Università
degli Studi
di Padova

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Fisica e Astronomia G. Galilei

Scuola di dottorato di ricerca in FISICA
Ciclo XXVII

Local sampling and statistical potentials for scoring protein structures

Direttore della Scuola : Ch.mo Prof. Andrea Vitturi
Supervisore : Ch.mo Prof. Antonio Trovato
Co-Supervisore : Ch.mo Prof. Flavio Seno

Dottorando: Stefano Zamuner

Abstract

Understanding and predicting the binding process of a protein with drug molecules or other proteins is nowadays of fundamental importance for boosting advancements in medicine and the production of protein-based therapeutic products. In this thesis we investigated three main aspects that are important for the comprehension and modeling of binding processes: (i) protein-solvent interactions (ii) change of the volume of the protein configurational space, and (iii) protein configurational modifications.

The study of the first two aspects led to the development of new methodologies in the field of knowledge-based potentials (KBPs). In particular we took into account entropic contributions to the binding affinity between proteins and demonstrated how this improves the performances of a simple KBP in estimating it. Moreover, the utilization of a simple physical-modeled KBP as a playground allowed us to recognize some of the limitations of this statistical approach: two new KBPs have been developed with the aim of overcoming these issues. The performances of one of these new methods in discriminating native state of proteins have proven to be comparable to that of other state-of-the-art methods.

In order to model the local protein configurational changes in the binding interface we developed a sophisticated numerical algorithm inspired by robotics. Through a generalized notion of *concerted rotation*, the developed method allows to locally perturb the backbone configuration of a certain region by modifying only an arbitrary number of degrees of freedom while leaving all the others unchanged. The efficiency of the methodology capitalizes on the inherent geometrical structure of the manifold defined by all chain configurations compatible with the fixed degrees of freedom. We then validated the proposed algorithm on few pedagogical examples.

Abstract

Comprendere il meccanismo di binding tra diverse proteine e tra proteine e piccoli ligandi è oggi di fondamentale importanza per accelerare lo sviluppo di prodotti terapeutici basati sull'utilizzo di proteine. In questa tesi sono state principalmente approfondite tre tematiche che sono basilari per la comprensione ed il modeling del processo di binding: (i) interazioni proteina-solvente, (ii) contrazione dello spazio conformazionale delle proteine dopo il binding e (iii) cambiamenti conformazionali delle proteine durante il processo di binding.

L'analisi dei primi due aspetti ha portato allo sviluppo di nuove tecniche nel campo dei potenziali statistici (knowledge-based potentials or KBPs). In particolare è stato considerato il ruolo di contributi entropici alla affinità di legame tra proteine ed è stato dimostrato come le performance di un semplice KBP nel predire le affinitaà di legame migliorino se si tiene conto di questi contributi. Inoltre l'utilizzo di un semplice KBP come banco di prova ci ha consentito di individuare alcune limitazioni di questo approccio: nel tentativo di superare queste limitazioni sono stati sviluppati due nuovi potenziali statistici. Le performance di uno di questi nuovi potenziali si sono rivelate pari o superiori a quelle di altri metodi all'avanguardia.

Al fine di descrivere i cambi conformazionali all'interfaccia tra proteine è stato sviluppato un algoritmo che generalizza la nozione di *rotazione coordinata* e che permette di modificare localmente un numero limitato di gradi di libertà interni alla catena, senza perturbarla globalmente. L'algoritmo sfrutta la geometria della varietà che definisce l'insieme delle possibili configurazioni compatibili con i vincoli di località imposti per assicurare la convergenza ad una soluzione. L'algoritmo e' stato verificato in diverse applicazioni.

Preface

Chapter 2. Section 2.2 is a summary of the work presented in [CGL+12], here presented with the permission of authors. Figure 2.2 is an adaptation of a Figure from the same paper. The work presented in Section 2.3.1 has been published in [SZC+13], of which I am an author.

Chapter 3. Paper in preparation [SZS+].

Chapter 5. The work here presented has been accepted for publication by PlosOne.

Table of Contents

Thesis outline

Proteins are large biological molecules consisting of a linear chain of amino acid residues serially linked together. They are main actors in all living organisms and participate in virtually every biological process within them. The vast array of functions protein perform include catalyzing chemical reactions (enzymes), providing structural and mechanical support, and ligand binding; moreover proteins are involved in DeoxyriboNucleic acid (DNA) replication and signal transduction.

Nowadays proteins are largely employed in many medical treatments and pharmacological products. Their applications in medical products have dramatically increased in number and frequency since the introduction of the first recombinant protein — human insulin — in 1982. Despite therapeutic proteins already have a significant role in almost every field of medicine, this role seems to be only in its infancy [LBG08].

The biological function of proteins has proved to be related to their three-dimensional *native* structure, which is protein natural conformation when put in solvent in biological conditions. Non native folding often is the leading cause of diseases in the host and, eventually, his death. Among human diseases caused by an incorrect folding (proteopathies) are Alzheimer's disease, Parkinson's disease, prion disease, type 2 diabetes and amyloidosis [WL00, CD06, CL97].

Since proteins activity is (at least partly) determined by their three-dimensional structure, huge efforts have always been dedicated to structure determination. Moreover many biological functions involve the formation of protein-protein or protein-ligand complexes, thus the understanding of such interaction is of fundamental importance for structure-based design of therapeutic proteins.

In the third decade of the twentieth century the development of X-ray spectroscopy has allowed scientists to experimentally resolve the structure of the first biological molecule (hexamethylenetetramine) while crystal structures of

proteins began to be solved only in the late 1950s, starting with the structure of the sperm whale myoglobin. This experiment earned Sir John Cowdery Kendrew the Nobel Prize in 1962. Since then, an increasing number of proteins have been resolved and more than 98000 protein structures are publicly available in the Protein Data Bank (pdb) at the beginning of 2015.
The availability of this large amount of data allows and encourages scientists from all over the world to try to deducing the rules governing protein folding and activity. During last decades significant progresses have been made in understanding some important processes in which proteins are involved.

Another field taking advantage of the large amount of experimental data available is that of native state recognition. Differently from other applications (Molecular Dynamics and Monte Carlo simulations for instance) this approach is not applied to understanding the *mechanisms* ruling processes in which proteins are involved but is aimed to capture those *characteristics* that distinguish proteins in their native configuration from abnormally folded or bounded ones. The recognition of native protein structures is of fundamental importance in *protein design* [KDI+03] (see Chapter 2) and, more generally, in protein structure prediction [RSMB04, SRK+99]. The instruments by excellence in this field are Knowledge Based Potentials (KBPs), empirical functions specifically devised for scoring/benchmarking the goodness of protein structures. The first KBP has been proposed by Miyazawa and Jernigan in 1985 [MJ85] and since then many valuable KBP have been introduced and reported in literature [SRK+99, LK00, KDI+03, RMF08, SR08, JTT92].

Aim of this thesis was to evaluate the feasibility of a new KBP for the study of protein-protein or protein-ligand binding processes. A good KBP for such applications should be able to recognize with good accuracy and reliability the correct binding between two monomers as well as to estimate their binding affinity.
Three main aspects of key importance to reach our aim have been examined:

- monomer-solvent interactions
- entropic contributions
- protein modifications upon-binding.

The Bayesian Analysis Conformation Hunt (BACH) KBP [CGL+12] was chosen as testing ground for examining and validating new developed concepts because of its simplicity and proven ability in recognizing native state of proteins.

The outline of the thesis is as follows.

In **Chapter 1** we briefly introduce proteins and recall some notions of biophysics. The risk in introducing such a vast topic is that of incurring in either errors or trivialities. Moreover experimental findings in biological areas are prone to be proven wrong or incomplete by successive studies [Ioa05, GG07, SZP+07, KPK+09, NFW11]. In the attempt to avoid these risks, proteins are introduced in a very abstract and unconventional way which should allow anyone to understand it. However, the reader who has familiarity with the topic will surely be facilitated in the comprehension of the chapter. This introduction is not intended to give a comprehensive knowledge on the topic but will deal mainly with those concepts that are useful for the understanding of central chapters, along with some hopefully interesting extras. For a more exhaustive introduction to proteins pleas consult for instance [Whi05, LNC08].

In **Chapter 2**, after a brief introduction on protein design and KBP we describe the BACH knowledge based potential that we used as benchmark for the development and implementation of new features aimed to improve its performances in recognize the native state of proteins. In the same chapter we also present and discuss two methods for determining macromolecule surface; the first method is a simplified version of the Linear Combination of Pairwise Overlaps (LCPO) algorithm [WSS99] while the second is based on the definition of the α-shape of a set of points. Both these methods have proved to be more precise and computationally efficient in determining the surface of protein complexes with respect to the original implementation of BACH.

In **Chapter 3** we discuss the importance of entropic contribution in the estimation of binding free-energy between two proteins. Starting from the structures of two monomers and of their complex we exactly computed the rotational and translational entropy loss upon binding. A vibrational contribution has been estimated by employing an Anysotropic Network Model (ANM), in which the spring constant has been obtained by matching the mobility profile of the structures obtained with the ANM with the ones obtained with short Molecular Dynamicss (MDs). Finally the interaction term has been estimated by using the BACH score. The total binding free-energies (ΔG) of 15 complexes, obtained as a weighted sum of these four contributions, have been compared to experimental values.

In **Chapter 4** we discuss the limitations hidden in the formulation of KBPs and propose two different potentials aimed to resolve the observed problematic.

In **Chapter 5** we investigate the possibility of perturbing a small set of backbone degrees of freedom in such a way that only a limited number of atoms are affected by the perturbation and no other degrees of freedom are modified. These *constrained backbone modifications* can be potentially employed in the study and modeling of conformational changes of protein-protein interfaces that occur upon-binding.

List of Acronyms

ANM Anysotropic Network Model. 3, 22, 45, 46, 48–50, 53, 54, 115, 116

BACH Bayesian Analysis Conformation Hunt. 2, 3, 28–35, 40, 41, 47, 49, 51, 56, 57, 59, 61, 63–66, 68, 114–116, 118

BLAST Basic Local Alignment Search Tool. 9

CASP Critical Assessment of protein Structure Prediction. 24, 28, 30, 31, 33, 34, 74

CGAL Computational Geometry Algorithms Library. 35, 37

DNA DeoxyriboNucleic acid. 1, 7, 9–11

DSSP Define Secondary Structure of Proteins. 29

FJC Freely Jointed Chain. 19

GCP Gaussian Chain Potential. 117, 118

IUPAC International Union of Pure and Applied Chemistry. 7, 8

KBP Knowledge Based Potential. 2–4, 24–26, 28, 31, 49, 59, 61–63, 65, 69, 71, 74, 114–118

LCPO Linear Combination of Pairwise Overlaps. 3, 32–34, 36, 37, 114, 115

LOO Leave One Out. 52, 57

MC Monte Carlo. 117, 118

MD Molecular Dynamics. 3, 21, 44, 45, 48–50, 54, 115, 118

MSE Mean Square Error. 46, 50

OLS Ordinary Least Squares. 46, 48

pdb Protein Data Bank. 2, 13, 14, 27, 48, 49, 51, 67

RMSE Root Mean Square Error. 52–57

RNA RiboNucleic Acid. 7, 10

SASA Solvent Accessible Surface Area. 27, 28, 30, 32, 34–36, 114, 115

SAW Self Avoiding Walks. 62

VMD Visual Molecular Dynamics. 30

Chapter 1

Proteins

Polymers (*poly*:many, *mers*:parts) are macromolecules that play a fundamental and ubiquitous role in living organisms due to their broad range of physical properties [Pai97]. They are composed of multiple repeating units of low relative molecular mass called monomers (*mono*:single). Depending on the nature of the monomers, polymers exhibit a variety of architectures (linear or branched). For instance carbohydrates can be both branched (e.g. amylopectin) and linear (e.g. amylose); on the contrary nucleic acids (DNA and RiboNucleic Acid (RNA)) and proteins are intrinsically linear polymers.

This chapter is mainly focused on the properties of proteins and partly on DNA. Proteins are composed of repeating units which are called *amino acid residues*, being the part which remains in the polymeric chain after amino acids are strung together: we will refer to them as both *residues* or *amino acids*. Units composing DNA are instead called *bases*.

1.1 Languages

The simple architecture of linear polymers is suitable to be described as a sequence of symbols that identify the order in which each monomer appears in the polymeric chain. Each monomer is in fact uniquely identified by a symbol, usually a character or a short string; the list of symbols necessary to label different possible monomers of a given class constitutes an alphabet α. Different classes of polymers require different alphabets in order to be appropriately described: as an example Tables 1.1 and 1.2 show the International Union of Pure and Applied Chemistry (IUPAC) alphabets for DNA and protein respectively.

Base Name	Symbol
Adenine	A
Cytosine	C
Guanine	G
Thymine	T

Table 1.1: IUPAC nomenclature for DNA bases

Name	Symbol		Name	Symbol	
	1 Letter	3 Letters		1 Letter	3 Letters
Alanine	A	ALA	Leucine	L	LEU
Arginine	R	ARG	Lysine	K	LYS
Asparagine	N	ASN	Methionine	M	MET
Aspartic Acid	D	ASP	Phenylalanine	F	PHE
Cysteine	C	CYS	Proline	P	PRO
Glutamine	Q	GLN	Serine	S	SER
Glutamic Acid	E	GLU	Threonine	T	THR
Glycine	G	GLY	Tryptophan	W	TRP
Histidine	H	HIS	Tyrosine	Y	TYR
Isoleucine	I	ILE	Valine	V	VAL

Table 1.2: IUPAC nomenclature for protein residues

The symbols defined in α, often referred to as *letters*, can be put together in a sequential way and according to some predetermined rules to form sequences, which virtually represent existing polymers of the corresponding class.
The whole set of possible sequences so generated is a *formal language* and the set of rules used in the formation process is called *grammar*. In simple cases, such this one, the grammar is identified by four elements [BHPS61, Cho56]:

- an alphabet α
- a set of *starting symbols*, that are symbols not belonging to α and usually represented inside chevrons $\langle\cdot\rangle$
- a list of production rules $\mathcal{P}$, that are used for substituting the starting symbols into combination of starting symbols and/or letters of the alphabet, here also referred as *terminal symbols*
- an initial symbol, chosen among the starting ones, that indicate how to

begin the substitution process.

An empty production rule is often employed to allow the process to end, and therefore the *stop symbol* ϵ is introduced.
For instance all the (finite and infinite) sequences composed by the letters of the alphabet α can be generated by the *grammar* $G = (\{\langle\mathrm{S}\rangle\}, \alpha, \mathcal{P}, \langle\mathrm{S}\rangle)$ with production rule $\mathcal{P}$:

$$\begin{aligned}\langle\mathrm{S}\rangle &\models \textit{any symbol defined in } \alpha, \langle\mathrm{S}\rangle\\ \langle\mathrm{S}\rangle &\models \epsilon,\end{aligned}$$

where the symbol $\models$ stands for "can be substituted into" or "is defined as" and $\langle\mathrm{S}\rangle$ is the initial symbol (as well as the only starting one). The first production rule tells that the sequences can be generated by appending any letter of α to the end of the existing sequence, in a recursive fashion. The process can go on indefinitely or can end if the empty production rule $\langle\mathrm{S}\rangle \models \epsilon$ is used. The number of letters that form the sequence is called the *length* of the sequence.

It was experimentally observed that each letter is not equally probable in DNA and protein languages and that this *usage bias* is specie-dependent. The study of the reasons of this bias is still an active research field in bioinformatics and biochemistry [JH03, SBG+05]. Despite the fact that some persistent characteristic in sequence composition have been found in the realm of prokaryotes [PMD06], the phenomenon is still under discussion. Among the factors which have been recognized to influence codon and, consequently, amino acid usage are the action of mutation, natural selection, and genetic drift [CWB06]. But many others factors, like translation efficiency [Mer03, Roc04], seem to bias both codon and amino acid usage.

The degree of identity between two sequences belonging to the same language is measured by employing a metric definition over the space of sequences. Basic Local Alignment Search Tool (BLAST) [AGM+90] is probably the most used method for determining the identity between two sequences (in this thesis we will refer to it whenever we talk about sequence identity) but various definitions of distance are used depending on the application.
Indeed the fast growing number of available sequences and the resulting need to process increasingly larger data sets have pushed toward the development of new and faster algorithms [Edg10]. A selection of other methods that are nowadays popular is reviewed in [Not02].

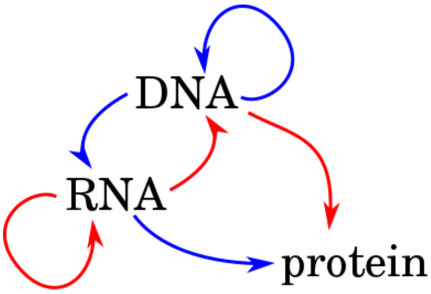


Figure 1.1: Graphical description of the central dogma of molecular biology. Arrows depict the information flow from nucleic acids to proteins. Blue arrows are related to general transfers (believed to occur normally in most cells) while the red ones are related to special transfers (known to occur, but only under specific conditions, e.g. in the case of some viruses or in the laboratory). The three other possible transfers are believed to never occur.

Sequence alignment is the preferred method for the determination of sequence homology and this is probably the reason why sentences like "$X\%$ homologous sequences" and "$X\%$ identity sequences" are commonly used interchangeably, despite the fact that only the latter is formally correct.

On the contrary to all spoken languages on Earth, protein and DNA sequences are often pangrammatic sequences, i.e. all symbols defined in their alphabet often appear at least once along the sequence. For a comparison, there are no pangrammatic sentences in this whole thesis.

DNA, RNA and protein sequences are closely related by the possibility to translate one into another. For instance DNA sequences can be translated into RNA ones and RNA can, in turn, be translated into a protein sequence. Among the possible nine translations between DNA, RNA and proteins only six naturally occur Fig. 1.1. The observation that the *sequential information* seems to flow from nucleic acid to protein and never back from protein to either protein or nucleic acid constitutes the *central dogma of molecular biology* [C+70].

The existence of the translation process DNA $\rightarrow$ RNA $\rightarrow$ protein allows to define the direction, or order, of RNA and protein sequences once the order of the corresponding DNA sequence is given. By convention DNA sequences are ordered from their $3'$ end to their $5'$ one and, as a consequence, proteins are ordered starting from their N-terminal monomer to their C-terminal one (see Section 1.3).

Even if DNA sequences are usually not directly translated into protein ones, it is possible to define the effective translation that is the result of the two combined translations DNA → RNA and RNA → protein. In this way each DNA sequence can be mapped to a corresponding protein sequence by sequentially mapping each successive triplet of bases (called *codons*) to one amino acid, as described by Table 1.3

Table 1.3: DNA to protein conversion table.

First Base	Second Base				Third Base
	T	C	A	G	
T	PHE	SER	TYR	CYS	T
					C
	LEU		ϵ	ϵ	A
				TRP	G
C		PRO	HIS	ARG	T
					C
			GLN		A
					G
A	ILE	THR	ASN	SER	T
					C
			LYS	ARG	A
	MET				G
G	VAL	ALA	ASP	GLY	T
					C
			GLU		A
					G

As can be seen from Table 1.3 different codons are translated into the same residue and as a consequence a certain amount of information is lost; sequential information contained in DNA is lossily compressed into a protein sequence during translation and the process can be more correctly seen as a lossy compression rather than a mere translation. This redundancy of the DNA language is of great importance for the evolution of species: a highly redundant sequence is indeed less prone to be affected by mutations of a single base (*single point mutations*) and is therefore more evolutionary stable.

1.1.1 A language of proteins

In this section we will focus on the properties of the sole protein sequences in light of the facts discussed above.

Physical and chemical properties of protein, and consequently its functions, are completely determined by its sequence of amino acids. For this reason

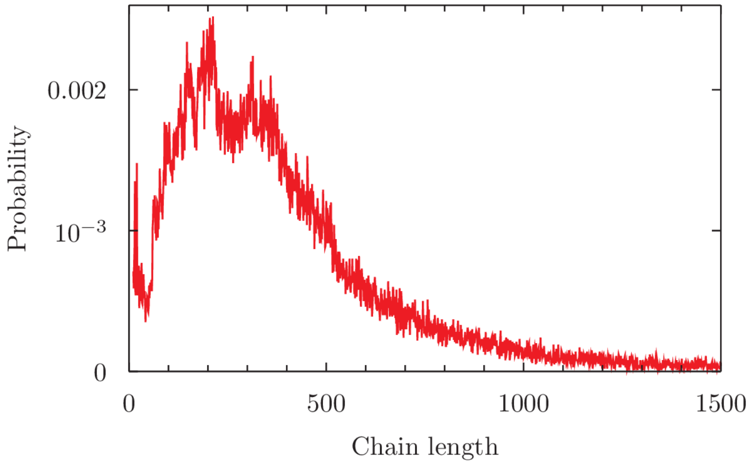


Figure 1.2: Protein length distribution computed by using 10^5 protein sequences from uniref50 database.

protein sequences perhaps constitute the most concrete example of what is known in ontology as a *substance*: a thing-in-itself, a property-bearer that must be distinguished from the properties it bears [Lan98].

