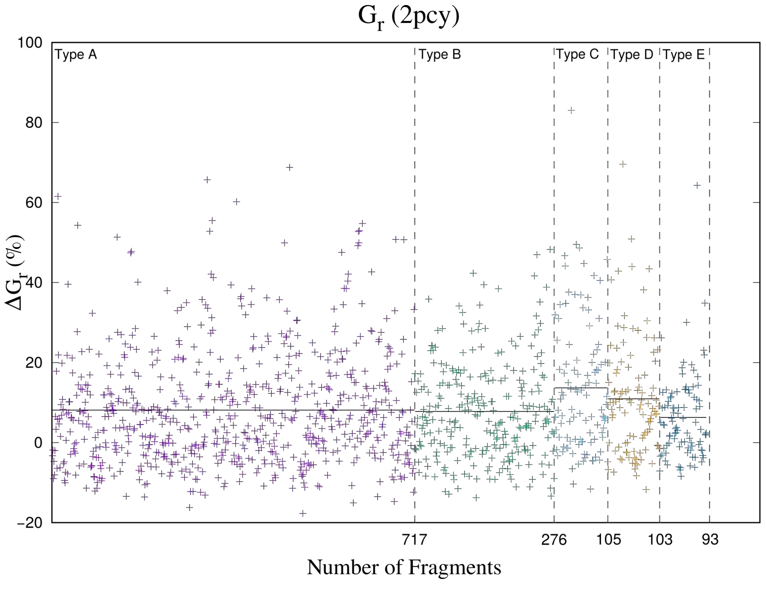


Figure 4.19: Variation in gyration radius value for type A-E for 4ouf decoys. In solid black line the average gyration radius for each type.

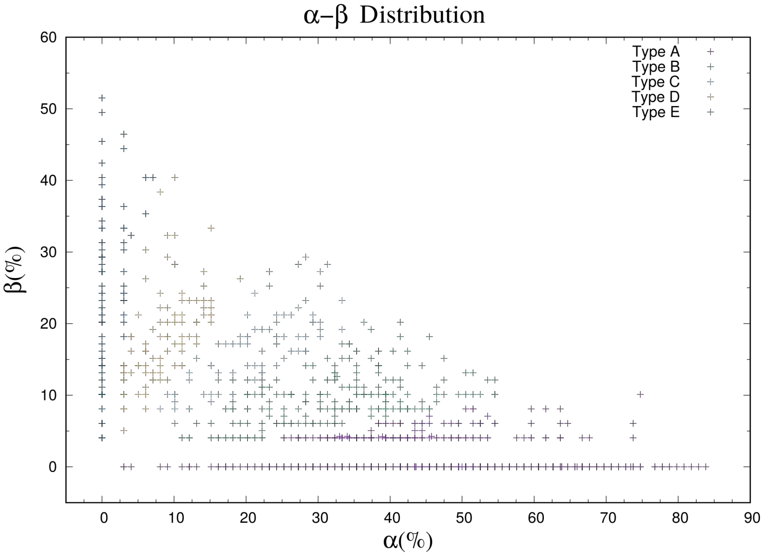


Figure 4.20: α-β percentage for 2pcy decoys generated.

Conclusion and perspective

The aim of the of this thesis was to understand the effect of solvent specificities on the stability of a specific protein. Water is a fundamental solvent of nearly, if not all, biological processes. Water was in fact selected by nature to this aim in view of this extraordinary properties, as briefly described in Chapter 1. Several studies have underlined the importance of considering other solvents to understand how live first emerged on Earth under probiotic conditions, as well as, how life could exist on other planets. Motivated by this studies, tackled the problem using two different approaches. In the first approach, we compute explicitly the change in free energy when specific amino acids are brought from a gas phase to water. To this aim, we used thermodynamic integration, a versatile tool that could be adapted to different solute, as full protein, and different solvent, as an apolar solvent. The results of this first part could be used to test the reliability of an alternating approach, called *morphometric approach*, to the calculation of this free energy change used in the second part.