The language that describes the universe of protein sequence is generated by the *grammar* $G = (\{\langle \mathrm{S} \rangle, \langle \mathrm{T} \rangle\}, \alpha_{prot}, \mathcal{P}, \langle \mathrm{S} \rangle)$ with production rule $\mathcal{P}$

$$
\begin{aligned}
\langle \mathrm{S} \rangle &\models \mathrm{M}, \langle \mathrm{T} \rangle \\
\langle \mathrm{T} \rangle &\models \textit{any symbol defined in } \alpha_{prot}, \langle \mathrm{T} \rangle \\
\langle \mathrm{T} \rangle &\models \epsilon.
\end{aligned}
$$

This is a slightly more complicated version of the production rule encountered in the previous section. The drawback arises from the fact that every protein sequences begin with a methionine residue at the N-terminal end, with the only exception of those proteins that undergo to some post-translation processes that cause a removal of the N-terminal portion of the original methionine-starting sequence.

The remaining positions in the chain are not constrained to any specific residue, but nonetheless the sequence of amino acids in proteins are not completely random. The probability of each amino acid to occur at a given position along the chain depends both on the position and on the type of previous and following residues.

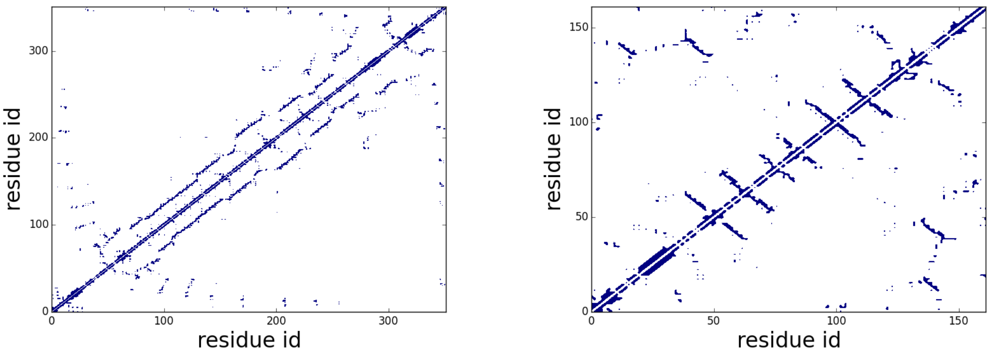
Usually protein length range from ~ 20 to $\sim 10^3$ residues. Lengths distribution is very peculiar and exhibits a peak in the range of $100 - 300$ with complex species having a peak at higher length with respect to less species [BK05].

1.2 Contact maps

The three-dimensional arrangement of a protein chain in space constitute the protein *tertiary structure*.
Describing a protein structure consists in characterizing the position of all its atoms in space. These information are usually stored in formatted files [BHN03] (pdb,pdbx,mmcif) and are freely accessible on the web from any of the Protein Data Bank [BWF+00] servers. A pdb code, or accession number, is assigned to each protein structure that has been experimentally resolved and uniquely identifies it.

However, more schematic descriptions can be drawn with much less amount of information. Indeed a first insight into protein structure can be obtained by using *binary contact maps*, that are square matrices having the same size N of the protein length, and whose elements Δ_{ij} are equal to one if and only if residues at positions i and j are *near* in space, whereas they are equal to zero otherwise. The vicinity relation employed in the definition of the contact map is usually determined by computing the distance between the residues and checking if it is lower than a certain threshold value D. For completeness we introduce in the notation a superscript characterizing the threshold used for computing the map: $\Delta \to \Delta^D$. Since the vicinity relation is symmetric the binary contact map Δ^D is symmetric too. The distance between the residues is typically defined as the distance between two reference atoms or the minimum distance between the atoms of the residues.
The binary contact map gives information about the topology of the protein but, in general, does not uniquely identify its tertiary structure [VMDL+08]; moreover not every possible binary matrix corresponds to a physical protein structure. Nonetheless contact maps are useful for studying in an effective way the possible folds that a protein can assume and are therefore used to describe similarity between protein structures [HS96].

As an example Fig. 1.3 shows the contact maps of two proteins (pdb codes $2PEC$ and $4OMX$): interacting residues are highlighted by a blue colored entry in the matrix; the main diagonal Δ_{ii} (from bottom-left to top-right), that describes the interaction of each residue with itself, is considered empty. The first characteristic that can be noticed by looking at contact maps is the presence of patterns. The most occurring patterns are given by contacts between pair of residues that follows the relation $(i+k, j-k)$ and $(i+k, j+k)$ with varying k and $i \neq j$. These patterns correspond to very specific structural



(a) Contact map of $2PEC$ pdb file (b) Contact map of $4OMX$ pdb file

Figure 1.3: Contact maps in which it is possible to recognize a parallel β pattern (left) or an antiparallel β pattern (right). Alpha helices are present in both contact maps. A threshold $D = 4.5$ Å was used in both cases.

arrangements that constitute the *secondary structure* of proteins.
The former one corresponds to an anti-parallel arrangement of the chain known as *anti-parallel β-sheet.*

Different structural arrangement instead contribute to the $(i + k, j + k)$ pattern.
The case $k = 1$ is trivial, being due to the fact that each residue is always connected with both the preceding and the following ones (with the obvious exception of the terminal residues). Structures corresponding to $k = 4$ are called *four-helices* or *α-helices* and finally the case $k > 4$ represents *parallel beta sheets* structure.

The probability of detecting a contact between two monomers of a chain rapidly drops when their separation along the sequence increases and is usually negligible for distances larger than $m = 30 \sim 35$ residues. The curve in Fig. 1.4 shows how the contact probability decreases as a function of the separation of two residues along the sequence. Even if the probability is a monotonic decreasing function of the separation distance m, an irregularity at $m \sim 27$ is evident. This irregularity has first been noted by Berenzovsky [BGT00] in 2000 and is at the base of Trifonov's theory of protein evolution [TKKB01, TB03].

1.3 Tertiary structure

Until now we depicted a very abstract image of proteins by focusing on the properties of sequences and contact maps. In this section we discuss in details

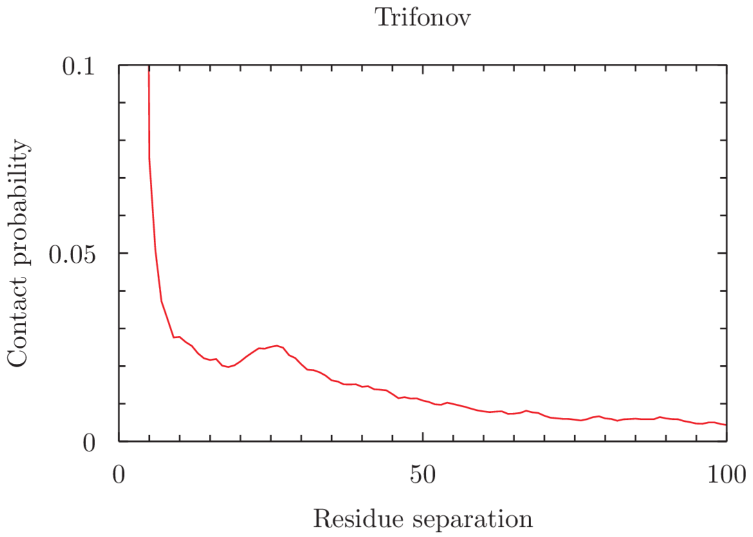


Figure 1.4: Probability of observing a contact between two residues at a fixed separation in sequence. The graph has been obtained by checking the number of pairs of alpha carbons whose distance is below the threshold of 7.0 Åin the whole top8000. The peak visible at $x \sim 27$ was firstly observed by Berenzovsky.

the characteristics of three-dimensional structure of proteins. The architecture of residues is first examined and subsequently used to understand the architecture of the whole chain.

Every amino acid shows a similar structure (see Fig. 1.5) with a central carbon atom, named α-carbon and labeled CA or Cα, which is bonded to an amine group NH_2, to a carboxyl group COOH and to a sidechain R that is residue-specific. Proline (PRO) constitutes the only exception to this scheme because its side-chain is bonded to the nitrogen of the amine group to form a ring.
Side-chains exhibit a large variety of chemical and physical properties that can be grouped in negatively charged (GLU, ASP), positively charged (ARG, LYS, HIS) or neutral. Moreover side-chains can be hydrophobic (ALA, VAL, ILE, LEU, MET, PHE, TYR, TRP) or polar (SER, THR, ASN, GLN) [BR03].

During protein synthesis the amino acidic chain is built by recursively join a residue to the already formed portion of the chain. During the polymerization a condensation process takes place between the carboxyl group of the former sequence and the amino group of the residue to be added. As a consequence a covalent bond, named peptide bond, is formed between the carbon and the nitrogen atoms. The significant delocalization of electrons gives the bond a partial double bond character, which is responsible for the planarity of the four atoms of the peptide group O=C−N−H.

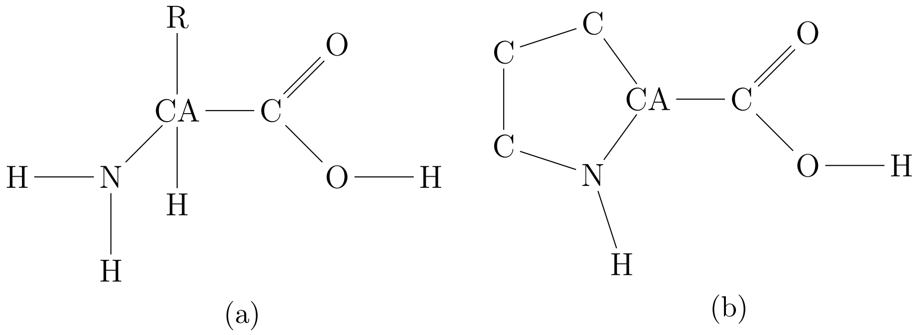


Figure 1.5: Amino acid structure. Fig. 1.5a shows the structure shared by all amino acids except proline. Fig. 1.5b shows proline chemical structure.

Figure 1.6 shows a sketch of the resulting chain. As can be seen the first residue exposed an amine group NH_2 while the last one exposes a carboxyl group ($COOH$): for this reason they are called N-terminal and C-terminal residues, respectively. Peptide linkages, along with N-C$_\alpha$ and C$_\alpha$-C bonds, form the so called *protein backbone* (highlighted with bold lines in Fig. 1.6). Bond length and bond angles assume a very narrow range of values about their average and can be thought to be approximately constants.

The planarity of the peptide bond implies that the torsional angle ω, defined as the angle between the planes CA,C,N and C,N,CA, is either 0 or π. The two possible isomer forms characterized by these two values of ω are named *cys* and *trans* isomers, respectively. In the folded state of proteins the trans isomer is overwhelmingly preferred in most peptide bonds, with rare exceptions. Prolin, on the contrary, is the only residue that is often found in the cys isomer (roughly 3:1 ratio in trans:cys population versus the roughly 1000:1 typical of the remaining residues [RM76]). The trans-cis isomerization of proline plays a key role in the rate-determining steps of protein folding [WWS02].
The conditions of constant bond angles and bond length, together with $\omega = \pi$ fix the distance of two consecutive α carbons to be 3.8 Å.
The remaining backbone torsional angles are usually labeled with the greek letters ϕ and ψ and can be used to fully parameterize the backbone since they are its only unconstrained degrees of freedom. The distribution of the pairs (ϕ, ψ) inside proteins has been firstly studied by Ramachandran who, in 1963, introduced the Ramachandran's plot [RRS63]. Ramachandran's plot has since then become a standard tool for the determination of protein structure [MMHT92] and the definition of secondary structures [MS94]. Using an analysis of local hard-sphere repulsions between atoms that are at least third

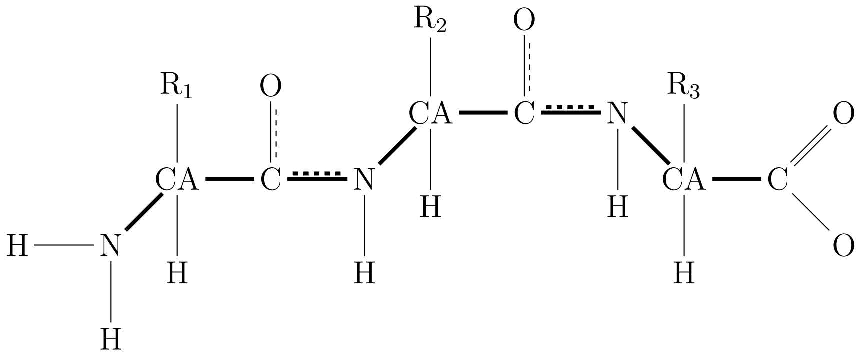


Figure 1.6: A sketch of the architecture of a protein. Residue side-chains R are labeled with consecutive integers. The protein backbone is highlighted in bold.

neighbors, Ramachandran constructed a steric map of the Ramachandran's plot that predicted the commonly allowed α_R, α_L and β regions and has become the standard interpretation of the Ramachandran's plot. Figure 1.7 shows the Ramachandran's plot obtained from the structures of the top500 database. Despite the fact that various studies have refined the calculation of the Ramachandran plot [Ram68] the shape of the allowed regions of the diagram is not perfectly understood and is yet a subject of study [HTB03].

Side chain configurations are certainly not frozen and different degrees of freedom are associated to different residue side chains. The free degrees of freedom of residue side chains play an important role in protein folding [BD94] and protein-protein interaction [Zac03]. Despite recent advantages in the understanding of the protein folding mechanism [MLM02, CT93, Bak00], the folding process of a protein in yet not well understood. Among the driven forces responsible for the correct folding are the hydrophobic effect and inter-atomic interactions between the atoms of the protein, but [Bak00] suggests that folding rates and mechanisms are largely determined by the topology of the native (folded) state.

Following the linguistic metaphor of Section 1.1, and in analogy with the definition of *word* in spoken languages as the "minimum unit with semantic or pragmatic content", we define a word in the context of *protein language* as the minimum (sub-)sequence that exhibit a well defined native structure, independently from the eventual presence and nature of other residues belonging the same chain. This concept correspond to the definition of *protein domains* in biology.

The thermodynamic properties of a polymer composed of N monomer, by virtue of the polymer linkage, is different from that of a gas of N isolated

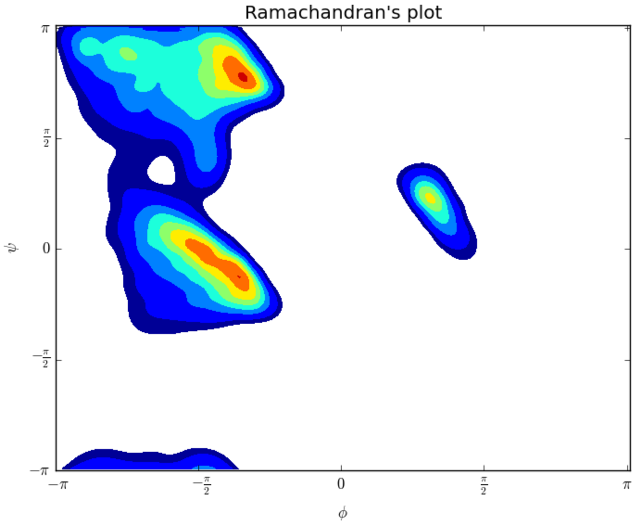


Figure 1.7: Ramachandran's plot obtained from the structures of top500. The density has been normalized: red regions correspond to highly favored values while blue regions are less favored. The highly disfavored configurations, whose probability is less than $10^{-}4$, are represented in white (see [CAH+09] for the definition of favored/disfavored regions).

monomers at any nonzero temperature T. Unlike gases, polymers would not occupy the whole available volume in order to maximize their entropy. This brings in a quantity very special to polymers, namely the equilibrium size of a polymer, in addition to the usual thermodynamic quantities.

The size of a protein, to be distinguished from its length, is quantified through two quantities called end-to-end distance R_e and radius of gyration R_g: the former measures the distance between the first and the last residues while the latter is the average root mean square distance of the protein atoms from their center of mass. If the three-dimensional position of the α-carbon of residue k is denoted as $\mathbf{r}_k^0$ the end-to-end distance and the radius of gyration can be computed as

$$R_e = || \left(\mathbf{r}_1^0 - \mathbf{r}_N^0\right) || \tag{1.1a}$$

$$R_g^2 = \frac{1}{N} \sum_k \left(\mathbf{r}_k^0 - \bar{\mathbf{r}}^0\right)^2 , \tag{1.1b}$$

where the vector $\bar{\mathbf{r}}^0$ is the average position of all the alpha carbons of the structure.

1.4 Protein size

The dependence of the polymer size on its length and how this dependence affects other thermodynamic properties are among the main topics in the field of polymer physics.
The success of exact methods, scaling arguments and the renormalization group crafted the statistical physics approach to polymer physics into a well defined and recognized field. These topics are covered in various monographs [Yam71, DCJ90, BGM13]; here we briefly introduce a simple model of polymer chain, called Freely Jointed Chain (FJC) model, and summarize some results of the famous Flory's theory [FV+69].

A FJC is a flexible chain composed by N monomers that are serially linked one to another to form a linear chain. The bonds variable $\boldsymbol{\tau}_j$, i.e. the direction of the vector connecting two consecutive monomers at positions $\boldsymbol{r}_j$ and $\boldsymbol{r}_{j-1}$, is supposed to have a fixed length $|\boldsymbol{\tau}_j| = b$; the flexibility is expressed as an absence of correlation between any two bonds

$$\langle \boldsymbol{\tau}_i \cdot \boldsymbol{\tau}_j \rangle = b^2 \delta_{ij}. \tag{1.2}$$

It can be shown [BHM05] that the end-to-end distance of a FJC follows the gaussian distribution

$$P(\mathbf{R}_e, N) = \left(\frac{3}{2\pi\sigma^2} \right)^{\frac{3}{2}} \exp\left(-\frac{3R_e^2}{2\sigma^2} \right), \tag{1.3}$$

in which the variance, given by $\sigma^2 = Nb^2$, also corresponds to the expected value of the square of the polymer size $\langle R_e^2 \rangle = Nb^2$. If one assume a functional dependence of the end-to-end distance on the chain length of the form

$$< R_e >= N^\nu \, b, \tag{1.4}$$

where ν is a size exponent, the FJC model predicts $\nu = \frac{1}{2}$.

The FJC model describes an ideal (non-interacting) chain in which the orientation of bonds are not correlated along the chain. A simple way to introduce the steric effects between non-consecutive monomers is to consider an hard-sphere repulsive energy proportional to the excluded volume V_{excl} of one monomer pair times the number of monomer pairs per unit volume

$$\textit{repulsive energy} \propto \frac{N^2}{R_e^3}\, V_{excl}; \tag{1.5}$$

notice that the available volume is supposed proportional to R_e^3. Once the entropy of the chain is computed from Eq. (1.3)

$$S(R_e, N) \propto -\frac{R_e^2}{Nb^2}, \tag{1.6}$$

the total free-energy of the system can be estimated as a linear combination of the energetic term in Eq. (1.5) and of the entropic term in Eq. (1.6):

$$F(R_e, N) = \beta_0 + \beta_1 \frac{N^2}{R_e^3}\, V_{excl} + \beta_2 \frac{R_e^2}{Nb^2}, \tag{1.7}$$

where β_0, β_1 and β_2 are three temperature-dependent parameters that account for the undetermined proportionality constants in Eq. (1.5) and Eq. (1.6) as well as for the remaining parts of the free-energy. The expected size of the polymer can be computed by minimizing the free-energy with respect to the end-to-end distance. As a result the size exponent ν for a purely repulsive chain is predicted to be $\nu = \frac{3}{5}$.

The case in which monomers attract each others can be solved by simply noting that, in a more or less compact packing of hard spheres, the volume enclosing the system is roughly proportional to the number of monomers, and therefore $\nu = \frac{1}{3}$.

It is therefore possible to distinguish three regions in the phase diagram of a polymer:

- a compact phase, described by a size exponent $\nu = \frac{1}{3}$, that is dominated by the effective attraction between the monomers

- a swollen phase, described by a size exponent $\nu = \frac{3}{5}$, in which the behavior of the polymer is dominated by the hard-sphere repulsion between monomers

- an intermediate phase, described by a size exponent $\nu = \frac{1}{2}$, in which the effects of attraction and repulsion between the monomers compensate each other and as a consequence the end-to-end distance distribution is that of an ideal gaussian chain.

1.5 Fluctuations

Proteins in the folded state are not rigid but fluctuate near equilibrium positions and sample numerous conformations in the neighborhood of their native conformation [FSW91, ADJ+01] . These fluctuations are usually small in magnitude, not exceeding several Ångstroms, and lie in the sub-nanosecond frequency range. The details of the atomic motion due to equilibrium fluctuation can be in principle elucidated by MD simulations and by using *ad-hoc* all-atom empirical potentials. This approach however becomes computationally inefficient when increasing the size of the system. Therefore coarse-grained protein models and simplified force fields therefore becomes particularly appropriate for describing the collective motions of large complex systems [BEJ+99, ADJ+01] that can not otherwise be investigated with atomic models [BEJ+99, BJ99, KG99]. A first approximation that can be done is considering an harmonic approximation to the real potential function and to describe the system as a network of coupled harmonic oscillators.