The analysis of the data generated from the thermodynamic integration calculations, using gromos54a7, implemented with the ATB optimized structure, nonbonded and bonded interaction, produced results in general consistent with the experimental and calculated values (Figs. 3.1-3.2). Respect of the value calculated by Villa et al. [51] customized gromos54a7 with *Automated Topology builder* (ATB) features results more able to fit the experimental data. In comparison with free energy value available from ATB the gromos54a7 results slightly more precise. This precision could be effered to a most modern forcefield or the more precise feature of GROMACS version. These lineup has been used to understand, using the variation of free energy in function of λ and the cumulative free energy that returns the ΔG_{Solv}, the non-bonded interaction between solvent and compound. As default the λ steps have been set as 20 ($\Delta\lambda = 0.05$). Amino acid have several different chemical characteristic and, consequently, define different results depending on the chemical group-solvent interaction during the decoupling. Using the Bennet's acceptance ratio and ancillary data generated we were able to refine the λ value as needed to improve precision in particular simulation runs. This hypothesis has been confirmed several trending pattern evaluable in change of free energy as function of λ, and by the evidence that chemically correlated amino acids generated similar data. As stated before, the main intent of this analysis was not the validation, or a comparative study on a force field respect to another. Results in this first part of the thesis can be defined as a starting point to explore the free energy decoupling ”fingerprints” of a particular chemical group in water, and the respective interaction with other polar solvent or non polar as ethanol or cycloehexane, respectively. Exhaustive understanding of these interaction could have a pivotal role in the comprehension of contribute of single amino acid in numerical methods free energy calculation. Collimating this results with a parallel study of the compound in other solvent results could able to clarify the transfer free with mixed solvent phase as from low dielectric, as cycloehexane, to high dielectric, as water. The understanding of these phenomena in molecular dynamic permit to simulate biological mixed solvent as the integral membrane proteins. Further research, could focus on the analysis in computational and experimental in several solvent, of little peptide, with different amino acid composition in the way to more comprehend the interaction. Beside the molecular dynamics calculation these little peptide could be used, subsequently, in morphometric method proposed in the second part.

The second part of the thesis take on the free energy of solvation exploration from a non numerical methods, strictly considering the work done, standpoint. The development of a collection of protein, with different amount of secondary structure from an unique sequence of amino acids, the primary structure, has been required to explore the configuration phase with aim to find the free energy difference respect of the native state of the protein studied. Protein with the same primary structure but different geometry are called decoys. The distinctiveness of this methods is that use pre-existing native structure, or more precisely the three-dimensional backbone structure, to generate possible stable conformation. Proposed morphometric calculation of free energy calculation needs less computational time than molecular dynamics and can be used to scan large protein population. Beside the hypothesis to discover a different native state, or a configuration with a similar free energy of solvation, this collection can be used to investigate qualitatively alternative folding in different solvents. Algorithm proposed has been able to generate 1048 alternative configuration for the 4ouf as following reported: 564 structures all-α, 279 70%-α, 84 50%-α, 61 30%-α, 60 all-β; and 1294 alternative state for 2pcy of which 717 all-α, 276 70%-α, 105 50%-α, 103 30%-α, and 93 all-β. The minimum and maximum RMSD values for 4ouf and 2pcy are 51.16 Å to 92.47 Å and 35.88 Å to 43.58 Å respectively, producing a wide change in geometry as expected. The proposed routine, even thought generated a generous amount of protein decoys, could be surely improved as the efficiency of decoys generated. Using different sampling logic or acceptance could be possible generate more precise restraint so that it can be possible range the decoys variety.

Appendix A

Experimental methods

A.1 X-ray crystallography

This technique is based on the analysis of the atomic and molecular structures of a crystal using diffraction pattern generated by the scattering of a electromagnetic radiation, in this case the x-ray. The X-ray, photons with a wavelength λ ranged between 0.1 and 100 Å, are generated colliding high accelerated electrons on a tungsten surface. The impact generate photons by Bremsstrahlung effect. In a regular array of atoms, where the scatter are due by the atom's electron, X-rays scattering produces a regular array of spherical waves These waves can cancel out other waves through destructive interference or add constructively in a few specific directions, determined by Bragg's law:

$$2d \sin\theta = n\lambda \tag{A.1}$$

The bragg's law is correlated with the vector of scattering **K** where the module is related with the scattering angle θ:

$$|K| = \frac{1}{\pi}\frac{\sin\theta}{\lambda} \tag{A.2}$$

and results included between 0 and $1/\pi\lambda$.