By assuming that each residue k at position $\boldsymbol{r}_k$ is assigned a mass m_k, the Lagrangian of the system can be written, with harmonic approximation, as

$$\mathcal{L} = \frac{1}{2}\dot{\boldsymbol{q}}^T \boldsymbol{M} \dot{\boldsymbol{q}} - \frac{1}{2}\boldsymbol{q}^T \boldsymbol{H} \boldsymbol{q}, \tag{1.8}$$

where $\boldsymbol{q}$ are a set of generalized coordinates describing the atom positions and $\boldsymbol{M}$ is the mass matrix

$$M_{ij} = \sum_k m_k \frac{\partial \boldsymbol{r}_k}{\partial q_i} \frac{\partial \boldsymbol{r}_k}{\partial q_j}. \tag{1.9}$$

The matrix $\boldsymbol{H}$ is instead the Hessian matrix of the potential energy with respect to the generalized coordinates. This system can be compared to a system of uncoupled harmonic oscillators by introducing the change of variables $\boldsymbol{q} = \boldsymbol{U}\boldsymbol{\theta}$ such that

$$\boldsymbol{U}^T \boldsymbol{M} \boldsymbol{U} = \mathbb{I} \tag{1.10a}$$

$$\boldsymbol{H}\boldsymbol{U} = \boldsymbol{M}\boldsymbol{U}\boldsymbol{\Lambda}, \tag{1.10b}$$

with $\boldsymbol{\Lambda}$ diagonal: the resulting system of harmonic oscillators is described by the equations $\ddot{\theta}_i = -\boldsymbol{\Lambda}_{ii}\, \theta_i = -\omega_i^2\, \theta_i$. We can take advantage of this mapping in order to compute the partition function $\mathcal{Q}$

$$\mathcal{Q} = \prod_i \frac{2\pi}{\beta h \omega_i} \tag{1.11}$$

and the free-energy F

$$\beta\, F = 3N \ln\left(\frac{\beta h}{2\pi}\right) + \sum_i \ln\left(\omega_i\right) \tag{1.12}$$

of the original system, where we denoted h the Plank's constant and indicated with β the inverse temperature, N is the number of residues and 3N the total number of degrees of freedom. We will be interested in the entropic contribution to the free energy:

$$\frac{s}{K_B} = -\sum_i \ln(\omega_i) - 3N \ln\left(\frac{\beta\hbar}{e}\right) \tag{1.13}$$

. Computationally the procedure to obtain the matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{U}$ is as follows. Defined $\boldsymbol{\Lambda}_M$ and $\boldsymbol{U}_M$ as the eigenvalue matrix and the eigenvectors matrix obtained from the diagonalization of the mass matrix $\boldsymbol{M}$, the matrix $\boldsymbol{\Lambda}$ is the eigenvalue matrix of $\boldsymbol{H}_M$, where

$$\boldsymbol{H}_M = \boldsymbol{\Lambda}_M^{-\frac{1}{2}} \boldsymbol{U}_M^T \boldsymbol{H} \boldsymbol{U}_M \boldsymbol{\Lambda}_M^{-\frac{1}{2}}. \tag{1.14}$$

The eigenvectors matrix $\boldsymbol{U}'$ of $\boldsymbol{H}_M$ defines instead the matrix $\boldsymbol{U}$

$$\boldsymbol{U} = \boldsymbol{U}_M \boldsymbol{\Lambda}_M^{-\frac{1}{2}} \boldsymbol{U}'. \tag{1.15}$$

In ANM, each pair of residues is assumed to be coupled by harmonic potential

$$V_{ij} = \frac{1}{2} k \left(r_{ij} - r_{ij}^0\right)^2, \tag{1.16}$$

where r_{ij} is the distance between residues i and j, while r_{ij}^0 is the equilibrium distance between the two residues in the native state. No distinction is made between different types of amino acids, so that a generic force constant k is adopted for the interaction potential between all pairs of sufficiently close residues. The success of ANM is due to its simplicity and, at the same time, to its ability to predict many equilibrium properties related to collective modes of motion.

Chapter 2

Knowledge Based Potentials

Since the early 1980's [MJ85], scientists have been trying to use the ever-increasing data availability about protein structure and function to redesign existing proteins, and more recently, to design entirely new proteins. Many promising designing methods have been proposed in the last decades [DWL89, BBL$^+$95, DM97, HPT$^+$98, KDI$^+$03].

Protein design consists the creation of novel protein sequences with arbitrarily chosen three-dimensional structures [RR89]. Given a target structure, the design process can be coarsely divided into two steps.

The first step consist in the proposal of a test sequence S that, during the second step, has to be tested with some suitable and specific method. The testing methods usually takes advantage of the observed regularities exhibited by protein structures in order to assess the quality of the proposed protein. Indeed different types of amino acids, having their own chemical features, are found in different strategic positions of the structure to stabilize the fold. A collection of carbon-rich amino acids, like leucine and phenylalanine, are usually placed inside the protein, and lock perfectly together. On the other hand, charged amino acids, such as lysine and aspartic acid, are typically spread across the surface to make the protein soluble in water. Hydrogen-bonding amino acids, such as serine and asparagine, are spread in strategic positions to tie different portions of the chain together. Finally, glycine and proline are added to redirect the chain in the proper direction. This combination of favorable forces locks the protein chain into a stable and compact structure.

The two steps described above are iterated several times and the sequence that better fit the target structure is finally named the *designed* sequence.

Ideally the iteration would happen over every possible sequence having the right length but this is practically impossible: indeed even enumerating all the sequences of a small protein composed of only 30 residues is unfeasible with modern computers. The solution is restricting the search space to some subset of sequences with desired characteristics such as a reasonable ratio of hydrophobic and hydrophilic residues. The test sequence is often obtained by modifying the sequence of a protein with a structure similar to the target one (protein redesign).

In order to design a protein that will successfully fold the previously described steps are not sufficient and to assure that the protein only has one stable structure is also necessary. Indeed any other fold compete with the desired stable structure thus interfering with it. Therefore designing a stable protein structure is not enough and the design of a protein structure that is unstable in every other conformation is needed.

We focused our attention on a class of methods that are most frequently employed in the second stage of protein design and that are referred to as KBPs. Despite their popularity and success [KDI+03] in the field of protein design, the application of these methods is also exploited in other areas where the validity and accuracy of a proposed tertiary structure is of fundamental importance, as for instance protein structure prediction and refinement. Indeed many of the several scoring functions that have been proposed for recognizing protein native structures [FVM07, DWL06, MSE03, ZZ02, FS06, RMF08] are rooted on KBPs. Broadly speaking KBP are energy functions derived from databases of known protein conformations that empirically aim to capture the key aspects of the physical chemistry of protein structure and function. As benchmarked during the biennial Critical Assessment of protein Structure Prediction (CASP) [MFK+09], the performances of KBPs in the recognition of native states are reasonably good, but significant room for improvement remains. For the advantage of the scientific community, many of these methods are publicly available through web servers so that users may have direct access to scoring functions for quality estimation of their own protein structural models.

2.1 Assess the quality of a structure

KBPs are potential functions that associate a real number to the three-dimensional configuration of a (macro-)molecule. The molecular structure is often represented by mean of a coarse-grained model, in which groups of atoms are merged together into a single effective unit or particle. The term *knowledge based* reflects the fact that the parameters modulating the potential are inferred from experimental data, typically from the knowledge of the three-dimensional structure of a large number of macromolecules. The aim of KBPs is to take advantage of the availability of such data in order to estimate how much a molecular structure resembles its (unknown) native structure. Since the native structure corresponds to the minimum energy conformation of the protein, KBP usually try to describe in an effective and coarse-grained fashion all the inter-atomic interactions that stabilize the protein. Similarly to physical potential energies KBPs assign a negative (favorable) score to events that happen more frequently than expected and positive score to rarer events. The expected probability of an event is called *a priori* probability and acts as a reference for distinguishing between favorable and non favorable interactions. The set of all *a priori* probability constitutes the *reference state* of the KBP.

All KBP are based on a common theoretical framework and are specified by four elements:

- a base knowledge
- chain representation
- event definition
- reference state

The knowledge on which the potential is based is usually constituted by a (preferably large) set of protein structures. The choice of the data-set is therefore fundamental and has to be performed carefully, since different classes of proteins (globular, fibrillar, ...) are expected to exhibit different effective interactions. A good KBP must be at least robust over the change of starting data-set, i.e. the parameters estimated from two different (but equivalent) data-set of protein structures must be the same within the errors.

The macro-molecule representation is at the core of the KBP formulation. Indeed particles used to represent the structure constitute the interacting units upon which the potential will act. In case the macromolecule is a protein the natural choice for such units are the residues themselves but other choices, such as functional groups or single atoms, are equally appropriate.

The interactions between structural units are usually defined in a coarse-grained way themselves, by choosing a set of events to be associated to the interactions. Typical events are *"A particle of type α and particle of type β are found at a distance $r \pm \Delta r$"* or *"A particle of type α and particle of type β are found at a distance $r < r_0$"* or again *"A particle of type α and particle of type β are hydrogen-bonded"*, but infinite other definitions are possible. Since the set of events is often chosen to mimic different interaction that can take place between pair of particles, we introduce the concept of *interacting class* in such a way that each event can be described by the sentence *"The interaction between particle of type α and particle of type β belong to interaction class ℓ"*. Each event can therefore be identified by three indexes: the first two (α and β) identify the pair of particles while the last one (ℓ) identifies the kind of interaction that occurs between the two. It is important to notice that the choice of the interacting classes is equivalent to imposing a functional form to the potential. The definition of the interacting classes is in general not *knowledge based* but motivated by physical or chemical reasoning.

Notably an intimately dependence of the definition of the interacting classes on the coarse-grained level used for describing the structure exists. For instance if particles are composed of a large group of atoms different relative orientations of two particles could in principle lead to different physical interactions between them. In addition the loss of details induced by the coarse grained should be integrated by specialize the interaction classes.

In order to estimate the statistical interaction associated to an event it is necessary to set an expected probability $\tilde{p}^{\ell}_{\alpha,\beta}$ for that event. The choice of this reference probability is highly non trivial and largely discussed in the literature [Mue02, ZZ02, LZZZ04]. The observed probability $p^{\ell}_{\alpha,\beta}$ is instead computed from the previously selected database. The score associated to the event is finally estimated as

$$\epsilon^{\ell}_{\alpha,\beta} = -K_B T \cdot \log\left(\frac{p^{\ell}_{\alpha,\beta}}{\tilde{p}^{\ell}_{\alpha,\beta}}\right), \tag{2.1}$$

as proposed in [Sip90].

2.1.1 Solvent

One aspect of the formulation of KBPs that we have omitted in the previous section but deserves to be examined is the role of interactions between the protein and the solvent [MJ85]. Indeed even if the folding process has not been understood yet (and is still far from being understood) it is known that

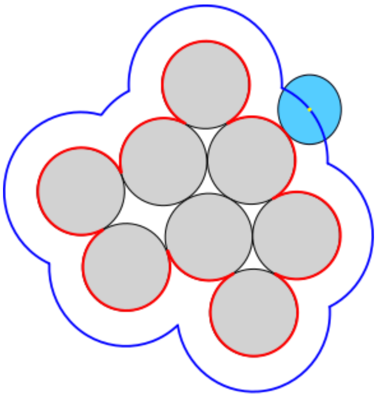


Figure 2.1: A graphical sketch of the solvent accessible surface (blue line) of a group of atoms (grey circles) compared to their van der Walls surface (red lines). The water molecule is assumed to roll around atoms: during this motion the center (yellow dot) of the water molecule (blue circle) slips onto the solvent accessible surface.

protein-solvent interactions are crucial in determining the correct fold of a protein [BB01, Jac06].

From a practical perspective modeling the interactions happening between protein and solvent is fundamentally different from the case of interaction between particles inside the protein structure.

The first challenge arises from the fact that experimental three-dimensional coordinates of solvent molecules are usually not present in pdb files. In order to overcome this problem, the solvent is often modeled as an homogeneous medium in which the protein is immersed (*implicit solvent*). However, this approach introduces new challenges, due to the continuum nature of the solvent. Indeed, since physical solvent molecules are missing it is not possible to talk about residue-solvent *interaction* but more precisely about the probability that a particle is interacting with the solvent. If this probability is zero we denote the particle as *buried*; otherwise it is *exposed to the solvent*. The degree of exposure of a particle to the solvent is measured by the area in Å^2 of the interface between the solvent and the particle itself.

We here consider Solvent Accessible Surface Area (SASA) as an approximation of the area of this interface.

The solvent accessible surface is a virtual surface that surround the protein and whose points constitute the center of a spherical probe in its maximum approach position to the protein (see Fig. 2.1). The value of SASA of a given group of atoms therefore constitute a measure of how much that group is exposed to the solvent. In last decades both analytic [FB98, HHS$^+$05, Con83] and approximated [HHS88, SM98] methods for computing SASA have been developed. Thanks to their reduced computational time, approximated methods are often preferred on the analytical ones in all those applications (MC, MD) in which SASA needs to be computed a large number of times; as usual the choice of the preferred algorithm has to consider the best compromise between reliability and computational efficiency.

As SASA is the most meaningful quantity that can be used to measure the degree of exposure of a particle to the solvent, it is not surprising that protein-solvent interaction in KBPs are often parameterized by this quantity.

2.1.2 Benchmark

KBPs are benchmarked by comparing their performances in recognizing the native state of a protein among a large set on non-native structures. During the biennial CASP competition hundreds of independent research groups try to predict the structure of a target protein by using state-of-the-art algorithms. The resulting decoys sets, composed by the native structure and by all the proposed models, show high structure similarity and therefore constitute a challenging test for KBPs [HKL09]. Although it is not trivial to definitely assess their absolute efficiency, many KBP perform quite well: Rosetta [SRK+99, TBM+03], a scoring function derived using an elegant Bayesian analysis, the composite scoring function QMEAN6 [BTS08] and the potential RF_CB_-SRS_OD introduced by Rykunov and Fiser [RF10] are particularly successful, even when tested on CASP targets.

2.2 BACH

In 2011 Cossio et al. developed a KBP called BACH [CGL+12]. This potential is constructed by analyzing a set of 500 experimentally resolved protein structures from top500 database, monitoring the probability of the event in which single residues or residue pairs are observed in different structural classes.

The BACH score function is obtained by adding together two contributions that account for residue-residue as well as for residue-solvent interactions. The residue-residue term is often denoted as *pair* contribution while the residue-solvent term is addressed as *solvation* term. The pair score is weighted with a positive parameter p that fixes the relative units of the two contributions:

$$S_{BACH} = p \times S_{PAIR} + S_{SOLV}. \tag{2.2}$$

The pairwise statistical potential is based on classifying all residue pairs within a protein structure in five different structural classes, labeled by integers from 1 to 5 in decreasing order of priority. One pair cannot be classified as belonging to one class if it is also classified in a second class with higher priority (lower label), in such a way that each pair only belongs to one class. Two residues may form a α-helical bridge (1), or an anti-parallel β-bridge (2), or a parallel β-bridge (3), or may be in contact with each other through side chain atoms (4), or may not realize any of the previous four conditions (5). A modified version of the Define Secondary Structure of Proteins (DSSP) algorithm [KS83] is used in order to detect α and β-bridges. The modified algorithm employs a more stringent energy threshold (-1 kcal mol^{-1} in place of the original -0.5 kcal mol^{-1}) to assess hydrogen bond formation. Since hydrogen atoms are often missing in experimental data, we used a simple geometrical rule borrowed from DSSP to reconstruct the position of backbone hydrogen atoms, which are required in the computation of the H-bridge energy. A residue pair is assigned to the side chain-side chain contact class if any inter-residue pair of side chain heavy atoms is found at a distance lower than 4.5 Å. If a pair of residues is not assigned to any class labeled with $\ell < 5$ it is automatically assigned to the non-interacting class.

Pairwise parameters can be stored in five symmetric matrices $\bar{\epsilon}^{\ell}$ whose entries $\epsilon^{\ell}_{\alpha,\beta}$ represent the score assigned to a pair of residue of given type α and β, whose interaction is classified as class ℓ. The number of parameters required by the pair-wise potential, once the symmetry $\epsilon^{\ell}_{\alpha,\beta} = \epsilon^{\ell}_{\beta,\alpha}$ is kept into account, are 1050.

The pairwise contribution to BACH statistical potential S_{PAIR} is computed as

$$S_{PAIR} = \sum_{i<j} \epsilon^{\ell}_{a_i,a_j}, \tag{2.3}$$

where a_i, a_j represent the type of amino acid in position i and j along the chain, and ℓ identifies the contact class between residues i and j. Notice that BACH does not score the structure of a protein but rather its contact map: small perturbations of the structure that do not affect its contact map do not affect the value of S_{PAIR} either.

Similarly, the solvation term is based on classifying all residues in two different environmental classes, either buried or solvent exposed. The environmental class is defined based on the evaluation of the SASA performed by the SURF tool of Visual Molecular Dynamics (VMD) graphic software. The SASA is computed by SURF for all heavy atoms of the protein chain by rolling a probe sphere (representing a water molecule) on the surface of the set of spheres centered at heavy atom coordinates. The radii of the probe sphere and of all atoms are set to 1.8 Å. The radius of water is higher than what is employed in VMD (1.4 Å) in order to avoid considering internal cavities as areas exposed to the solvent. The output of SURF is the number of triangle vertexes associated to each atom of the protein. These vertexes are used in the triangulated representation of the protein surface employed by VMD, and the area associated with each vertex is approximately 0.15 Å. By summing over all atoms of a given residue, the number of vertexes t associated to that residue (which is approximately proportional to its SASA) is obtained. The single residue statistical potential S_{SOL} requires two separate parameter sets, for overall 40 parameters.

The reference state used for defining BACH pairwise parameters $\epsilon^{\ell}_{\alpha,\beta}$ corresponds to the observed probability for a pair of residues to be in interaction class ℓ, without distinguishing their type α, β. Similarly the solvation reference state $\tilde{p}^{\ell}_{\alpha}$ is simply given by the probability of observing a residue in the environmental class (buried or exposed to the solvent) ℓ. In both cases the reference state is computed from the same set of proteins used for computing BACH parameters.

The parameter p introduced in Eq. (2.2) is used to tune the contribute of the pairwise scoring function with respect to the solvation part. Its value has been decided in such a way that the standard deviation of the solvation contribution to the scoring of each structure in top500 is the same as that of the pairwise contribution and results to be $p = 0.6$.

Notice that since top500 is a database of globular proteins, BACH parameters could be suitable to only describe effective interactions inside globular proteins.

Figure 2.2 shows the performance of BACH in discriminating the native state of a set of 33 decoys from CASP 8 and CASP 9 competitions [MFK+09]. For each decoy, all structures are ordered in increasing value of BACH score: the normalized score is computed as the position of the native structure (its rank) divided by the total number of structures in the decoy. The lower the normalized score, the better the performance of the KBP. Normalized ranks of all 33 decoys are subsequently ordered from the lower to the higher, as in Fig. 2.2.

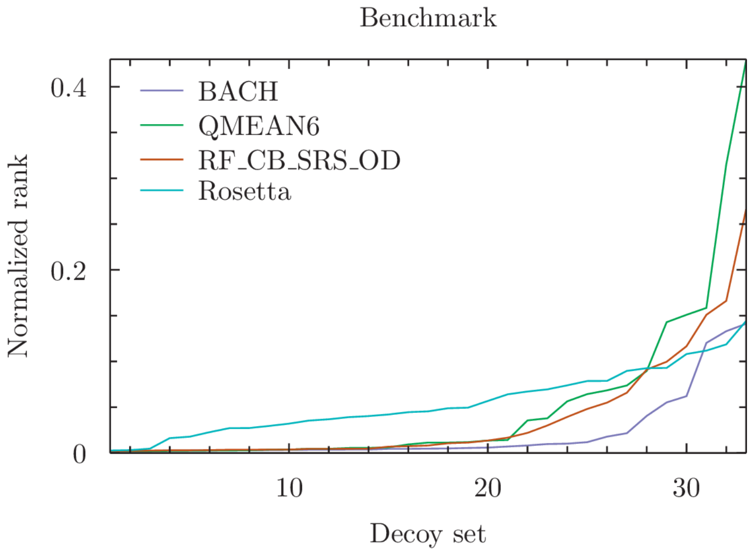


Figure 2.2: Performances of different KBPs on CASP8-9 decoys (the lower the better).

BACH has proved to outperform other state-of-the-art methods in discriminating the native state of proteins among a large set of misfolded configurations of the sequences.