The X-ray crystallography determine the density of charge $\rho(\mathbf{r})$ throughout the crystal, where r represents the three-dimensional position vector within the crystal. X-ray scattering is used to collect data about its Fourier transform $F(\mathbf{K})$. From this value is possible calculate the density defined in real space, using the formula

$$F(\mathbf{K}) = \int d^3r\rho(\mathbf{r}) \exp\left[-i\mathbf{K}\mathbf{r}\right] \tag{A.3}$$

$$\rho(\mathbf{r}) = \frac{1}{(2\pi)^3}\int d\mathbf{K}F(\mathbf{K})\exp\left[i\mathbf{K}\mathbf{r}\right] \tag{A.4}$$

The structure factor $F(\mathbf{K})$ is a complex number, therefore:

$$F = |F|^2\exp[i\phi] \tag{A.5}$$

$$F = F_{\mathrm{R}} + iF_{\mathrm{i}} \tag{A.6}$$

The intensity $I(\mathbf{K})$ calculated by the instrument is real number and is the product between the structure factor $F(\mathbf{K})$ and his complex conjugate $F^*(\mathbf{K})$

$$I(\mathbf{K}) = F(\mathbf{K})F^*(\mathbf{K}) = |F|^2 \tag{A.7}$$

From the structure factor is not possible calculate the phase parameter. Protein crystallization is predominantly carried out in water. Protein crystallization is generally considered challenging due to the restrictions of the aqueous environment, difficulties in obtaining high-quality protein samples, as well as sensitivity of protein samples to temperature, pH, ionic strength, and other factors. Proteins vary greatly in their physiochemical characteristics, and so crystallization of a particular protein is rarely predictable

After the crystallization, to find the unknown phase, isomorphic crystals needs to be preprepared. This kind of crystal are generated by diffusion of heavy atoms by a regent in the protein crystal. Silver nitrate ($AgNO_3$) is able to bind the with the thiol ($-SH$). The structure factor of the isomorphic protein crystal $F_{PH}(\mathbf{K})$ is the sum of the protein structure factor $F_P(\mathbf{K})$ and the heavy atoms $F_H(\mathbf{K})$:

$$F_{PH}(\mathbf{K}) = F_P(\mathbf{K}) + F_H(\mathbf{K}) \tag{A.8}$$

The amplitude $|F(\mathbf{K})|$ can be approximated using the difference between the difference between the isomorphic crystal and the real protein crystal. The Patterson can be calculated by:

$$\Delta P = \frac{1}{V}\sum_{\mathbf{K}} ||F_{PH} - |F_P||^2\exp[i\mathbf{K}\mathbf{R}] \tag{A.9}$$

where R is the atoms position. All the value of F_P are arranged on a circumference of radius $|F_P|$ centered in the origin. In the same way, all the value of F_{PH} are arranged on a circumference of radius $|F_{PH}|$ centered in the origin. The intersection points between these circumference that fulfill the equation are two phase ϕ_a and ϕ_b. To choose the right phase can be prepared more isomorphic crystal with a several different heavy atoms. According with the resolution achieved different details can be defined,

Resolution	Details
6 Å	α-helix appear like sticks. The protein can be divide in sub-uni.
3 Å	The backbone can be defined
2.5 Å	Almost all side chain is visible. The carbonyl of the peptide bond permit. to define the plane orientations
1.5 Å	All atom are visible

Table A.1: Resolution and details achieved in X-ray diffraction.

A.2 Calorimetry

As Anfinsen shown, protein folding is reversible thermodynamic processes. To study this fold-unfold process several experimental ultra-sensitive technique has been developed.

A.2.1 Differential scanning calorimetry

The differential scanning calorimetry, or DSC is a thermo analytical procedure that is based on the comparison about the amount of heat necessary to maintain at the same temperature a interest sample and a reference. The difference in temperature in a DSC experiment is curve of heat in function of temperature. Working in isobaric condition the change in heat is equal to the change in enthalpy.

$$\left(\frac{dQ}{dT}\right)_P = \frac{dH}{dT} \tag{A.10}$$

the entalpy can be obtained integrating

$$\Delta H_{\text{sample}} = \int \frac{dH_{\text{sample}}}{dT} dT \tag{A.11}$$

$$C_{\text{prot}} = C_{\text{prot}}^{(\text{ref})} \frac{\left(\frac{dQ}{dT}\right)_P}{\left(\frac{dQ_{\text{ref}}}{dT}\right)_P} \frac{m}{m_{\text{ref}}} \tag{A.12}$$

where m and m_{ref} are the masses of the sample and the reference, respectively.