Authors also showed how using the mean value of BACH score obtained during a short molecular dynamics trajectory enhanced the performance of the algorithm in those cases in which the native structure was not correctly recognized. BACH seemed to be sensitive to rather small fluctuation of proteins around their native structure: this suggests that fluctuations cannot be neglected in order to correctly determine the native state of a protein.

The simplicity of the model employed in BACH scoring function and its performances compared to other state-of-the-art knowledge-based-potentials encouraged us to develop the ideas beneath it and to devise other potentially interesting applications in the field of protein bio-physics.

2.3 Enhancing BACH

The modifications we propose to enhance the performances of BACH are focused on the development of new algorithms for determining the environmental class of each residue. Indeed there are three main disadvantages in the usage of the rolling-ball algorithm implemented in SURF:

- the usage of the external routine results in a limited ability to modify the rolling-ball algorithm as needed

- the computation of SASA is not efficient since the routine can not be perfectly integrated into the BACH source-code
- the rolling-ball algorithm itself is relatively slow

For these reasons we implemented and tested two different algorithms that are able to determining if a group of atoms is buried or exposed to the solvent. The first algorithm we present estimates SASA by using a modified version of the LCPO algorithm [WSS99]. The algorithm presented in Section 2.3.2, on the contrary, determines if a particle is exposed to the solvent without computing its SASA and therefore saving computational time.

2.3.1 Linear Combination of Pairwise Overlaps

We here propose to assign a residue to an environmental class by estimating its SASA in a simpler manner, following the LCPO algorithm presented in [WSS99]. This approach approximates the accessible surface of a solute as a linear combination of the surfaces of its atoms, modeled as spheres of radius r. The working principle is to remove from the sum of the whole surface contribution of each atom the estimated overlap of the surfaces of nearby atoms. The exposed surface of the atom i is approximated as follows:

$$\begin{aligned} A_i \;=\; & P_1 4\pi r_i^2 + P_2 \sum_{j\in N(i)} A_{ij} + P_3 \sum_{j,k\in N(i) k\in N(j) k\neq j} A_{jk} + \\ & + P_4 \sum_{j\in N(i)} A_{ij} \left(\sum_{k\in N(i) k\in N(j) k\neq j} A_{jk} \right) , \end{aligned} \tag{2.4}$$

where

$$A_{ij} = 2\pi r_i \left(r_i - \frac{d_{ij}}{2} - \frac{r_i^2 - r_j^2}{2d_{ij}} \right) . \tag{2.5}$$

The quantity r_i is the radius of atom i, $N(i)$ stands for the list of atoms that overlap with atom i and d_{ij} is the center-to-center distance of atom i and j. In the original work [WSS99] the four parameters P_1-P_4 depended on the hybridization of the atom and on its neighborhood and were estimated by linear regression of a heterogeneous database of analytically calculated cases.

In order to simplify the implementation of the method, we decided to attempt a coarse approximation, using for all the heavy atoms of the protein a single value of the radius r_{LCPO}, and single set of parameters P_i, the one

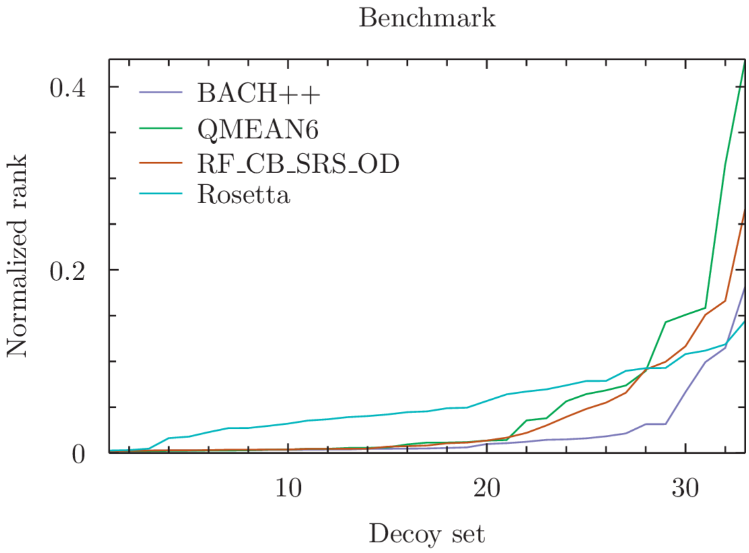


Figure 2.3: BACH++ benchmark on 33 decoy sets from CASP 8/9 competition; the performance is comparable to that of other state-of-the-art methods.

introduced in [WSS99] for a sp_3 carbon bound with three heavy atoms. Since the parameters P_2 and P_3 are negative the value of A_i can also be negative. By visual inspection, we checked that this normally happens when the atom is deeply buried in the protein core. We therefore assumed that a residue is solvent exposed if the sum of the values A_i of its atoms is larger than zero.

The only free parameter in the functional form (2.4) is thus the radius of the heavy atoms. To determine it, we defined a *coherence score* as the fraction of residues over all the top500 structures for which SURF and the LCPO algorithm agree on the environmental class assignation (b or e). We then found the radius parameter that maximizes the value of this score. We reached a maximum 87% overall accordance for the value of $r_{LCPO} = 3.09$ Å. This value roughly corresponds to the sum of the Van der Waals radius of an aliphatic carbon (1.6 Å) and the radius of a water molecule (1.4 Å). We named this upgraded version of BACH as BACH++.

We benchmark this functional form of S_{solv} by the same procedure followed in [CGL$^+$12]. In particular, we consider a selection of 33 decoy sets from CASP 8/9 competition [MFK$^+$09]. For all the decoy sets we first compute the value of S_{sol} for all the decoy structures and for the native structure. We define the normalized ranking as the rank in the sorted list of the native structure, divided by the number of structures of the decoy set. The smaller this number, the better the solvation function discriminated the native among the decoys. In Fig. 2.3 the rankings of BACH and BACH++ are compared to that of other

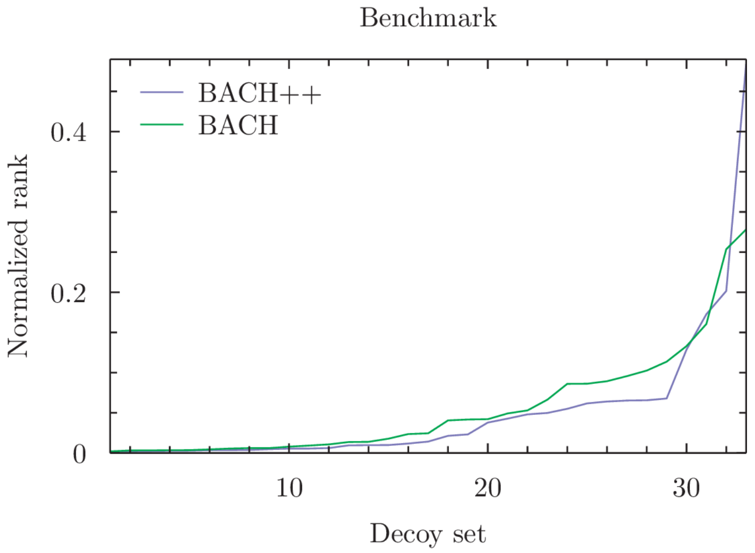


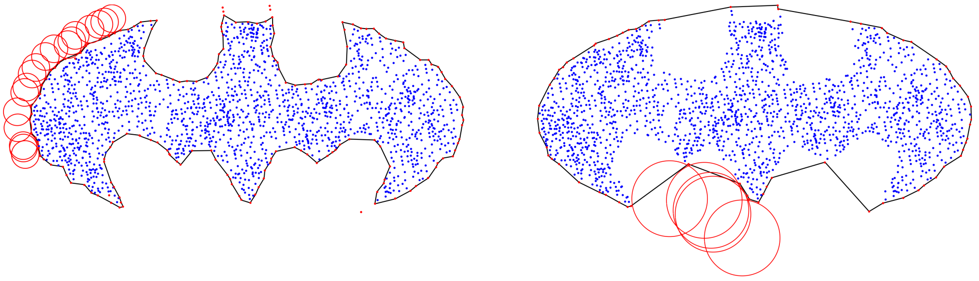
Figure 2.4: Solvation term contribution to BACH++: benchmark on 33 decoy sets from CASP 8/9 competition; the performance is comparable to that of BACH.

state-of-the-art methods. In Fig. 2.4 the sole contribution of the solvation term is instead considered. We can see that this new LCPO-based procedure guarantees a slightly more accurate prediction of the native structure with respect to BACH.

2.3.2 Alpha shapes

As seen in Section 2.1.1 in order to classify a group of atoms as buried or exposed to the solvent the value of its SASA $\mathcal{S}$ is often used . The classification is decided by comparing the surface area to some pre-determined and fixed threshold value $\mathcal{S}^\star$: if $\mathcal{S}$ exceeded $\mathcal{S}^\star$ the group of atoms is considered exposed to the solvent, otherwise it is considered buried.
In this section we will present a technique that greatly enhances this procedure from a computational efficiency perspective and produces, at the same time, a more exact classification.

The main problem in estimating SASA is computational time. Indeed existing algorithms (see previous section on LCPO) need to iterate on every pair of atoms belonging to the structure and, as a consequence, the time required is proportional to the square of the number of atoms considered.
Fortunately it is not necessary to compute SASA of a group of atoms in order to determine if it is located on the protein surface: the information stored in the sentence “SASA is $\mathcal{S}$” is much more than the information required by



(a) Appropriate probe radius.

(b) Large probe radius

Figure 2.5: Two different alpha-shapes (black lines) of a random set of points (blue) in the plane. In panel (b) the larger radius of the probe (red circles) is responsible for the loss of details of the reconstructed shape.

BACH, which can be summarized in the sentence “Is SASA greater than $\mathcal{S}^{\star}$?”. Computing the surface area requires a great amount of computational resources and time that can be saved by an algorithm devised to compute just the right amount of information needed.

The algorithm we proposed computes the *shape* of the desired protein or protein complex and manage to classify every atom of the structure as belonging to the surface in sub-quadratic time.

The shape of an object is a vague notion and there are probably many possible interpretations; among these we utilize that of α-shape. The concept of α-shape is relatively new [EKS83] but is nowadays largely used for shape reconstruction in many different fields like computer sensing [FM09, BB97], topography [VKMT10], and biology [BNL03, TAD06]. Alpha-shapes are piecewise linear surfaces made up of triangles. A triangle is drawn for all triplets in the input set that sit on the boundary of a probe sphere of radius $\alpha^{\frac{1}{2}}$ that contain no other points in the input set. The radius of the probe sphere plays a fundamental role in the accuracy of the algorithm and need to be decided with meticulous attention: indeed if a too large value is used most of the details of the reconstructed surface would be lost. On the contrary the usage of a probe that is too small if compared to the typical distance between points will result in a disconnected shape. Notice that the alpha shape obtained with a infinitely large probe ($\alpha = \infty$) corresponds to the convex hull of the original set of points. Figure 2.5 shows pictorially how the reconstructed shape can depend on the choice of probe radius.

The algorithm for reconstructing the protein surface uses the α-shape implementation of the Computational Geometry Algorithms Library (CGAL) library [cga]. The set of input points that is used for determining the shape

of the macromolecule is formed by the position of every heavy atom of the structure; as usual hydrogen atoms are not considered because information about their position is often missing from experimental data. In the context of protein surface reconstruction the probe sphere used by the algorithm mimic the presence of a solvent molecule and therefore its size should reflect the steric effects between (heavy) atoms of the structure and solvent molecules. The solvent is assumed to be water and so we expect the correct radius to be greater than the typical dimension of an H_2O molecule, i.e. 1.4 Å.
In order to determine the best probe radius value we studied how the classification of residues of a set of 50 different globular proteins randomly chosen among those of the top500 database changes as a function of the probe radius. We considered the *exact* classification to be the one obtained using an analytical method for computing SASA [FB98] and by setting a threshold value $\mathcal{S}^{\star} = 0$. The probe radius was therefore chosen in such a way that the error committed in classifying residues as buried or exposed to the solvent is minimized. Figure 2.6 shows how the relative error varies as a function of the probe radius: the minimum is obtained for a probe radius of $r_p = 3.2 \pm 0.01$ Å that, as expected, is greater than the typical value used as radius of a water molecule ($r_{H2O} = 1.4$ Å). The larger effective radius of the solvent molecule accounts for the steric effects between the solvent and structure heavy atoms as well for the steric effects due to the presence of hydrogen atoms, that are not considered in the model.

We also compared the error obtained by using the α-shape algorithm with the one obtained by using the LCPO one.
We recall that LCPO algorithm estimates the SASA of a group of atoms as a linear combination of terms composed by pairwise overlaps of hard spheres of fixed radius r_{LCPO} and centered on heavy atoms. Notice that the sphere used by the LCPO algorithm has the same physical interpretation than the probe used in the α-shape method, its size having to mimic the steric effects of a water molecule and atoms inside proteins. The same Fig. 2.6 also shows the relative error obtained in the classification of residues by the modified version of LCPO as a function of r_{LCPO}. We noticed that the minimum error obtained with LCPO is *twice* the error produced by the α-shape algorithm. Moreover the fact that the width of the minimum obtained with α-shape is considerably larger than the one obtained with LCPO indicates that the classification is more stable, and therefore reliable, in the former case.

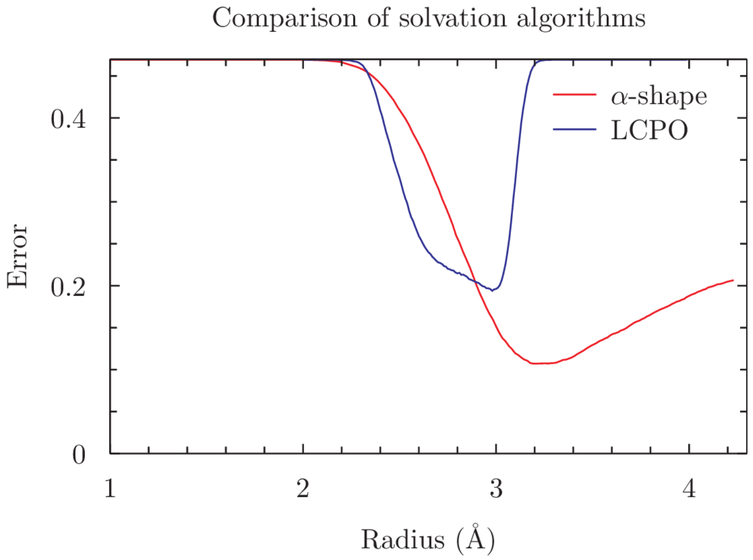


Figure 2.6: Error committed in the classification of residues as buried or exposed according to LCPO (blue line) and *alpha*-shape (red line) algorithms.

The improved performances of the α-shape based algorithm allow the application of the same technique for the classification of even smaller sets of atoms, for which the classification proposed by LCPO method becomes unreliable.

The α-shape algorithm implemented in CGAL also computes the tetrahedralization of the space induced by the input points, that is a subdivision of the protein inner space obtained by connecting near atoms, similarly to what is done by the Delaunay tetrahedralization algorithm. This space partition can be used to determine for each atom a list of neighboring atoms: the corresponding graph describing the vicinity relation between input points can be used during the computation of the pairwise scoring function for limiting the search of interacting residues to those connected in the graph. In this way the computation of the pairwise potential only require a time that grows linearly with the protein size.

Chapter 3

Estimating binding affinities

Protein–protein interactions are involved in almost all biological processes [JT96, NT03, JBC08] and the comprehension of their thermodynamical and structural properties is of paramount importance to gain a quantitative understanding of biological function and to enable the design of proteins, small molecules and other compounds to modulate their strength [AW04, ZC05]. Moreover, the characterization of protein interaction networks is the main goal of functional genomics in order to understand the complex relationship between genotype and phenotype on a global, genome-wide scale [HHL+08, KZV10, ZPNH10].

A crucial step to get a deeper insight in the protein-protein interaction mechanisms is the calculation of protein-protein binding affinities, e.g. a measure of the capability for a pair of proteins of forming a stable complex [AR06, Del09, KMH+11, MAB11]. For a binary complex, the binding affinity translates in physical–chemical terms into an equilibrium dissociation constant (K_d), which may be measured at equilibrium or derived from the reaction kinetics, and the related Gibbs binding free energy ΔG. Determining computationally the binding affinity starting from the knowledge of the structure of the complex and of the two unbound sub-units alone, in the case of non-obligatory complexes, is therefore an important benchmark for our understanding of the determinants which are driving the protein binding and it can be also a valuable tool for many efforts in structural bioinformatics ranging from the design of peptides for therapeutic purposes [KVS+10] to protein engineering [SYDS11].

Given its huge practical relevance, the construction of accurate and efficient binding free energy functions attracted a large variety of studies based on somewhat different methods in terms of physical plausibility, accuracy and computational cost. These methods range from highly accurate but numerically costly approaches such as thermodinamic integration [Kol93] and Molecular Mechanics/Poisson–Boltzmann Surface Area (MM/PBSA) methodologies [GZ07] to empiric energy functions, mainly based on statistical potentials, where a statistical analysis of the relative position of residues observed in experimental structures is used to infer a potential of mean force. While the former strategies use extensive molecular dynamics sampling and are usually only applicable when the bound and unbound state states have a significant overlap, the latter are much more flexible and their performances can be boosted by optimizing relative weights with linear regression against known binding affinities [BYY+11].

The strong limitations in data sets available have been a major drawback both for training and for benchmarking the quality of numerical predictors of binding affinities. Such problems have been overcame by a databank recently obtained [KMH+11] by collecting the largest and most diverse set of experimental binding free energies to date, covering 144 non-redundant interactions, with structural cross-referencing to both the bound complex and its unbound constituents. While in the past all the methods have been parameterized and/or evaluated on a narrow range of proteins, this new databank, which cover several order of magnitude of K_d, provides an unprecented opportunity for the construction and evaluation of empirical binding free energy models. By using this databank, high accuracy predictions previously obtained[HL92, JGM+02, MWLZC02, AS07, ZL08, SZX+09, BYY+11] turned out to be an artifact of the unreliability of former data sets. New methods have then been introduced [MAB11, VHPW12] but significant improvement over previous empirical free energies has only been obtained throughout very complicated consensus energy functions based on up to 200 molecular descriptors [MAB11].

In this chapter we want to study the problem by taking into account the entropic terms due to the fluctuations of the structures which can be significantly different between the unbound states and the complex. In our approach the physical processes governing the association of proteins, such as van der Waals interactions, electrostatic interactions, hydrogen bonding and solvation are modeled by using the BACH-SixthSense energy function. This potential, that is based on the hypothesis [SGS+15] that protein complexes are stabilized by the same fundamental interactions of monomeric proteins, works equally well in discriminating correctly the native structure among a competing set of decoys for both monomeric proteins and protein dimers. SixthSense is grounded on the BACH knowledge-based potential [CGL+12, SZC+13], where a pairwise contact potential, based on 5 different contact classes, is combined with a one-body solvation potential and a term introduced to consider the effect of steric clashes. However, as expected, the mere application of SixthSense to compute binding affinities show a low correlation coefficient with experimental measurements. This correlation dramatically increased as soon as fluctuations of the structures around their native conformations are taken into account. This can be done by an accurate matching between the vibrations modes, as computed with a coarse grained elastic network model and with short molecular dynamic simulations run from the native three dimensional structures of the complex and of the unbound sub-units. With this setup our approach give a level of accuracy of the same quality of most up to date methods underlying the relevant role of fluctuations in estimating bind affinities of not rigid bodies.

3.1 Contributions to binding Free-Energy

As seen in Section 2.2 fluctuations of the structure around its native conformation are critical in the scoring process. We expect the role of fluctuations to be particularly evident in the case of protein-protein interaction, as the space available to the monomers for their fluctuation to occur shrinks upon binding. Unfortunately the contribution of fluctuations to the score of a structure cannot be simply computed from the molecular structure but instead relies on the ability of generating a molecular dynamics trajectory starting from the native conformation. As molecular dynamics requires a set of preliminary settings that can not be automatized it results to be time consuming and, above all, delicate.

We devised a procedure that allow to estimate the ΔG of binding of protein structures in water with an accuracy comparable to that of other state-of-the-art methods. The range of free-energy differences on which the proposed procedure has been tested spans 12 orders of magnitude, corresponding to binding free energies from $-18\ \mathrm{kcal} \cdot \mathrm{mol}^{-1}$ to $-6\ \mathrm{kcal} \cdot \mathrm{mol}^{-1}$.

In order to compare our results with that of other methods we employed a largely used database constituted by non-obligatory complexes for which the structures of its unbound monomers as well as their binding affinity is available [KMH+11, MAB11]. A large fraction of the structures in this dataset is not complete, missing some heavy atoms or some residues. Often missing residues in the monomers are instead present in the complex structure, or *vice versa*. In these cases we deliberately delete those residues and only deal with *maximum common structures* in which all residues are participating to both the monomer and the complex.

We therefore selected those complexes that:

- have no missing heavy atoms (but there can be whole residues missing)
- the contribution of the residues of the interface to the SixthSense score do not change after the maximum-common-structure procedure.