A.2.2 Isothermal titration calorimetry

The isothermal titration calorimetry, or ITC is a used to determine the thermodynamic parameters of interactions in solution. This kind of routine is used to study the binding energy of little ligands in macromolecules, Like the energy of interaction of compound, like a drug, with an active site. In protein folding studies the titration is intended to unfold the protein. The experiment runs into a titration of a solution by a titrating, monitoring heat released or absorbed by the reaction. The instrument is composed by identical cells made of thermally conducting and chemically inert material. This cells are protected by an adiabatic jacket. Thermopile, an electronic device that converts thermal energy into electrical energy, are used to detect temperature differences between the reference cell and the sample cell. The reference cell, filled with buffer or water, Prior to addition of ligand, is preheat applying a constant power. Measurements consist of power required to maintain equal temperatures between the sample and reference cells. This plot is time dependent. Repeating the experiment at different temperature is possible calculate the sample heat capacity. From a typical plot of C_{P} can be obtained several information. The peak area is the change in enthalpy derived by the protein unfolding. From the shape of the peaks can be calculated the van't Hoff enthalpy $\Delta H_{\mathrm{van'tHoff}}$. The van't Hoff equation calculate the variation in the equilibrium constant K_{eq} of a reaction to the change in temperature T, given the standard enthalpy change, $\Delta H^{\ominus}$, for the process.

$$\frac{d\ln K_{\mathrm{eq}}}{dT} = -\frac{\Delta H^{\ominus}}{R}. \tag{A.13}$$

From the isothermal titration calorimetry data the van't Hoff entalpic can be rewritten as:

$$\Delta H_{\mathrm{van'tHoff}} = \frac{4RT_t^2 C_{\mathrm{p}}^{(\mathrm{max})}}{\Delta H_{\mathrm{calc}}} \tag{A.14}$$

where T_t is the transition temperature between the folded and the unfolded state, R is the universal constant of gas $C_{\mathrm{p}}^{(\mathrm{max})}$ is the maximum heat capacity and ΔH_{calc} is the area of the peak in the transition. In the event that the van't Hoff enthalpy $\Delta H_{\mathrm{van'tHoff}}$ and the calculated change in enthalpy ΔH_{calc}, so the change in free energy is zero, is possible to calculate the folding entropy.

$$\Delta H = T\Delta S + V\Delta P \tag{A.15}$$

and

$$\Delta Q = T\Delta S \tag{A.16}$$

definifg the entropy and enthalpy difference between the denature state and the native, so the contribute due to the unfolding process, as $\Delta_N^U S(T_t)$ $\Delta_N^U H(T_t)$, respectively, the unfolding s entropy $\Delta_N^U S(T_t)$ can be written as:

$$\Delta_N^U S(T_t) = \frac{\Delta_N^U H(T_t)}{T_t} \tag{A.17}$$

Know the various heat capacity at different temperature, the entropy and enthalpy of folding can be calculated in function of the temperature T:

$$\Delta_N^U H(T) = \Delta_N^U H(T_t) + \int_{T_t}^{T} \Delta_N^U C_p(T')dT' \tag{A.18}$$

and

$$\Delta_N^U S(T) = \frac{\Delta_N^U H(T_t)}{T_t} + \int_{T_t}^{T} \Delta_N^U C_p(T')d\ln T' \tag{A.19}$$

This procedure is dependent by the experiment condition therefore is possible obtain different folding temperature. Changing reducing agent or pH can lead to different results. The dependence of the temperature of transition T_t with the pH can be used to calculate the amount of proton species released during the experiment.

$$\Delta v_t = -\frac{\Delta H_{\mathrm{calc}}(T_t)}{2.303RT_t^2}\frac{dT_t}{d(\mathrm{pH})} \tag{A.20}$$