Among the 144 complexes available in the original database only 74 satisfy these requests.

3.1.1 Thermodynamics

We assume that the partition function Q of one monomer can be written as

$$Q = q_i \cdot q_t \cdot q_r \cdot q_v, \tag{3.1}$$

where q_i denotes the contribution of intra-molecular interactions, q_t denotes the translational contribution, q_r the rotational contribution and finally q_v the vibrational contribution. Notice that we neglect other major possible contributions such as, for instance, solvation effects. In Eq. (3.1) we also assume that each of the terms in the r.h.s. of Eq. (3.1) is independent from each other.

Schematically, in the binding process, two molecules A and B associates into the complex C. The free-energy lost in the binding process can be easily computed by applying Eq. (3.1) to the complex C and to both the monomers A and B

$$\Delta G = -K_B T \log\left(\frac{Q^C}{Q^A\, Q^B}\right) \tag{3.2}$$

$$= \Delta G_i + \Delta G_t + \Delta G_r + \Delta G_v, \tag{3.3}$$

where ΔG_i corresponds to the difference in free energy due to the intra-molecular contributions and similarly ΔG_t, ΔG_r and ΔG_v are due to translational, rotational and vibrational contributions, respectively.

The translational and rotational contributions can be exactly computed from the three-dimensional structure of the monomers and of the complex:

$$\Delta G_t = - K_B T \log \left(\frac{h^3}{(2\pi K_B T)^{\frac{3}{2}}} \frac{V_C m_C^{\frac{3}{2}}}{V_A m_A^{\frac{3}{2}} V_B m_B^{\frac{3}{2}}} \right) \tag{3.4}$$

$$\Delta G_r = - K_B T \log \left(\frac{1}{8\pi^{\frac{1}{2}}} \frac{h^3}{(2\pi K_B T)^{\frac{3}{2}}} \frac{(I_x^C \cdot I_y^C \cdot I_z^C)^{\frac{1}{2}}}{(I_x^A \cdot I_y^A \cdot I_z^A)^{\frac{1}{2}} (I_x^B \cdot I_y^B \cdot I_z^B)^{\frac{1}{2}}} \right), \tag{3.5}$$

where h is the Plank's constant and I_x, I_y and I_z denote the principal moments of inertia for rotation about the three orthogonal axes and m_A, m_B and m_C the masses of both the monomer and of the complex ($m_C = m_A + m_B$).

The remaining terms cannot be deduced in a trivial way from the three-dimensional structures of the proteins but need to be estimated with some simplified models.

We recall that the *binding affinity* K_d is defined in term of the binding free energy as

$$K_d = c_0 \exp \left(\frac{\Delta G}{K_B T} \right), \tag{3.6}$$

where $c_0 = 1 \text{ mol} \cdot \text{L}^{-1}$ is a reference concentration.

3.1.2 Vibrations

Although a quantitative analysis of vibrations requires structural information and a detailed potential energy function, we can get a good qualitative understanding by means of a simplified approach.

We estimate the vibrational contribution ΔG_v by describing the unbound monomers and the complex with coarse-grained anisotropic network models [ADJ+01]. In this model we depict the protein chain as a network of ideal springs connecting the α-carbons of different residues whose distance is lower than 12 Å. The elastic constant of each spring can have two different values, depending on the relative position of the considered residues: if the residues are consecutive along the chain (i.e. they are connected with a peptide bond) an large constant k_{pep} is used; if, on the contrary, the two residues are not consecutive a smaller elastic constant k_{int} is used. In order to mimic the rigidity of the peptide bonds the constants are chosen in such a way that

$$\frac{k_{pep}}{k_{int}} = 10. \tag{3.7}$$

In case that the coordinates of the α-carbon of one or more residues are missing from experimental data, no special care is taken and the chain is left disconnected. The same strategy is followed in MD simulations. As seen in Chapter 1, anisotropic network models can be used to predict both the mobility profile of a molecular system and other thermodynamic quantities, such as the partition function. We determined the most appropriate value of k_{int} by comparing the mobility profile predicted by the anisotropic network model with the one computed from an MD trajectory in explicit solvent and we utilized the obtained value for estimating the vibrational contribution to the binding free-energy.

Molecular Dynamics

A MD simulation has been performed for each of the input structures, i.e. 57 complexes and the corresponding unbound monomers, by using the GROMACS 4.5.4 package [HKVDSL08, PPS+13]. All MD simulations are based on the AMBER99SB force field and on the TIP3P water model.

The hydrogen atoms of the proteins have been replaced (or added, if missing in the input file) employing H++ automated system through its web interface (version 3.1) [GMF+05, AAO12]. The protonization process has been performed according to the pH value and salt concentration specified in the Supplementary Material of [KMH+11] while the values of protein dielectric constant and solvent dielectric constant were 10 and 80 respectively. H++ automated system also assigned the most favorable protonation states of histidine residues, as well as the favorable orientations of glutamine and asparagine side-chains.

An octahedron box was used for delimiting the system: the protein was centered into the octahedron, whose size was chosen in such a way that the minimal distance between the solute and the box was 10 Å. Water molecules were added to the system and Na and Cl atoms were inserted in accordance with the value of the experimental salt concentration.

In order to fix steric clashes between atoms of the protein and water molecules, we employed an energy minimization procedure based on the steepest descent algorithm. Finally the system was gradually heated from 0 to 300 K by using a velocity-rescaling thermostat with different time-coupling constants for protein and water/ion molecules.

After the preparation part described above we performed 1 ns-long equilibration runs by using a time-step of 1 fs. The final runs, during which the

productive trajectories were collected, were instead 5 ns-long and employed a time-step of 2 fs. As a result a total of 2500 different configurations belonging to the native state have been generated for each input structures.
These configurations have been clustered by using the gmx built-in clustering utility in GROMACS.

Estimating the vibrational contribution

The mobility of each residue is extracted from the MD trajectory as the mean square deviation of the position of any alpha carbon around a reference structure, which is taken to be the center of the most populated cluster, as determined by the gmx utility built-in in GROMACS. The mobility profile μ^{md} we obtained from the MD simulation is used as a reference mobility to determine the best value of the elastic constant employed in the anisotropic network model. Usually the mobility profile obtained from the network model is, to some extent, proportional to the one computed from the MD simulation: low mobility and high mobility areas are often recognized by both approaches with a good accordance. Nonetheless the two profiles often appear to be globally rescaled one with respect to the other and moreover some regions exhibit very different relative mobilities.

As a starting point we fixed the elastic constant $k_{int} = 1$ N/Å and we compute, as a N-component vector, the corresponding mobility profile $\boldsymbol{\mu}^0_{anm}$

$$\boldsymbol{\mu}^0_{anm} = \frac{1}{3} \sum_{i>6} \frac{K_B T}{\lambda_i} \left(\boldsymbol{u}_{x,i} + \boldsymbol{u}_{y,i} + \boldsymbol{u}_{z,i} \right), \tag{3.8}$$

where λ_i is the i^{th} eigenvalue and $\boldsymbol{u}_{x,i}, \boldsymbol{u}_{y,i}, \boldsymbol{u}_{z,i}$ are N-component vectors that together form the corresponding $3N$-eigenvector, as obtained from solving the eigenvalue problem described in Section 1.5. Notice that the summation is extended to all oscillating modes except the first 6 ones, that correspond to rigid roto-translations. Their contribution can not be estimated with an ANM model and has already been taken into account in Eqs. (3.4) and (3.5). The vibrational contribution to the free energy that corresponds to this initial choice of k_{int} is given by the following relation

$$\hat{G}^0_v = K_B T \sum_{i>6} \log \left(\frac{\hbar\, \lambda_i^{\frac{1}{2}}}{e\, K_B T} \right). \tag{3.9}$$

Modifications of the spring constant correspond to a rescaling of this initial profile: we therefore try to match the MD and ANM profiles by mean of a global rescaling of the latter one

$$\boldsymbol{\mu}_{anm} = \mathcal{C}\boldsymbol{\mu}^0_{anm} \tag{3.10}$$

by computing the rescaling factor $\mathcal{C}$ that minimize their Mean Square Error (MSE). This can be done exactly by employing some estimators like, for instance, Ordinary Least Squares (OLS). But in order to eliminate the contribution of those regions that can not be matched (usually loops and chain ends) we decided to proceed by using an iterative procedure that allows us to automatically recognize and discard those undesired regions.

Within this procedure, on every step $k > 1$ the scaling constant $\mathcal{C}^k$ is computed by minimizing the sum of squared residuals as

$$\mathcal{C}^k = \frac{\boldsymbol{\mu}^k_{md} \cdot \boldsymbol{\mu}^k_{anm}}{\boldsymbol{\mu}^k_{md} \cdot \boldsymbol{\mu}^k_{md}}, \tag{3.11}$$

where $\boldsymbol{\mu}^k$ is the mobility of all those residues that have not been excluded during the preceding steps. Once the constant $\mathcal{C}^k$ is computed, the residue whose mobility exhibits the biggest difference between the two profiles is excluded and not considered in the successive steps. The procedure ends when the MSE between the profiles is smaller than $\Delta\mu = 4$ Å^2 and $\mathcal{C}$ is finally set to be the last computed $\mathcal{C}^k$. Even if there are cases in which most of the residues are excluded from the matching procedure, as in Fig. 3.1, it can be seen that the final mobility profile has been appropriately rescaled and therefore the factors $\mathcal{C}$ can be safely used for estimating the vibrational contribution ΔG_v. The correction to the binding free energy due to the introduction of these rescaling factors is given by

$$-\frac{K_B T}{2} \log\left(\frac{\mathcal{C}_C^{3N_C-6}}{\mathcal{C}_A^{3N_A-6}\,\mathcal{C}_B^{3N_B-6}}\right), \tag{3.12}$$

where $\mathcal{C}_A$, $\mathcal{C}_B$ and $\mathcal{C}_C$ are the rescaling constants obtained by minimizing the MSE of the mobility profiles of both the unbound monomers A, B and the complex C. The final estimation

$$\Delta\hat{G}_v = -\frac{K_B T}{2} \log\left(\frac{\mathcal{C}_C^{3N_C-6}}{\mathcal{C}_A^{3N_A-6}\,\mathcal{C}_B^{3N_B-6}}\right) + \Delta\hat{G}_v^0 \tag{3.13}$$

is obtained by summing the estimation initially obtained by fixing the elastic constant to $k^0_{int} = 1\ \mathrm{N} \cdot \mathrm{\AA}^{-1}$ (given in Eq. (3.9)) with the correction computed in Eq. (3.12). The missing contribution of the 6 roto-translational modes is taken into account by the translational and rotational terms in Eqs. (3.4) and (3.5). Notice that this interpretation is correct only if the number of residues of the complex is equal to the sum of the number of residues of its monomers.

3.1.3 Intra-molecular contribution

The remaining term in Eq. (3.3) depends on the intra-molecular interactions and is given by

$$\Delta G_{int} = -K_B T \log \left(\frac{q^C_{int}}{q^A_{int} q^B_{int}} \right). \tag{3.14}$$

We decided to estimate this term by using the modified version of the BACH knowledge based potential presented in [SGS$^+$15]. This version, named *SixthSense* employs a more refined definition of side-chain interaction between residues and distinguishes between cases in which an interaction takes place between non-polar groups from cases in which at least one of the groups involved in the interaction is polar. Following the approach already employed in [SGS$^+$15], we only considered the contribution to the SixthSense score due to those residues that, in the complex, belong to the interface.
The interface between the two bound monomers is defined as the set of residues that in the complex interact (according to BACH) with at least one residue belonging to the other monomer.

The ΔG_{int} is assumed to be proportional to the difference between the SixthSense score of the complex $S_{BACH}(AB_{bound})$ and that of the two unbound monomers $S_{BACH}(AB_{unbound})$ and is therefore estimated as

$$\Delta \hat{G}_{int} = S_{BACH}(AB_{bound}) - S_{BACH}(AB_{unbound}). \tag{3.15}$$

Notice that residue pairs from different monomers that interact at the complex interface will be assigned the non-interacting SixthSense score in the unbound state.

Complex		Monomer 1		Monomer 2	
pdb id	chains	pdb id	chains	pdb id	chains
1AK4	A:D	2CPL	A	1E6J	P
1BVN	P:T	1HOE	A	1PIG	A
1DFJ	E:I	2BNH	A	9RSA	B
1DQJ	AB:C	1DQQ	CD	3LZT	A
1GL1	A:I	1PMC	A	4CHA	ABC
1GPW	A:B	1K9V	F	1THF	D
1MAH	A:F	1FSC	A	1J06	B
1MLC	AB:E	1MLB	AB	3LZT	A
1PPE	E:I	1LU0	A	2PTN	A
2PTC	E:I	2PTN	A	9PTI	A
2SIC	E:I	1SUP	A	3SSI	A
3SGB	E:I	2QA9	E	2OVO	A

Table 3.1: Training set used for determining parameters β_0, β_1 and β_2 in Eq. (3.16). For each pdb code the list of chains that constitute the complex are provided.

3.1.4 All together

The total binding free energy is estimated by using the following linear combination of the contributions computed in the previous sections:

$$\Delta\hat{G} = \beta_0 + \beta_1 \Delta\hat{G}_i + \beta_2 \Delta\hat{G}_{vtr}, \tag{3.16}$$

where $\Delta\hat{G}_{vtr} \equiv \Delta\hat{G}_v + \Delta G_t + \Delta G_r$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ are scalar parameters we introduced in order to take into account for the presence of systematic errors in the estimation of ΔG_i and ΔG_v as well as for the role of other contributions. The parameters have been computed by minimizing the mean square residuals between $\Delta\hat{G}$ and ΔG, by mean of an OLS regression, on a set of 12 complexes that we used as *training* set. The training set comprises 12 complexes and has been chosen among the structures constituting the training set in [MAB11] in such a way to maximize the range of experimental binding affinities. It is described in Table 3.1. Other 15 complexes (listed in Table 3.2) have been used as test set and have been used to benchmark the accuracy of the method proposed. The test set has been chosen among the structures constituting the test set in [MAB11], with the same criteria followed for the training set.

Complex		Monomer 1		Monomer 2	
pdb id	chains	pdb id	chains	pdb id	chains
1ACB	E:I	4CHA	ABC	1EGL	A
1B6C	A:B	1D6O	A	1IAS	A
1BVK	DE:F	1BVL	BA	3LZT	A
1CBW	ABC:D	4CHA	ABC	9PTI	A
1E6J	HL:P	1E6O	HL	1A43	A
1EAW	A:B	1EAX	A	9PTI	A
1F6M	A:C	1CL0	A	2TIR	A
1FSK	BC:A	1FSK	BC	1BV1	A
1JIW	P:I	1AKL	A	2RN4	A
1RV6	VW:X	1FZV	AB	1QSZ	A
2AQ3	A:B	1BEC	A	1CK1	A
2I25	N:L	2I24	N	3LZT	A
2JEL	HL:P	1POH	A	2JEL	HL
2PCC	A:B	1CCP	A	1YCC	A
2TGP	Z:I	1TGB	A	9PTI	A

Table 3.2: Test set used to benchmark the accuracy of the method. For each pdb code the list of chains that constitute the complex are provided.

3.2 Results

We start by showing a practical example of how the matching between the mobility profiles estimated by MD and ANM allows to obtain the scaling constant $\mathcal{C}$ (by employing Eq. (3.11)) and consequently the correct elastic constant

$$k_{int} = k^0_{int}/\mathcal{C}\,; \tag{3.17}$$

as introduced in Section 3.1, $k^0_{int} = 1\ \mathrm{N}\cdot\text{Å}^{-1}$ is the initial estimation of the elastic constant used to compute the residue mobility profile with ANM (see Eqs. (3.8) and (3.9)). It is important to observe that by using the above procedure we get different elastic constants for different structures. Therefore, when considering a complex and the two unbound subunits, we compute three different elastic constants.

Figure 3.1 graphically depicts the matching procedure, as it happens for the complex 1BVK and its unbound subunits 1BVL and 3LZT. The residues excluded from the rescaling procedure are signaled by grey bars, to highlight the fact that such regions are most typically excluded because the ANM grossly underestimations their MD mobility. Our interpretation is that ANM fails to

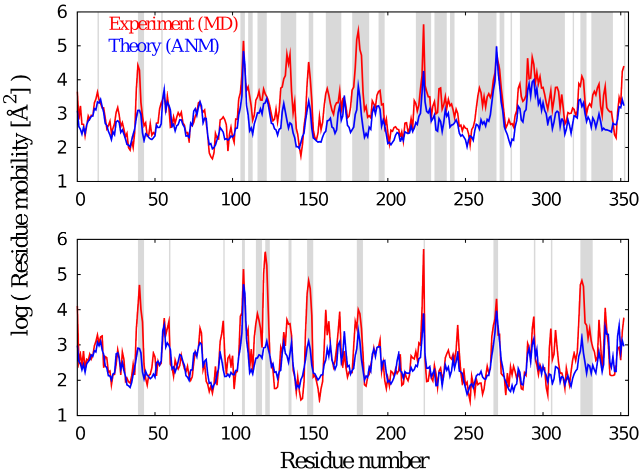


Figure 3.1: Best matching between the residue mobility profile obtained from the MD trajectories and the one estimated from the coarse-grained elastic network model after rescaling, as resulting from the iterative procedure of excluding residues whose MSE differs more than 4 Å^2. The logarithm of the mobility is shown on the vertical axis, in order to better appreciate the similarities between profiles after rescaling. Upper panel: the MD (red) and ANM (blue) residue mobility profiles for the complex 1BVK (residues covered with the vertical grey lines are excluded from the matching, and the remaining ones are used to determine the scaling constant (see Eq. (3.17))). Lower panel: the same as in the upper panel for the two unbound subunits 1BVL and 3LZT; the two sequences are pasted to allow a direct comparison with the complex 1BVK.

correctly describe those regions because their behavior is beyond the linear elastic regime. Therefore, since we want to use the results obtained by an ANM approximation to estimate the vibrational entropy, we deliberately exclude those regions from the rescaling procedure, even if they can eat up a good fraction of the whole sequence (see the upper panel of Fig. 3.1).

In order to appreciate the crucial role of a correct estimation of the vibrational entropy contribution to the binding free energy, we proceed step by step. First we use only the BACH KBP score of the interface of the protein complex as an estimate of the binding free energy.

The results are shown in Fig. 3.2, together with the corresponding Pearson linear correlation coefficient obtained for three choices of the data set: (i) 74 complexes chosen among those used in [MAB11] as described in Section 3.1, (ii) the 12 complexes that we will use in the following as our training set (red points in the plot), and (iii) the 27 complexes that include both the training and the test set used in the following (red and blue points in the plot). Both the training and the test sets are subsets of the larger corresponding sets employed in [MAB11]. The full set of 27 complexes that we will consider in the following

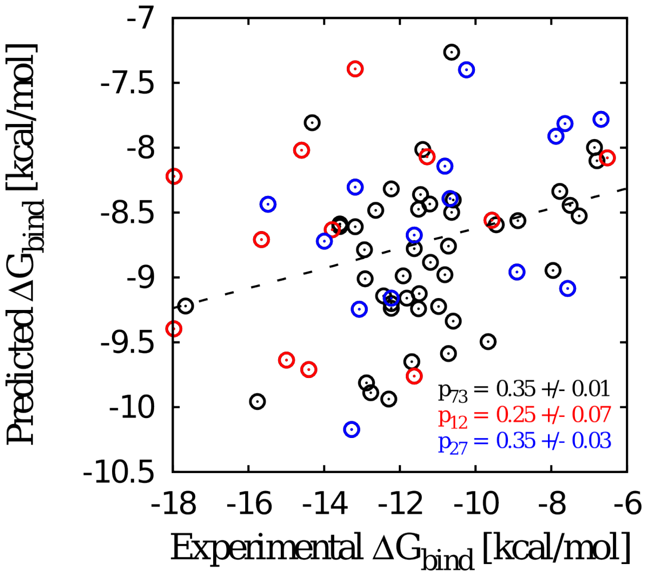


Figure 3.2: Correlation between the experimental values of the binding free energy ΔG and the estimate given by the BACH interface score. All the data points (colored in black, red and blue) correspond to 74 complexes, that are the subset of the benchmark data set of Kastritis *et al.* Ref. [KMH+11] for which no heavy atom was missing in the pdb file and for which the value of the BACH interface score remained unchanged upon the usage of “maximum similarity” protein models, instead of the original pdb files. The data shown in red correspond to 12 complexes that belong to our training set, while the data colored in blue correspond to the 15 complexes that form our test set. The Pearson correlation coefficient is shown for the 74, 12, and $27 = 12 + 15$ sets. The dashed black line is the linear interpolation for all 74 points.

was chosen to span all possible experimental values of binding affinity, with a deliberate preference for the worst outliers in Fig. 3.2.
A mild correlation appears, notably even less evident for the training set that corresponds to the “validated” complexes, for which the experimental measure of the binding affinity should be more reliable, according to [MAB11].