Appendix B

General algorithm for decoys construction

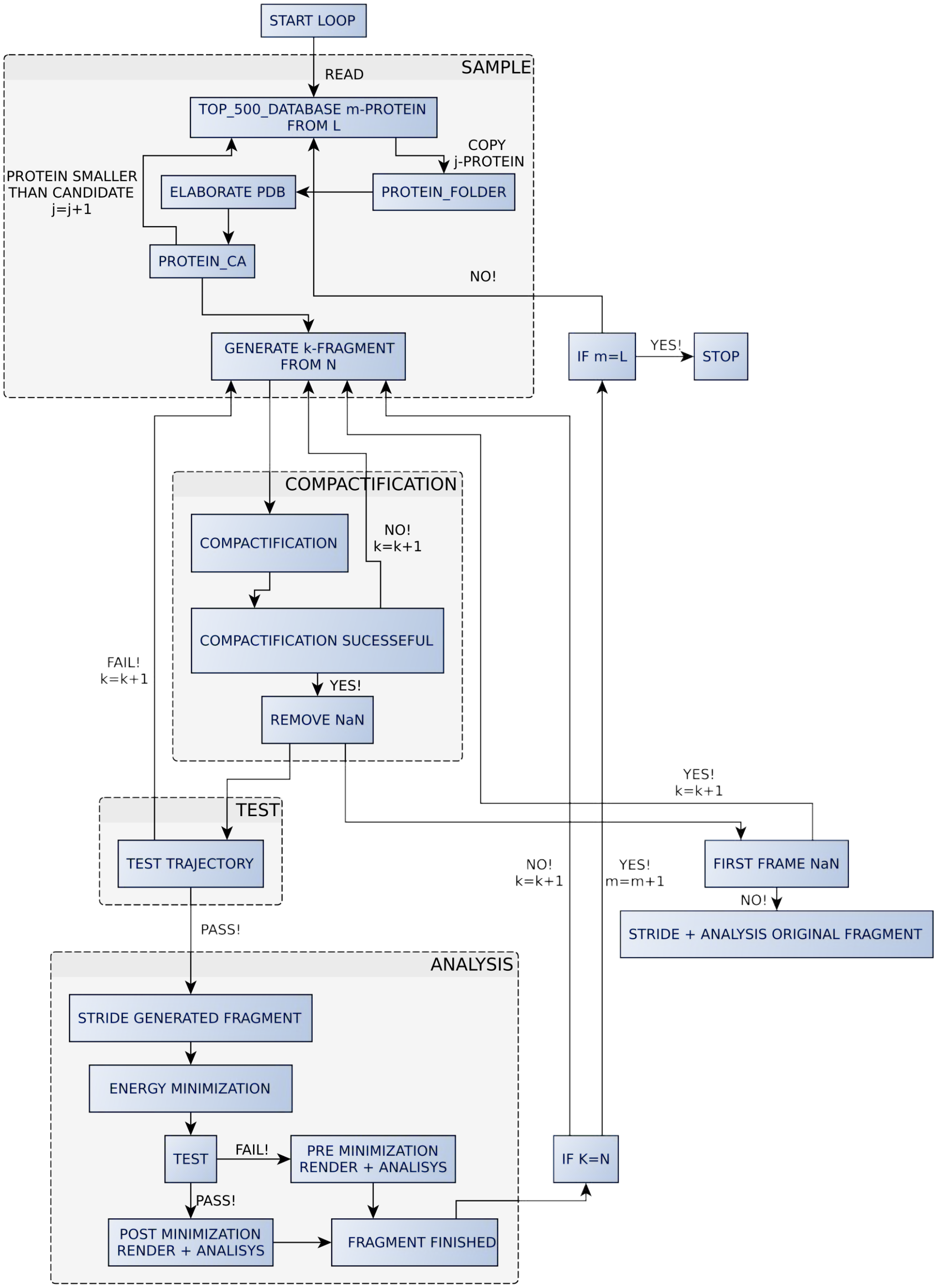


Figure B.1: General algorithm for decoys construction.

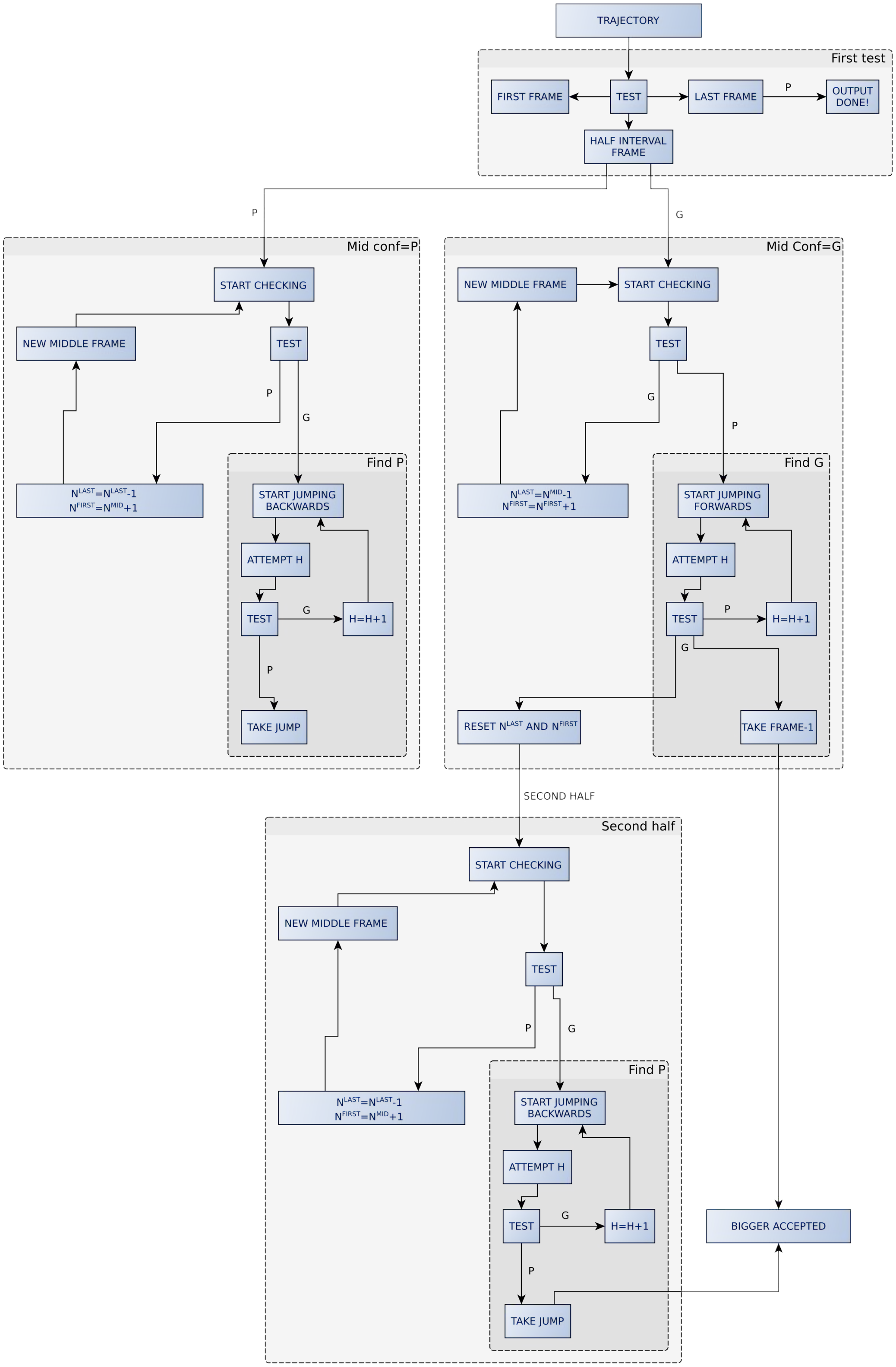


Figure B.2: Test algorithm procedure.

Bibliography

[1] D. Schulze-Makuch; S. Haque; M. R. de Sousa Antonio, D. Ali, R. Hosein, Y. C. Song, J. Yang, E. Zaikova, D. M. Beckles, E. Guinan, H. J. Lehto,S. J. Hallam. Astrobiology, Volume 11, Issue 3, 241-258 (2010).

[2] Cox) D. L. Nelson, M. M. Cox "Lehninger Principles of Biochemistry". W.H.Freeman & Co Ltd (2013)

[3] R. Horton, L. A. Laurence G. Scrimgeur M. Perry D.Rawn "Principles of Biochemistry" fourth edition Pearson Education (2006)

[4] S. Freeman "Biological Science" Prentice Hall; 2nd Edition edition (2005)

[5] K. Bosco, R. Brasseur. BMC Structural Biology, 5, 14 (2005).