We therefore tried to enhance this mild correlation by introducing entropic corrections. In the following, we did not simply verify the linear correlation between experimental values and theoretical estimations, but also tried to predict quantitatively the binding free energy of different complexes. Along this line, we introduced a set of three scalar parameters ($\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$) that enter our theoretical estimations. These parameters have been optimized by minimizing the Root Mean Square Error (RMSE) between the theoretical estimations $\Delta\hat{G}$ and the corresponding experimental values ΔG on the training set of complexes, and the performance of the method has been verified on the full (training and test) set.

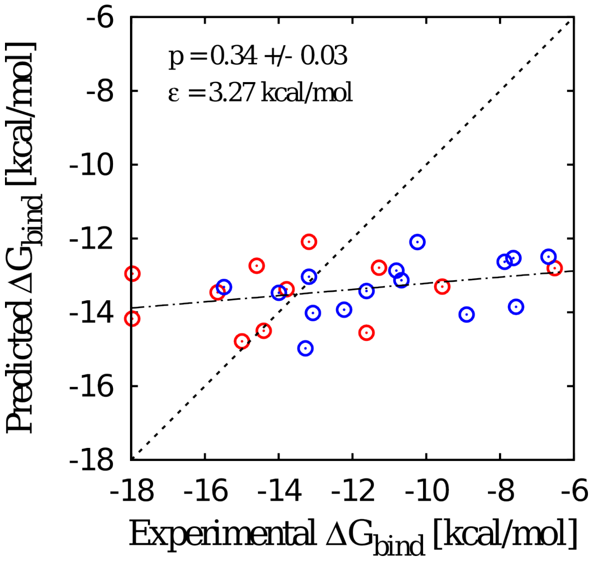


Figure 3.3: Experimental and predicted binding free energies; no estimate for entropic contribution. Eq. (3.18) is used for the theoretical prediction, with optimized parameters $\beta_0 = -12.1223$ kcal $\cdot$ mol^{-1} and $\beta_1 = 0.1670$. Red points: training set. Blue points: test set. All points should lie on the dashed black diagonal line for a perfect predictor. The dashed-dotted black line is the result of the linear regression for all 27 data. The reported Pearson linear coefficient (p) and RMSE ϵ are computed for all 27 data.

Again, we started by not including entropic contribution in our theoretical estimate:

$$\Delta\hat{G} = \beta_0 + \beta_1 \Delta\hat{G}_i \;, \tag{3.18}$$

where β_0 and β_1 are the parameters to be optimized. The reported Pearson correlation coefficient is obtained with the Leave One Out (LOO) procedure as the average of the 27 different results obtained for all possible subsets with 26 data. The resulting standard deviation is reported as well. The LOO procedure is used for the RMSE ϵ as well, the standard deviation being 0.01 kcal $\cdot$ mol^{-1} in all the cases shown in this work. The mild correlation that is found is the same as seen in Fig. 3.2, and the quantitative prediction is definitely poor.

As a second step, we considered the estimation of the entropic contribution $\Delta\hat{G}^0_v$ obtained through Eq. (3.9), by using the same elastic constant k^0_{int} in all ANMs:

$$\Delta\hat{G} = \beta_0 + \beta_1 \Delta\hat{G}_i + \beta_2 \Delta\hat{G}_{vtr} \;, \tag{3.19}$$

where $\beta_0, \beta_1, \beta_2$ are the parameters to be optimized. As can be seen in Fig. 3.4, the linear correlation coefficient slightly improves, but the performance in quantitative prediction is essentially as poor as in the previous case. Even

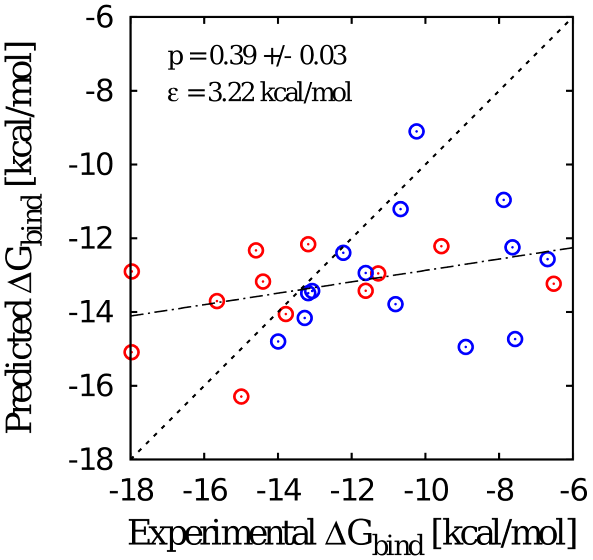


Figure 3.4: Experimental and predicted binding free energies; estimate of entropic contribution with the same elastic constant for all structures. Equation (3.19) is used for the theoretical prediction, with optimized parameters $\beta_0 = -3.5509$ kcal $\cdot$ mol^{-1}, $\beta_1 = 0.0604$ and $\beta_2 = -0.2091$. Red points: training set. Blue points: test set. All points should lie on the dashed black diagonal line for a perfect predictor. The dashed-dotted black line is the result of the linear regression for all 27 data. The reported Pearson linear coefficient (p) and RMSE ϵ are computed for all 27 data.

worse, the negative sign of β_2 implies that the estimation of the entropic contribution is grossly mistaken.

As a third step, we considered the estimation of the entropic contribution obtained through Eq. (3.13), that is by using in the ANM model a different elastic constant k_{int} obtained through the rescaling (see Eq. (3.17)) with all residues included in the matching procedure that determines the scaling constant $\mathcal{C}$. The optimized parameters are again $\beta_0, \beta_1, \beta_2$, according to Eq. (3.19). The sign of β_2 is now positive, but the overall performance is the same as in the previous case, both for the linear correlation coefficient and for the RMSE.

Finally, as a fourth step, we consider the estimate of the entropic contribution obtained through Eq. (3.13), by using in the ANM model a different elastic constant k_{int} obtained through the rescaling (see Eq. (3.17)) where some residues are iteratively excluded from the matching procedure that determines the scaling constant $\mathcal{C}$, until the RMSE between the MD and the ANM mobility profiles of the remaining residues is less than $\Delta\mu = 4$ Å^2 (see Section 3.1). The optimized parameters are again $\beta_0, \beta_1, \beta_2$, according to Eq. (3.19). The performance is now drastically improved for both estimators.

All performance obtained separately for training, test, and full (test and

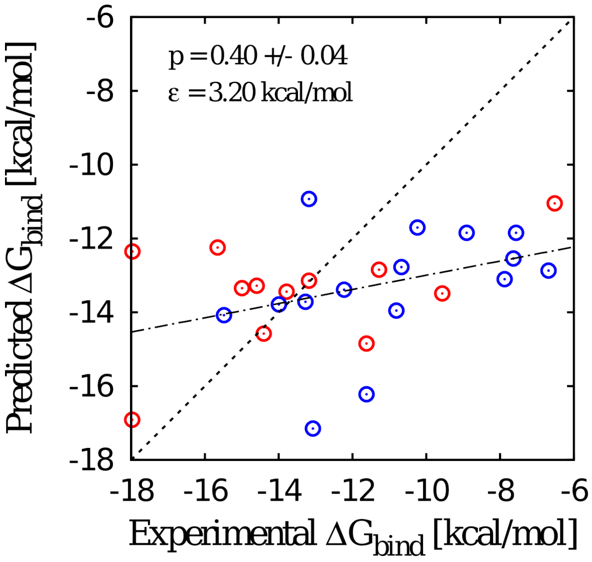


Figure 3.5: Experimental and predicted binding free energies; estimate of entropic contribution with the different rescaled elastic constant for different structures and all residues included in the rescaling procedure with $\Delta\mu = 4$Å.Equation (3.19) is used for the theoretical prediction, with optimized parameters $\beta_0 = -11.7356$ kcal $\cdot$ mol^{-1}, $\beta_1 = 0.1922$ and $\beta_2 = 0.0123$. Red points: training set. Blue points: test set. All points should lie on the dashed black diagonal line for a perfect predictor. The dashed-dotted black line is the result of the linear regression for all 27 data. The reported Pearson linear coefficient (p) and RMSE ϵ are computed for all 27 data.

training) are shown in detail in Table 3.3

A crucial parameter in our estimation of entropic contribution is $\Delta\mu$, the threshold value of the RMSE between MD and ANM mobility profiles, below which the iterative residue exclusion procedure is stopped. The value $\Delta\mu = 4$ Å^2 used in Fig. 3.6 was chosen after a careful analysis, as shown in Figs. 3.7 and 3.8. $\Delta\mu = 4$ Å^2 is the threshold value where both the Pearson correlation coefficient and the RMSE between experimental and predicted binding free energies show the best performance on both the training and the full set. As expected, the performance is better on the training set for both estimators.

Finally, we compare our results with a state-of-the-art method for the prediction of protein complex binding affinities, introduced in [MAB11]. The comparison is obviously taken on the 27 complexes that we were able to address within our methodology. We remind that our training and test sets were deliberately chosen as subsets of the training and test sets, respectively, used in [MAB11]. As can be seen in Fig. 3.9, our performance for both estimators is just slightly worse. This is a remarkable result, given that in our method two independent quantities are combined, the BACH interface score and the

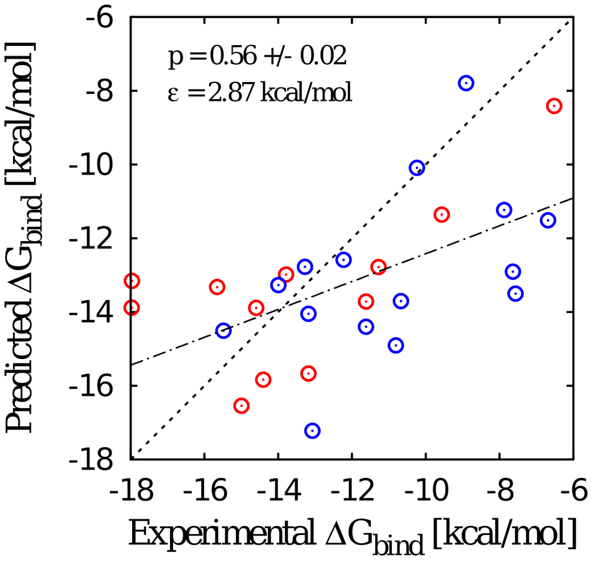


Figure 3.6: Experimental and predicted binding free energies; estimate of entropic contribution with the different rescaled elastic constant for different structures and residues excluded in the rescaling procedure. Equation (3.19) is used for the theoretical prediction, with optimized parameters $\beta_0 = -12.5834 \text{ kcal} \cdot \text{mol}^{-1}$, $\beta_1 = 0.1606$ and $\beta_2 = 0.0297$. Red points: training set. Blue points: test set. All points should lie on the dashed black diagonal line for a perfect predictor. The dashed-dotted black line is the result of the linear regression for all 27 data. The reported Pearson linear coefficient (p) and RMSE ϵ are computed for all 27 data.

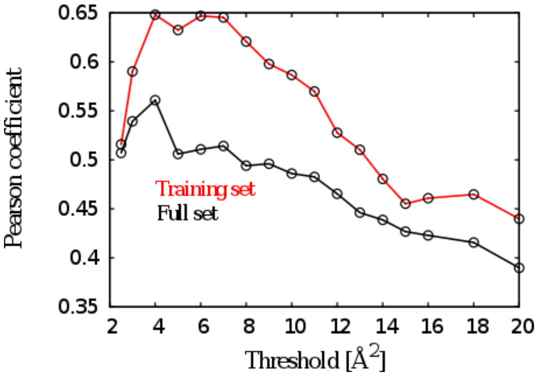


Figure 3.7: Pearson correlation coefficient with varying $\Delta\mu$. Red: training set; black: full set.

(A) BACH contribution only

Complexes	Pearson coefficient	RMSE $(\mathrm{kcal} \cdot \mathrm{mol}^{-1})$
train set	0.26±0.07	3.08
test set	0.42±0.06	3.42
full set	0.34±0.03	3.27

(B) Same *k* for all structures

Complexes	Pearson coefficient	RMSE $(\mathrm{kcal} \cdot \mathrm{mol}^{-1})$
train set	0.37±0.07	2.96
test set	0.46±0.07	3.41
full set	0.39±0.03	3.22

(C) No excluded residues ($\Delta\mu = \infty$)

Complexes	Pearson coefficient	RMSE $(\mathrm{kcal} \cdot \mathrm{mol}^{-1})$
train set	0.43±0.13	2.85
test set	0.40±0.05	3.46
full set	0.40±0.04	3.20

(D) Excluded residues: $\Delta\mu = 4$ Å^2

Complexes	Pearson coefficient	RMSE $(\mathrm{kcal} \cdot \mathrm{mol}^{-1})$
train set	0.64±0.10	2.42
test set	0.50±0.03	3.19
full set	0.56±0.02	2.87

Table 3.3: The values of Pearson coefficients and of RMSE between experimental and predicted binding affinities for different predictors corresponding to Figs. 3.3 to 3.6; for the training set of 12 complexes, the test set of 15 complexes and the full set of 27 complexes.

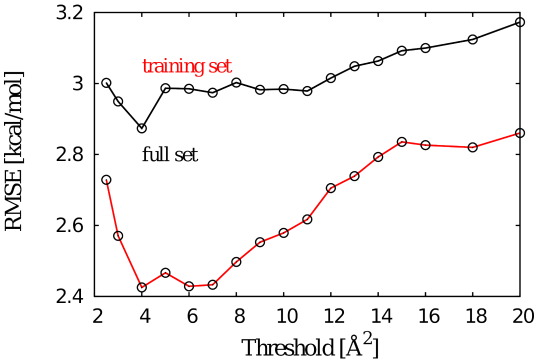


Figure 3.8: RMSE between experimental and predicted binding free energies $(\mathrm{kcal} \cdot \mathrm{mol}^{-1})$ with varying $\Delta\mu$. Red: training set; black: full set.

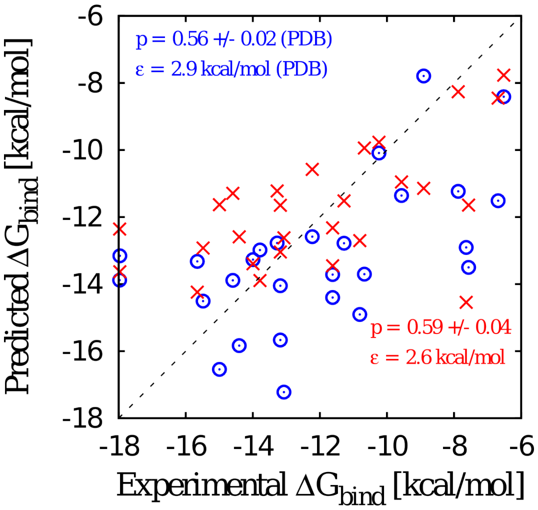


Figure 3.9: Experimental and predicted binding free energies ($\mathrm{kcal} \cdot \mathrm{mol}^{-1}$). Data for all 27 complexes used in this work are shown. Blue circles: our method, as in Fig. 3.6; red crosses: results from [MAB11]. The Pearson linear correlation coefficient and the RMSE ϵ between experimental and predicted values are computed using the LOO procedure. The Pearson coefficient from [MAB11] on the training set is 0.69 ± 0.10, while on the test set is 0.41 ± 0.09, to be compared with the corresponding performances of our method reported in Table 3.3 (D).

entropic contribution, with only 4 parameters being optimized ($\beta_0, \beta_1, \beta_2, \Delta\mu$), compared to the combination of beyond 200 descriptors used in [MAB11]. The good performance of our method underlines that it is crucial to take into account the entropic contribution to the binding affinity. It underlines as well that such contribution can be estimated to a good approximation by using coarse-grained elastic network models, provided that the elastic constant is rescaled carefully, by excluding from the matching procedure the more mobile residues whose behavior cannot be described within the harmonic regime.

Chapter 4

Devising new potentials

KBPs have proved to be a simple and fast tool for recognizing the native state of proteins (see Chapter 2) and also perform well in estimating the free energy contribution of the binding affinity due to the interactions inside proteins (see Chapter 3).

In this chapter we try to test the ability of KBP to capture even more detailed information about protein structures and interactions. With this in mind we investigated some of the limitations of the standard implementation of BACH (described in Section 4.1) and tried to overcome them in two different KBP implementations (presented in Sections 4.2 and 4.3).

4.1 Lessons learned

In applying the BACH scoring function in the contexts of native state recognition and protein binding we investigated the behavior of this simple potential in facing the many situations we put in front of him. One of the strong points of BACH is its simplicity and reduced number of parameters: it is often very easy to locate the problem when it does not work; as a testing ground, BACH has proven to be educative.

4.1.1 Untold hypotheses

In order to understand what are the approximations implicitly made in BACH formulation we tried to derive an equation similar to Eq. (2.1) starting from a very simple physical model of two interacting particles.

We assume that the theoretical probability $\hat{p}^{\ell}_{\alpha,\beta}$ of observing two different particles of type α and β respectively in a contact of type ℓ is

$$\hat{p}^{\ell}_{\alpha,\beta} = \frac{q^{\ell}_{\alpha,\beta} \exp\left(-\frac{\epsilon^{\ell}_{\alpha,\beta}}{K_B T}\right)}{\sum_{\ell'} q^{\ell'}_{\alpha,\beta} \exp\left(-\frac{\epsilon^{\ell'}_{\alpha,\beta}}{K_B T}\right)}, \tag{4.1}$$

where $\epsilon^{\ell}_{\alpha,\beta}$ is an interaction energy, K_B the Boltzmann's constant and T the temperature. Parameters $q^{\ell}_{\alpha,\beta}$ are intended to be the contact probability between the particle if the energies were zero or, equivalently, if the temperature was infinite. Being a probability, $\sum_{\ell} q^{\ell}_{\alpha,\beta} = 1$.

We now try to find the best values for the energy parameters that are compatible with our experimentally observed number of particles of type α and β in contact class ℓ, denoted by $\{n^{\ell}_{\alpha,\beta}\}$. If we also assume that the probability of observing the data $\{n^{\ell}_{\alpha,\beta}\}$ given the value of energies $\{\epsilon^{\ell}_{\alpha,\beta}\}$ is the multinomial distribution

$$p\left(\{n^{\ell}_{\alpha,\beta}\}|\{\epsilon^{\ell}_{\alpha,\beta}\}\right) = \frac{\left(\sum_{\ell} n^{\ell}_{\alpha,\beta}\right)!}{\prod_{\ell} n^{\ell}_{\alpha,\beta}!} \prod_{\ell} \left(\hat{p}^{\ell}_{\alpha,\beta}\right)^{n^{\ell}_{\alpha,\beta}} \tag{4.2}$$

we can easily estimate the best values of energies by maximizing the log-likelihood $\mathcal{L}$

$$\mathcal{L} = \sum_{\alpha,\beta} \sum_{\ell} \log\left[p\left(\{n^{\ell}_{\alpha,\beta}\}|\{\epsilon^{\ell}_{\alpha,\beta}\}\right)\right] \tag{4.3}$$

with respect to the parameters $\epsilon^{\ell}_{\alpha,\beta}$. As a result we obtain

$$\frac{q^{\ell}_{\alpha,\beta} \exp\left(-\frac{\epsilon^{\ell}_{\alpha,\beta}}{K_B T}\right)}{\sum_{\ell'} q^{\ell'}_{\alpha,\beta} \exp\left(-\frac{\epsilon^{\ell'}_{\alpha,\beta}}{K_B T}\right)} = \frac{n^{\ell}_{\alpha,\beta}}{\sum_{\ell'} n^{\ell'}_{\alpha,\beta}}, \tag{4.4}$$

where of course the ratio $p^{\ell}_{\alpha,\beta} = \frac{n^{\ell}_{\alpha,\beta}}{\sum_{\ell'} n^{\ell'}_{\alpha,\beta}}$ represents our best estimation of the theoretical probability $\hat{p}^{\ell}_{\alpha,\beta}$. In order to exactly estimate the energy parameters we define a reference class $\ell = 0$ and divide both the l.h.s. and the r.h.s. of Eq. (4.4) by $p^{0}_{\alpha,\beta}$ to obtain

$$\epsilon^{\ell}_{\alpha,\beta} = \epsilon^{0}_{\alpha,\beta} - K_B T \log\left[\frac{p^{\ell}_{\alpha,\beta}\, q^{0}_{\alpha,\beta}}{p^{0}_{\alpha,\beta}\, q^{\ell}_{\alpha,\beta}}\right], \tag{4.5}$$

that is the same as Eq. (2.1) once we identify

$$\tilde{p}^{\ell}_{\alpha,\beta} = \frac{p^0_{\alpha,\beta}\, q^{\ell}_{\alpha,\beta}}{q^0_{\alpha,\beta} \exp\left(-\frac{\epsilon^0_{\alpha,\beta}}{K_B T}\right)} \tag{4.6}$$

as the reference probability.

This model is very instructive because it makes explicit the hypotheses that have been necessary to obtain a KBP-like formula from a simple physical model.