[6] L. Pauling, R. B. Corey, H. R. Branson. Proceedings of the National Academy of Sciences, 37 (4) 205-211 (1951)

[7] G. N. Ramachandran, C. Ramakrishnan, V. Sasisekharan. Journal of Molecular Biology. 7: 95–9 (1963)

[8] https://commons.wikimedia.org/wiki/File%3AAlpha_helix_neg60_neg45_sideview.png

[9] https://commons.wikimedia.org/wiki/File%3AAlpha_helix_neg60_neg45_topview.png

[10] https://commons.wikimedia.org/wiki/File%3A3_10_helix_neg49_neg26_sideview.png

[11] https://commons.wikimedia.org/wiki/File%3A310_helix_topview.png

[12] https://commons.wikimedia.org/wiki/File%3APi_helix_neg55_neg70_sideview.png

[13] https://commons.wikimedia.org/wiki/File%3APi_helix_topview.png

[14] P.Y. Chou,G.D Fasman. Biochemistry. 13 (2): 222–245 (1974).

[15] J. Garnier, D. J. Osguthorpe, B Robson. J Mol Biol 120:97-120 (1978).

[16] M. Sela, F. H. Jr. White, C. B. Anfinsen. Science. 125 (3250): 691–692 (1957).

[17] C. Levinthal, J. Chem. Phys 65, 44 (1968).

[18] M. Mézard, G. Parisi, and M. Virasoro. ”Spin Glass Theory and Beyond”. World Scientific, Singapore (1987).

[19] C. Branden, J. Tooze. ”Introduction to Protein Structure”. New York: Garland Publishing (1991).

[20] J. D. Bryngelson, P. G. Wolynes. 84:7524-7528 (1987).

[21] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, P. G. Wolynes. Proteins, 21: 167–195. (1995)

[22] R. Dawkins ”The Blind Watchmaker”. New York: W. W. Norton & Company, Inc (1986).

[23] P. E. Leopol, M. Montal, J. N. Onuchic. Proc. Natl. Acad. Sci. USA. 89 (18): 8721–5. (1992).

[24] D. Baker, Nature 405, 39 (2000).

[25] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science 267, 1619 (1995).

[26] K. A. Dill and H. S. Chan, Nat. Struct. Biol. 4, 10 (1997).

[27] P. C. Hiemenz. ”Principle of colloids and surface chemistry”, Second edition Marcel Dekker, New York (1986).

[28] R. L. Baldwin,G. D. Roseg. Curr Opin Struct Biol. 23 (1): 4–10 (2013).

[29] K. A. Dill. Biochemistry. 24 (6): 1501–9 (1985).

[30] B. W. Matthews, H. Nicholson, W. J. Becktel. Proc. Natl. Acad. Sci. USA. 84(19), 6663–6667 (1987).

[31] D. Shortle. FASEB J. 10(1):27-34 (1996).

[32] J Mol Biol. 20;238(5):777-93 (1994).

[33] R. Ludwig. Angewandte Chemie International Edition, 40: 1808–1827 (2001).

[34] W. Kauzmann. Adv. Protein Chem, 14, 1–63 (1959).

[35] J. Chen, W. E. Stites. Biochemistry 40, 15 280–15 289 (2001).

[36] C. N. Pace. Biochemistry 40, 10–313 (2001).

[37] T. J. Taylor, I. I. Vaisman. BMC Structural Biology, 10(Suppl 1):S5 (2010).

[38] C. N. Pace, S. Treviño, E. Prabhakaran, J. M. Scholtz (2004). Philosophical Transactions of the Royal Society B: Biological Sciences, 359(1448), 1225–1235 (2004).

[39] B. J. Alder and T. E. Wainwright. The Journal of Chemical Physics 27, 1208 (1957).

[40] D. Fraankel, B. Smit. ”Understanding molecular simulation. From algorithm to applications” Academic press, USA (2002)

[41] A. R. Leach ”Molecular modelling. Principles and applications” second edition, Person Education Limited, England (2001)

[42] M. Abraham, B. Hess, D. van der Spoel, E. Lindahl. ”GROMACS Reference Manual, Version 2016” Department of Biophysical Chemistry, University of Groningen (2016)

[43] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. 21, 1087 (1953).

[44] J. G. Kirkwood. J. Chem. Phys., 3:300-313 (1935).