The parameters $q^{\ell}_{\alpha,\beta}$ have been supposed to only depend on the type of residues α and β and on the interaction class ℓ. This assumption seems to be poor, as we already know from Fig. 1.4 that the probability contact of two particles depends on their distance along the polymer chain. Introducing a dependence on the inter-particle distance in q can be easily done but, nonetheless, BACH does not consider this dependence in its parameters.

Even if Eq. (4.2) seems a reasonable assumption for describing the probability of observing the data $\{n^{\ell}_{\alpha,\beta}\}$, the multinomial distribution is the probability distribution of a certain number of *independent* Bernoulli trials, with the same probability of success on each trial. Employing Eq. (4.2) is therefore equivalent to assume that there are no correlations inside the chain. But the presence of secondary structures induces high correlations; for instance Fig. 4.1 shows the probability density for the distance between two alpha carbons located in two different strands of the same β-sheet.

Finally, Eq. (4.5) shows how critical is the definition of the reference state $\tilde{p}^{\ell}_{\alpha,\beta}$ to a correct and unbiased estimation of the parameters. It is also clear that, at least in this case, the correct reference state depends on α and β. On the contrary, the reference state chosen by BACH is

$$\frac{\sum_{\alpha,\beta} n^{\ell}_{\alpha,\beta}}{\sum_{\alpha,\beta} \sum_{\ell'} n^{\ell'}_{\alpha,\beta}}$$

and therefore does not depend on residue types.

4.1.2 The role of correlations

In order to get some insight on the functioning of the KBP we devised a simple experiment. The aim of the experiment is that of showing how the chain topology can affect the estimation of interaction parameters in KBPs. We

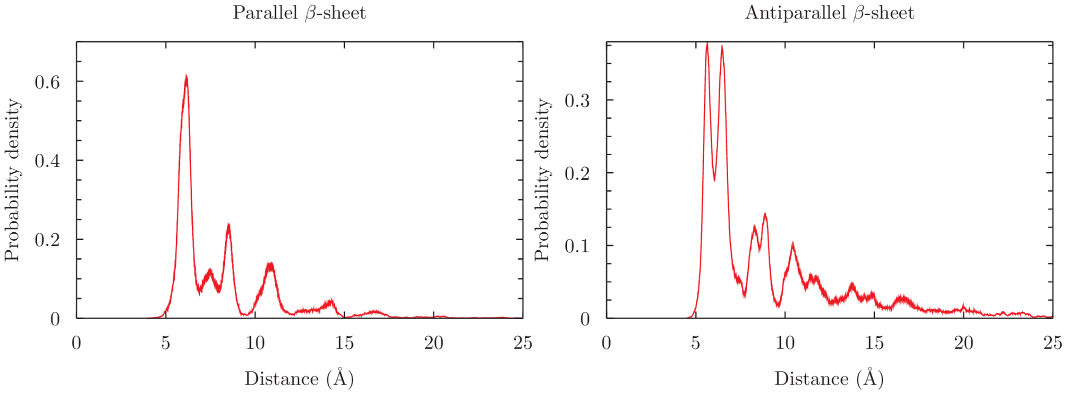


Figure 4.1: The effect of correlations inside β secondary structures. The figures show the probability density for the distance between two alpha carbons located in two different strands of the same β-sheet; pair of residues that take part to the same H-bond are not considered.

therefore consider an ideal case in which we have a perfect knowledge on the system (for instance we know the *Hamiltonian* and we also have access to every possible protein sequence and protein structure) and we ask what happens if we try to compute the parameters of a KBP by using Eq. (2.1). With this in mind we studied a revisited version of the famous HP-model on a bi-dimensional square lattice $\mathcal{Z}^2$ [Dil85, Dil90]. This approach is similar in spirit to the one in Ref. [TCPB04].

In the context of on-lattice protein models each site of the lattice is seen as connected to its adjacent sites: each site can host a protein residue only if it is not occupied by another residue and if the preceding residue is hosted on one of the adjacent sites. As a consequence a protein appears as a self avoiding polygonal chain, briefly referred as Self Avoiding Walks (SAW), connecting adjacent sites of a regular lattice. Two sites are said to be *adjacent* to each other (or *nearest-neighbors*) if they lie on the same side of one square defined by the lattice; they are said to be *next-nearest-neighbors* if the lie on the same diagonal of a square. In the case of a square lattice each site has exactly 4 adjacent sites and other 4 nearest-neighbor sites.

As in the original HP model each residue can be polar (P) or hydrophobic (H). The pairwise potential $V_{\alpha,\beta}(r)$ is attractive in case two H residues are nearest-neighbors (but not consecutive along the chain), repulsive if two H residues are next-nearest-neighbors and zero otherwise. Table 4.1 lists all parameters that define the energy of the system.

In order to have access to all possible information about the system we

Residue pair	Potential nn	Potential nnn	Potential nc
PP	0.	0.	0.
PH	0.	0.	0.
HH	$-1.$	$\frac{\pi}{17}$	0.

Table 4.1: Parameters describing the interactions between pair of residues in the modified version of HP model studied. *Potential nn* refers to the case in which two residues are nearest-neighbors while *Potential nnn* to the case two residues are next-nearest-neighbors. Finally *Potential nc* describe the interaction between all other pairs that are not consecutive along the chain.

restrict the SAW to have a fixed length $N = 16$. Indeed in this case we are able to enumerate easily all the possible sequences as well as all the possible polymer configurations; moreover we are able to associate an energy to every possible combination of structure-sequence: we therefore have access to the whole universe of proteins.
We defined the native state of a protein its minimum energy configuration *if* this configuration is unique; notice that with this definition not every sequence has a native configuration. In this context the ensamble of native structures is intended to mimic a database of experimentally resolved structures and we indeed employed it to determined the parameters of a BACH-like KBP on the lattice. We suppose that the definition of the interacting classes of this hypothetical potential perfectly matches that of the real interactions in such a way that any difference between the parameters of the KBP and that of the real interaction must be solely attribute to the effects of correlations and/or to the choice of the reference state.
Table 4.2 shows the parameters obtained by the described procedure: it is evident how the estimated potential is completely different from the real one. In particular we see how the nearest-neighbor interaction between a polar and an hydrophobic residue appears to extremely repulsive while the repulsive next-nearest-neighbor interaction between two hydrophobic residues appears to be attractive.

This example teaches us that it is possible that correlations completely alter and bias the estimation of the interacting parameters in KBPs. We can not know if the effects in real life scenarios are as drastic as in this simple model but nonetheless it is important to be aware of this possibility.

Residue pair	Potential nn	Potential nnn	Potential nc
PP	0.7521863	0.5344563	−0.1378130
PH	2.5862754	0.0459283	−0.0954614
HH	−0.95	−0.2416561	0.2454912

Table 4.2: The same as in Table 4.1 but here the interaction parameters have been computed with a BACH-like algorithm.

4.1.3 Interfaces

As seen in the previous section, interactions between residues located at the interface between two different monomers are critical in estimating their binding affinity and deserve to be discussed in more details.

The term *interface* between two macromolecules is somehow vague and intuitively define the space in which the two monomers can be consider to be in contact. Therefore we coherently classify a residue as being located on the interface if it interact with at least another residue belonging to the other monomer. From simple geometric consideration we can see that, in the case of compact proteins, while the number of residues scales with the third power of the protein size, the number of residues that belong to the interface scales linearly with it. As a consequence the interface is populated by a relatively low fraction of residues that, nonetheless, play a fundamental role in the binding process. It is not difficult to see that inaccurate classification of these interactions during the protein scoring and/or the usage of inaccurate scoring functions can easily result in wrong estimation of the binding affinity.

We verified that, in some cases in which BACH completely failed in estimating the binding, affinity the interactions that were taking place at the interface occur very rarely in the protein bulk, where BACH parameters are computed. In retrospect it is not surprising that the effective pairwise interactions depend on the environment (protein bulk or protein surface). Indeed, while in the protein bulk contacts are stabilized by pairwise interactions and by topological constraints, this is not true on surfaces were not only topological constraints are different but also there are entropic effects due to the presence of the solvent. The first attempt that has been made in order to capture this behavior is that of splitting the side-chain side-chain class of interaction into two different classes that distinguish between non-polar - non-polar interactions and non specific

ID	Description	Formula
ALC	Hydroxyl	ROH
CH2		$RC(H_2)R'$
CH3		RCH_3
CXY	Carboxyl or Carboxylate	$RC(=O)OH$ or $RC(=O)O^-$
SC3	Methionine functional group	SCH_3
THI	Thiol	RSH
ARG	Arginine functional group	$RN(H)C(=NH)NH_2$
PRK	Group NH in primary ketimine	$RN(H)R'$
PAM	Primary amine	RNH_2
PRA	Protonated amine	RNH_3
PHE	Phenil	RC_6H_5
IMD	Imidazole ring in Arginine	$n1c(H)[nH]c(H)c1$
TRP	Tryptophan rings	$c1ccc2c(c1)c(c[nH]2)CR$
PEP	Peptide group	$RC(=O)N(H)R'$

Table 4.3: List of moieties employed in the definition of mBACH.

interactions [SGS+15]. The resulting KBP, hereafter referred as SixthSense, has proven to be of fundamental importance in the prediction of binding affinities.

4.2 Moieties

The poor performance of BACH, with respect with its improved six-classes version, in the study of binding processes suggests that the coarse grained made in considering each residue as a whole is probably inaccurate.
We therefore studied a statistical potential in which the interacting units are chemical moieties with a more defined chemical behavior. Using such units is important: indeed if two groups of atoms α and β have a well defined behavior we can expect that the event "α and β are in interacting class ℓ" can be considered as favored or disfavored with a certain confidence and independently on their mutual orientation. This is not true if large set of atoms (such as residues) are employed as interacting units as two units can expose to the other different functional groups, depending on their mutual orientation. The first four classes of BACH and the two upgraded classes in SixthSense are devised to account this issue.

We identified 14 groups of atoms that satisfy the following properties:

1. have a well defined chemical behavior
2. cover all possible protein structures, without overlapping

Following the evidences that interactions can be different on the interfaces with respect to the protein bulk, we changed the functional form of BACH in such a way that pairwise interactions between moieties are environment-dependent. This is achieved by distinguishing between four cases:

00 - Both the moieties are buried

01 - The first moiety (α) is buried while the other (β) is exposed to the solvent

10 - The first moiety (α) is exposed to the solvent while the other (β) is buried

11 - Both moieties are exposed to the solvent

Besides these four interacting classes we consider the non-interacting (**nc**) class, for a total of five pair-wise interacting cases. Notice that the information stored in this classes is redundant as $\epsilon^{\mathbf{01}}_{\alpha,\beta} = \epsilon^{\mathbf{10}}_{\beta,\alpha}$, $\epsilon^{\mathbf{11}}_{\alpha,\beta} = \epsilon^{\mathbf{11}}_{\beta,\alpha}$, $\epsilon_{\alpha,\beta} = \epsilon^{\mathbf{00}}_{\beta,\alpha}$ and $\epsilon^{\mathbf{nc}}_{\alpha,\beta} = \epsilon^{\mathbf{nc}}_{\beta,\alpha}$.
We denote this upgraded version of BACH as mBACH, where m resembles the fact that the interacting units are moieties, rather than residues.

The score in mBACH is computed as

$$S_{mBACH} = \frac{1}{2}\sum_{i \neq j} \epsilon^{\ell}_{a_i,a_j}, \tag{4.7}$$

where the summation in the r.h.s. is extended over all the moieties of the structure and ℓ labels the pair-wise interacting class ($\ell \in \{00, 01, 10, 11, nc\}$). Parameters $\epsilon^{\ell}_{a_i,a_j}$ are computed as usual by applying Eq. (2.1) to the counting obtained by analyzing a large database of high-resolution protein structures top500. As in BACH the reference state is chosen as the average interacting probability and is computed on the same set of protein structure. Each moiety is checked to being exposed to the solvent or buried by using the alpha shape based algorithm described in Section 2.3.2: a moiety containing at least one heavy atom that belongs to the alpha shape of the protein is considered exposed to the solvent. The total number of parameters needed in mBACH is 490, less than a half of the ones employed by the standard version of BACH.

The usage of moieties as atomic units in the development of mBACH allows our statistical potential to be applied not only on protein structures but also

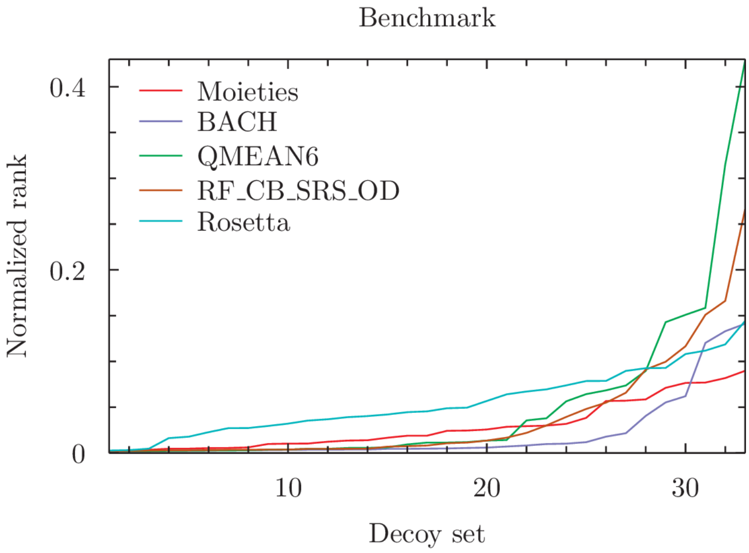


Figure 4.2: Comparison of the performances of mBACH with other state-of-the-art methods.

on other molecules such as drugs and ligands. This potentially interesting application is complicated by the need of automatically recognize functional groups as part of larger molecular structures. The drawback arises from the fact that different nomenclatures are used for labeling atoms in pdb files and from the fact that important structural information (such as bound valency and aromaticity) are not described in pdb files at all. Even if all information that are potentially useful to recognize a moiety as par of a larger molecule can be extracted from the list of atom properties (for instance bonds can be assigned by comparing the inter-atomic distance with the vdW radii of the two atoms) this approach is not efficient nor reliable. Fortunately the Protein Data Bank provide connectivity information, together with additional atom properties, in separate mmcif files.

We used these information in order to build an undirected graph that describe the structure of the target moleculeand in which vertexes represent atoms while edges represent inter-atomic bonds. Each vertex is described by two properties: an *element* descriptor, which uniquely identify the atom element, and a boolean *aromaticity* flag that classify the atom as aromatic or not aromatic. Bonds can be both single or double and aromatic or not aromatic, as described respectively by the *type* and *aromaticity* bond descriptors. Once the exact structure topology is built, moieties can be recognized by checking for the existence of some sub-graph isomorphism between the topology of the structure and that of the moiety. A sub-graph of the structure topology S is said to be isomorphic to the graph that describe the moiety M if and only if there is a bijection f from a subset of vertexes of S to the vertexes of M such that, for every pair of vertexes u,v:

1. $f(v)$ and $f(u)$ are connected in M if and only if u is connected to v in S
2. each vertex $f(u)$ has the same *element* and *aromaticity* of u
3. the bond connecting $f(u)$ to $f(v)$ has the same *type* and *aromaticity* as the respective bond between u and v.

Since we require that every atom belongs to at most one moiety, we have to pay attention to the order in which we assign an atom to a moiety. Consider, for instance the two moiety $CH2$ and $CH3$: it is clear that since there are three possible different sub-graph isomorphism between the two functional groups, an atom that belongs to a $CH3$ functional group can also be assigned to a $CH2$ one. The problem is simply overcame by checking the existence of isomorphisms starting from the biggest moiety (i.e. the one with the highest number of atoms) and by proceeding with order until a sub-graph isomorphism between the structure and the smaller moiety is checked. We optimized the process by removing from the large structure those vertexes that have been already assigned to a moiety, in such a way that subsequent steps of the algorithm proceed faster due to the ever decreasing size of the structural graph.

4.3 A gaussian chain potential

Simplifications usually made in the formulation of knowledge based potentials like BACH are based on the assumption that the presence of contacts between two groups of atoms inside a macro-molecule is due only to their properties and not to topological reasons. For instance the fact that residues are serially linked together is often neglected and events like "Residue i is in contact with residue j" and "Residue i' is in contact with residue j'" are considered as independent. This is obviously not true, as we already know that, inside β-sheets, the linear topology of the protein structure induces topological contacts between residues that are not related to any physical interaction (see Fig. 4.1). But the biggest effects in the estimation of pairwise parameters is probably due to the fact that we forget about the dependence of the distance between two group of atoms on their separation along the chain: it is indeed clear that the contact probability between two residues is a decreasing function of their separation along the chain, as shown in Fig. 1.4.

Another limitation of the KBP is the adoption of an *a priori* functional form of the potential which is implicitly determined by the set of events used to classify different interaction: its choice is often based of physical/chemical considerations and is not deduced from experimental data.

The effects introduced by these simplifications are not easily quantifiable; hopefully the reference state can be tuned to partly correct the error induced in the estimation of the parameters but nonetheless neglecting to model a key aspect of the protein structure such as it linear topology has consequences in the estimation of parameters in knowledge-based potentials (see Section 4.1.2).

We here propose a new knowledge-based potential that is ideally unaffected by the presence of topological contacts and that keep into account the dependence of the pairwise distance between residues on their separation along the chain. Moreover the method is devised is such a way that the dependence of the score on the pairwise distance r is not imposed *a priori* but is instead directly recovered from experimental data.

In 2005 Banavar et al. [BHM05] showed that the end-to-end distances of m-residues-long protein fragments are gaussian distributed according to the law

$$\tilde{P}_\sigma(r,m) = \frac{4\pi r^2}{\left(\frac{2}{3}\pi\sigma^2 m\right)^{\frac{3}{2}}} \exp\left(\frac{3r^2}{2\pi\sigma^2 m}\right), \tag{4.8}$$

with the assumption that m is sufficiently high ($m \geq 48$) to *forget* the presence of secondary structures and sufficiently low ($m < N^{\frac{2}{3}}$) to *forget* the presence of the protein surface (as indicated in [LBG04]). This scaling is the one predicted by Flory's theory at the θ point [FV⁺69, BGM13], for a polymer melt [LBG04], and, of course, by a gaussian chain model with average bond length σ. The existence of the gaussian scaling of Eq. (4.8) is usually explained by a sort of compensation between the inter-atomic repulsion inside the chain that tend to swollen it and the idrophobic effects due to the presence of the solvent, that tend to compact the protein. As a consequence the end-to-end distance of real-protein fragments behave *as if* fragments were ideal gaussian chains and is ideally unaffected by interactions between residues located in the central part of the fragment, or by other topological effects.

We reproduced the results presented in [BHM05] by using a larger data-set

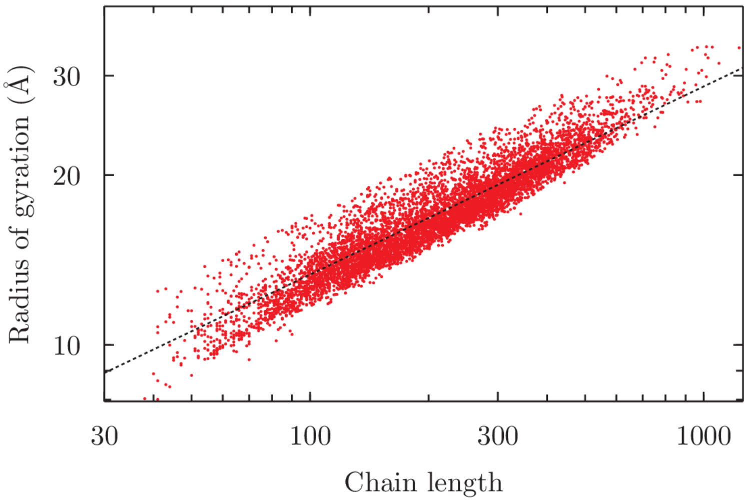


Figure 4.3: Distribution of the gyration radius on the selected subset of top8000. The slope of the fitted line (in log-log-scale) is $b = 0.359 \pm 0.002$ Å while the intercept is $a = 2.493 \pm 0.001$ Å.

of globular proteins, selected from top8000 in such a way that their gyration radius were proportional to the cube root of the chain length (see Fig. 4.3). We computed the end-to-end distance distribution of protein fragments of length $N^{\frac{2}{3}}$, where N is the total chain length. We also verified that the condition $m \geq 48$, suggested in the original paper, can be safely relaxed to $m \geq 35$. Figure 4.4 shows the distribution of experimental data obtained from the structures of top8000 database. End-to-end distances have been computed as the distance between the terminal C_α carbons and have been divided by a factor $\sqrt{m}$ (x axis); y axis has instead been rescaled by a factor $\sqrt{m}$ in such a way that data collapse on the curve of equation

$$y(x) = \frac{4\pi x^2}{\left(\frac{2}{3}\pi\sigma^2\right)^{\frac{3}{2}}} \exp\left(\frac{3x^2}{2\pi\sigma^2}\right), \tag{4.9}$$

which is independent from m. The variance σ of the experimental distribution is $\sigma^\star = 3.84 \pm 0.01$ Å, a value surprisingly similar to the distance between two consecutive α carbons. For each experimental data (x_i, m_i) collected we estimated the corresponding value of probability density by considering the number of data occurring in the interval of width $\Delta = 0.01$ Å and centered in (x_i, m_i), i.e. $\left[(x_i - \frac{\Delta}{2}, m_i), (x_i + \frac{\Delta}{2}, m_i)\right]$.