[45] C. H. Bennett. Journal of Computational Physics 22 : 245–268 (1976)

[46] S. Bruckner, S. J. Boresch. Comput. Chem., 32: 1303–1319 (2011)

[47] https://atb.uq.edu.au/

[48] K.B. Koziara, M. Stroet, A. K. Malde, A. E. Mark. Journal of Computer-Aided Molecular Design, 28, 221-233 (2014)

[49] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker,M. Winger,A. E. Mark, W.F van Gunsteren. European Biophysics Journal, 40, 843-856 (2011).

[50] A.K. Malde, L. Zuo, M. Breeze, M. Stroet,D. Poger,P. C. Nair, C. Oostenbrink, A. E. Mark. Journal of Chemical Theory and Computation, 7(12), 4026-4037 (2011).

[51] A. Villa, A. E. Mark. J. Comput. Chem., 23: 548–553. (2002).

[52] C. Jaeeon, M. L. Abraham, S. I. Sandler. The Journal of Physical Chemistry B 111 (8), 2098-2106 (2007).

[53] R. Wolfenden, L. Andersson, P. M. Cullis, and C. C. B. Southgate. Biochemistry. 20 (4), 849–855 (1981).

[54] R. C. Rizzo, T. Aynechi, D. A. Case, I. D. Kuntz†. Journal of Chemical Theory and Computation 2 (1), 128-139 (2006).

[55] P. D. Thomas, K. A. Dill. Proceedings of the National Academy of Sciences 93 (21) 11628-11633 (1996)

[56] B. Parka, M. Levitt. Journal of molecular biology. 258, (2) 367-92 (1996).

[57] J. Zhu, Q. Zhu, Y. Shi, H. Liu, Proteins, 52: 598–608 (2003).

[58] M. Kinoshita, J. Chem. Phys. 128, 024507 (2008).

[59] T. Yoshidome and M. Kinoshita, Phys. Chem. Chem. Phys. 14, 14554 (2012).

[60] M. Kinoshita, Biophys. Rev. 5, 283 (2013).

[61] H. Oshima and M. Kinoshita, J. Chem. Phys. 142, 145103 (2015).

[62] T. Hayashi, S. Yasuda, T. Škrbić, A. Giacometti, and M. Kinoshita. The Journal of Chemical Physics 147, 125102 (2017).

[63] P. M. König, R. Roth, and K. R. Mecke. Phys. Rev. Lett. 93, 160601 (2004).

[64] R. Roth, Y. Harano, and M. Kinoshita. Phys. Rev. Lett. 97, 078101 (2006).

[65] C. Das, S. Roy, S. Namjoshi, C.S. Malarkey, D.N. Jones, T.G. Kutateladze, M.E. Churchill, J.K. Tyler. Proc. Natl. Acad. Sci. USA 111: E1072-E1081 (2014).

[66] T.P Garrett, D.J. Clingeleffer, J.M. Guss, S.J. Rogersm, H.C. Freeman. J.Biol.Chem. 259: 2822-2825 (1984).

[67] B. Lee, F. M. Richards. J Mol Biol 55(3):379-400 (1971).

[68] M. L. Connolly, J. Am. Chem. Soc. 107, 1118 (1985).

[69] H. Deng, Y. Jia, and Y. Zhang, Bioinformatics 32, 378 (2016).

[70] T. A. Kunkel Proceedings of the National Academy of Sciences. 82 (2): 488–92 (1985).

[71] L. Serrano, L. J. Neira, J. Sancho, A. R. Fersht. Nature. 1992 Apr 2;356 (6368):453-5 (1992).

[72] http://kinemage.biochem.duke.edu/databases/top500.php

[73] J. Callaway, M. Cummings, B. Deroski, P. Esposito, A. Forman, P. Langdon, M. Libeson, J. McCarthy, J. Sikora, D. Xue, E. Abola, F. Bernstein, N. Manning, R. Shea, D. Stampf, and J. Sussman. Brookhaven National Laboratory (1996).

[74] T. Škrbić, T. X. Hoang, A. Maritan, J. R. Banavar, and A. Giacometti. (unpublished)

[75] P. Rotkiewicz, J. Skolnick. 29(9):1460-1465 (2008).

[76] D. Frishman, P. Argos. Proteins 23(4):566-79 (1995).

[77] W. Humphrey, A. Dalke, K. Schulten. J. Molec. Graphics, 14, 33–38 (1996).