Despite the overall good agreement, experimental data substantially differ from the theoretical distribution in the short-distance regime, where the experimental probability is lower than the theoretical one, and in the long-range regime

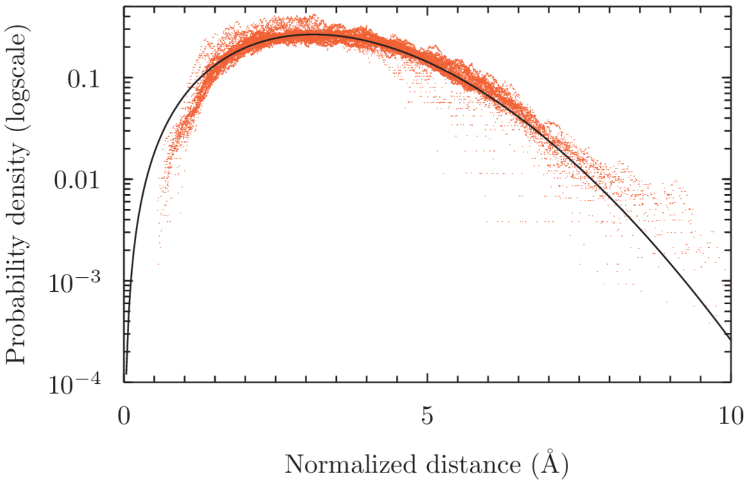


Figure 4.4: Distribution of experimental data (red point) as compared to the expected curve Eq. (4.9) and $\sigma = \sigma^\star$. Axis of ordinates is in log-scale.

where, on the contrary, the experimental probability is slightly higher than the theoretical one.

We assume that the observed distribution of end-to-end distances $P(r, m)$ is given by

$$P(r,m) = \frac{\exp\left(-\frac{V(r)}{K_BT}\right)\ \tilde{P}_{\sigma^\star}(r,m)}{\int_0^\infty dr\ \exp\left(-\frac{V(r)}{K_BT}\right)\ \tilde{P}_{\sigma^\star}(r,m)}. \tag{4.10}$$

Indeed the overall good agreement of the theoretical curve in Eq. (4.8) with $\sigma = \sigma^\star$ with experimental data allow us to identify Eq. (4.8) as the *expected* probability of finding a contact between residues given their separation along the chain and to use it as *reference state* in a KBP. According to Eq. (2.1) we therefore compute a pairwise statistical potential that is a (continuous) function of the distance between α-carbon atoms by computing

$$\epsilon(r,m) = -K_BT\ \log\left(\frac{P(r,m)}{\tilde{P}_{\sigma^\star}(r,m)}\right). \tag{4.11}$$

From Eq. (4.10) we can reduce the previous equation to

$$\epsilon(r,m) = V(r) + K_BT\ \log\left(\int_0^\infty dr\ \exp\left(-\frac{V(r)}{K_BT}\right)\ \tilde{P}_{\sigma^\star}(r,m)\right), \tag{4.12}$$

which assures us that the estimated function $\epsilon(r)$ differs from the *real* inter-particle potential by only a constant.

Figure 4.5 shows an ensamble of 150000 experimental values of $\epsilon(r, m)$ com-

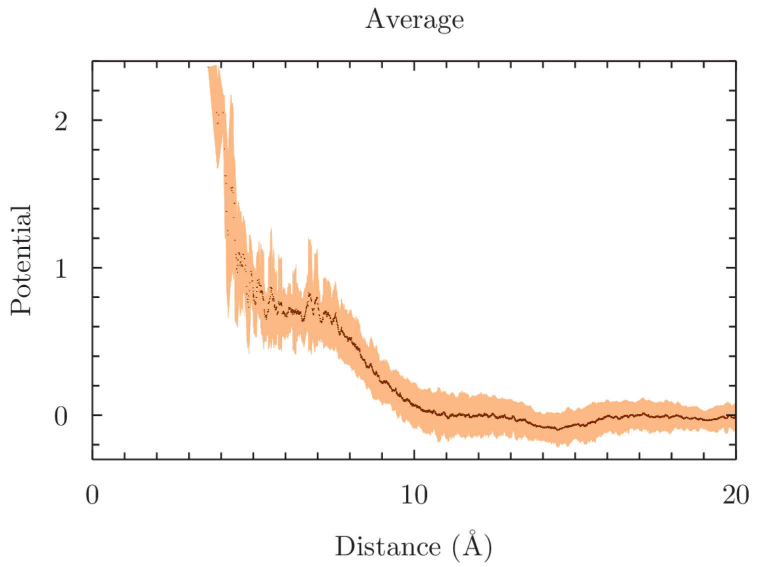


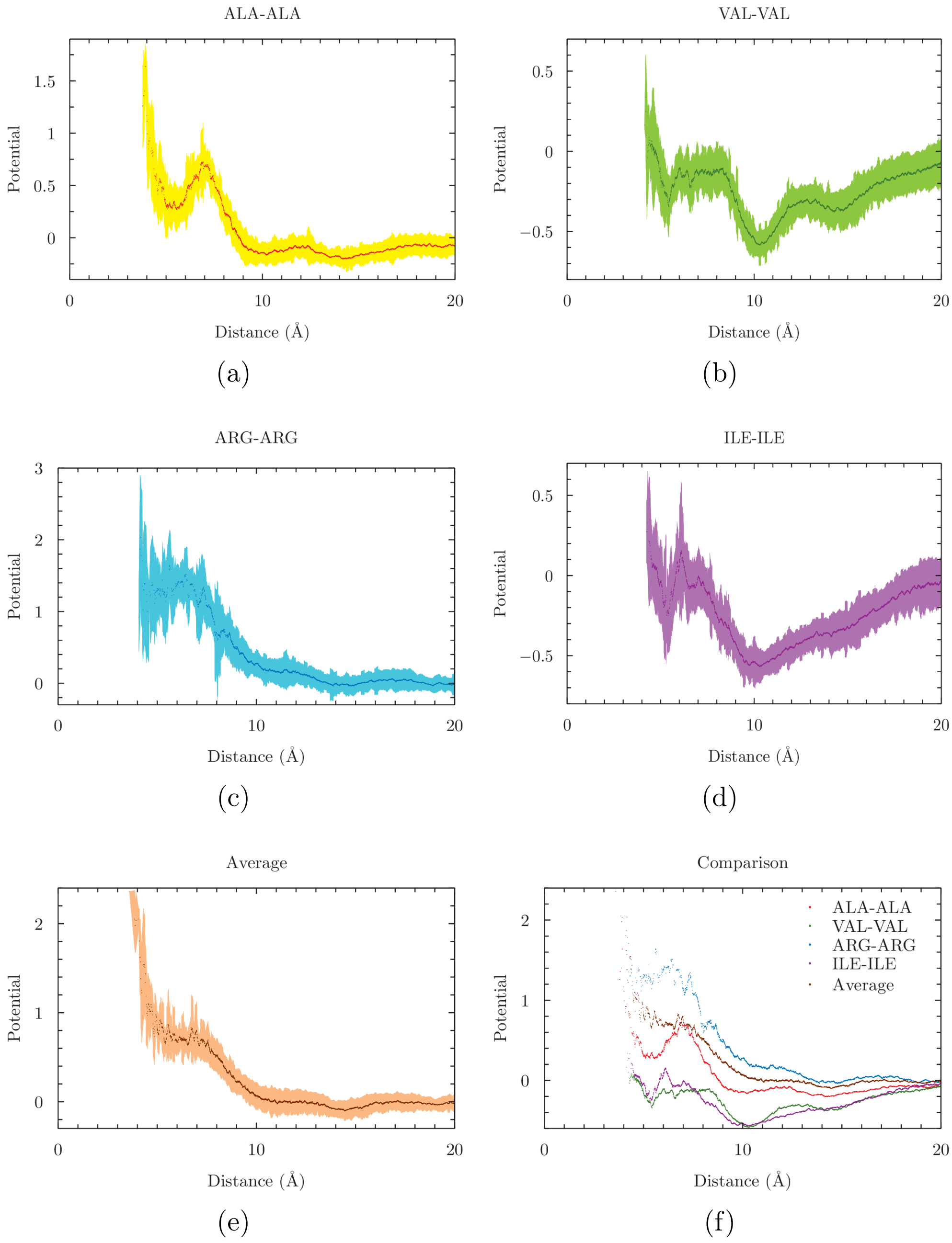
Figure 4.5: Average potential between two arbitrary residues inside a globular structure. The colored shadow represent the error associated to the average curve.

puted as described above; all points collapse on a well defined curve, independently from their value of m. This fact allows us to identify $\epsilon(r) \equiv \epsilon(r, m)$ as an effective potential. We therefore estimated $\epsilon(r)$ by averaging the data obtained from different values of m: the curve superposed to the data is a running average of the points and defines the function $\epsilon(r)$. At this stage, we intentionally neglect the dependence of the potential on the type of the interacting residues: the curve depicted in Fig. 4.5 is an average effective-potential that do not distinguish between different type of amino acids.

Data suggest that proteins can be described as repulsive chains with an hard-sphere repulsion (for distances $r \lesssim 4$ Å) and a soft shell (for distances 4 Å$\lesssim r \lesssim 10$ Å). Residues feel no interactions, on average, for distances greater than 10 Å.

We recover the dependence on the type of the interacting residues by restricting the set of fragments used for computing the contact probability $P^{\alpha,\beta}(r, m)$ to the ones whose first and last residues were respectively of type α, β of β, α. Beta sheets have been identified as a source of spurious contacts and the effects of these contacts have been canceled. In order to avoid biases in the computation of $P^{\alpha,\beta}(r, m)$ due to spurious contacts inside β-sheets (see Fig. 4.1) we did not consider pair of C$_\alpha$ carbons that belong to the same β-sheet and that are not hydrogen-bonded to each other.

Again we used theoretical distribution $\tilde{P}_{\sigma^\star}(r, m)$ as reference state and we compute the pairwise parameters of the scoring function by employing Eq. (2.1)



(a) (b) (c) (d) (e) (f)

Figure 4.6: A selection of pairwise potentials. Panel 4.6a describe the interaction between two Alanine residues, panel 4.6b the one between two valine residues, panel 4.6c the one between two arginine residues and finally panel 4.6d depicts the interaction between two isoleucine residues. Panel 4.6e represents the effective interaction between residues, independently from their type and panel 4.6f shows a direct comparison of all the potentials shown in the other panels. In all figures solid dots are the running average of the data while the shaded region is the error associated with the average in each point: its width is twice the standard deviation of the set of data that contribute to the running average.

as

$$\epsilon^{\alpha,\beta}(r) = -\log\left(\frac{P^{\alpha,\beta}(r,m)}{P_{\sigma^\star}(r,m)}\right) \tag{4.13}$$

where indexes $\{\alpha, \beta\}$ label the amino acid type and $P^{\alpha,\beta}(r,m)$ represents the end-to-end distance of fragments whose first residue is of type α and the last one is of type β, or vice-versa. The resulting potentials (some of which are shown in Fig. 4.6) exhibit features that are peculiar of the residues involved in the interaction: notice for instance how the interaction between two charged ARG is repulsive, as expected. We also notice that the value of $\epsilon(r)$ at large values of the distance is approximately zero. This make us to suppose that the constant $\epsilon(r) - V(r)$ is small when compared to typical values of $\epsilon(r)$.

In view of more appropriate choice of the reference state, the KBP introduced in this section correctly accounts for the dependence of the contact probability on the separation along the chain.

Even if the technique described in this section seems promising for the development of reliable KBPs, it is necessary to point out how this specific potential is not able to distinguish the different possible pairwise interactions between two residues, e.g. does not depend on their relative orientation. As a consequence a benchmark on the same CASP targets employed in [CGL+12] and in Sections 2.3.1 and 2.3.2 shows its limited performances in recognizing the native structure of proteins (see Fig. 4.7).

We can then conclude that categorize the possible interactions between residues i smore crucial than to take into account the correlations inside the protein chain. Further or going work is planned to develop this strategy.

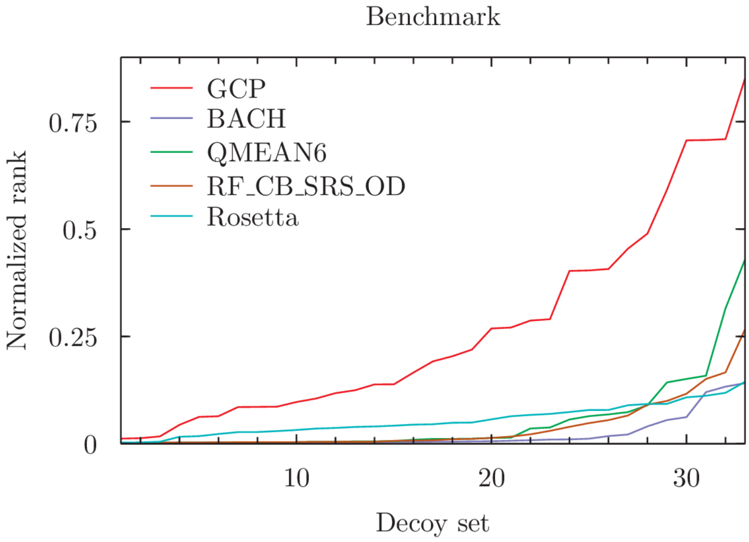


Figure 4.7: Comparison of the performances of Gaussian Chan Potential (GCP) with other state-of-the-art methods.

Chapter 5

Backbone movements

We consider the problem of the local movements of a chain molecule where a small subset of degrees of freedom, e.g. dihedral angles, bonds angles or bond lengths, are concertedly modified inside a specific portion of the chain, in such a way that *only* the atoms in that region are moved while all the others are kept fixed. We do not place any constraints on the degrees of freedom that are modified: they can be chosen everywhere along the part of the chain we want to move without the necessity of belonging to atoms/bonds which are consecutive along the polymer. The issue of local movements is related to the loop closure problem, i.e. finding conformations of a segment of consecutive atoms in a chain molecule that are geometrically consistent with the rest of the chain structure. These questions arise in the context of the control of robotic manipulators made up of serially connected joints, where in many common applications one end is fixed and the other must be positioned at a specific location and with a given orientation [HH92], but they are also topics of paramount importance in structural chemistry and in computational biology. For instance, effective loop closure tools can enhance the performances of homology modeling where segments of insertion or deletion have to be predicted while the rest of the protein structure is reasonably well known from structures of homologous proteins. The ability of moving efficiently a chain may as well have useful applications in the Monte Carlo dynamics for large scale simulations of dense polymeric systems [LCOS08]. In such situation, the efficiency of Monte Carlo simulations relies heavily on the kinetic algorithm used to sample the various possible conformational states of the molecule and the introduction of a concerted move which restricts itself to modifying atom locations only in limited positions of the molecule might play a key role to boost performances, by reducing the hindering effect due to excluded volume constraint. Even the theoretical study of conformational flexibility can benefit from the use of local collective movements that, at variance with Cartesian moves, avoid geometric distorsions of the chain giving the possibility to explore all the possible local arrangements of the flexible molecule.

In the biological context, the problem was often reduced to modify dihedral angles that are the only soft degrees of freedom of the system. A first analysis was due to Go and Scheraga [GS70] with an analytical approach and with the developing of equations for determining the allowed dihedral angles when all the rotating bonds are connected and when a subset is separated by rigid bonds in the trans conformation. This approach was further extended by Theodoru and coworkers [DBT93] to take into account the necessary requirements to ensure Boltzmann distributed sampling for Monte Carlo simulations and by Dinner [Din00] to generalize the formalism to allow fixed dihedral angles that sequentially interrupt the rotating bonds to be non-planar. Another refinement was later proposed with methods [FIS01, UJ03, UJ04, Bet05] in which the inclusion of bond angle variations and of local constraints improved the efficiency of the algorithm. Different formalisms were proposed by Hoffmann and Knapp [HK96] to derive equations for dihedral angles and to apply these moves to study a protein-like model that had the topology of polyalanine with rigid peptide planes and by Coutsias et al. [CSJD04, CSWD06] with a robotics inspired approach. The latter approach was succesfully incorporated in state-of-art protein modeling tools, allowing sub-angstrom accuracy in loop reconstruction [MCK09, SK13]. More recently, an efficient numerical method to solve the analytical solution of the classic chain closure problem was introduced with the specific purpose of optimizing Monte Carlo performances for dense molecular systems [BBEJ$^+$12].

Broadly speaking, all these methods are proposing efficient solutions, while violating the general rules of the problem: either by imposing restrictive conditions on the degrees of freedom that can be used (e.g. moving only specific angles) or by relaxing the boundary constraints (e.g. not keeping completely fixed all the other degrees of freedom).

On the contrary, here we develope a numerical strategy that, exploiting the knowledge of an initial configuration of the chain, allows for an exhaustive exploration of all the possible configurations that can be obtained by modifying *only* $n > 6$ degrees of freedom, and that perfectly adapts to the frozen part of the chain. The choice of the degrees of freedom themselves is completely free and any combination of bond and dihedral angles and of bond lengths can be selected, resulting in a very rapid and efficient search algorithm.

Starting from a geometrical description of the chain inspired by robotic language [HD64], similar in spirit to the one introduced by Go and Scheraga [GS70], we derive six numerical equations, as a function of the n free variables, which have to be fulfilled to satisfy the boundary conditions. Therefore, if there is no degeneracy, the solutions that have to be found lie on a manifold with dimension $n - 6$. The novel idea we here present consists in an extremely powerful strategy to explore these manifolds, based on moving slightly out along their tangent space and on coming back along the orthogonal space towards a new configuration which satisfies all the equations and constraints. This is viable by means of an appropriate double change of coordinates and by employing mathematical algorithms to optimize the computational time. Moreover, the algorithm is designed in such a way that the detailed balance is quite easily satisfied for a very general choice of the modified degrees of freedom.

The efficiency of the approach is remarkable and it makes possible, for instance, to estimate the volume of the manifold which corresponds to the number of possible conformations that are compatible with the constraints, in the simplest $n = 7$ case. While at this stage the method is presented for an ideal chain, without taking into account excluded volume or other energy functions, such features can be introduced in a straightforward manner.

Although the method is completely general and can be applied to any sort of linear object, it is intriguing to think about its applications to protein chains. In such context bond angles and bond lengths can be considered constant and the ψ and ϕ dihedral angles (Ramachandran's angles) are the natural degrees of freedom to be modified: the algorithm we propose thus becomes a generalized crankshaft move involving a portion of the chain of desired length. Some possible applications on proteins are shown, such as the estimation of their backbone mobilities and local structure refinement.

5.1 Mathematical framework

5.1.1 Denavit-Hartenberg parameters for chain description

In this section we introduce the parametric representation of a linear chain used to derive the equations at the core of our algorithm. We consider a linear chain composed of $N + 1$ atoms linked serially in which each of the N bonds can be labeled with numbers from 1 to N. We describe the chain by using the Denavit-Hartenberg (DH) notation [HD55]. According to DH a local reference system $\mathcal{O}_i$ can be built on each bond composing the chain: the $\hat{z}_i$ axis lies on the bond while $\hat{x}_i$ is oriented as $\hat{z}_{i-1} \times \hat{z}_i$. The $\hat{y}_i$ axis is given, as usual, by the right-hand rule. The origin o_i of each reference frame is always located along $\hat{z}_i$: in case $\hat{z}_i$ is co-planar to $\hat{z}_{i-1}$, it lies on the first atom defining the bond; otherwise it lies on the common normal to $\hat{z}_i$ and $\hat{z}_{i-1}$.

A vector $\vec{a}^j$ in the reference frame $\mathcal{O}_j$ can be expressed relative to $\mathcal{O}_{j-1}$ as

$$\vec{a}^{j-1} = \boldsymbol{R}_j^{j-1} \cdot \vec{a}^j + \vec{S}_j^{j-1}, \tag{5.1}$$

where $\boldsymbol{R}_j^{j-1}$ is a 3×3 orthogonal matrix that expresses the orientation of $\mathcal{O}_j$ relative to $\mathcal{O}_{j-1}$ and $\vec{S}_j^{j-1}$ is a vector describing the position of the origin o_j with respect to $\mathcal{O}_{j-1}$. The matrix $\boldsymbol{R}_j^{j-1}$ and the vector $\vec{S}_j^{j-1}$ can be completely described by using four parameters named *link offset*, *link twist*, *link length* and *joint angle*. These are defined by the following rules:

- the *link offset* d_i is the distance along $\hat{x}_i$ from o_i to the intersection of the $\hat{x}_i$ and the $\hat{z}_{i-1}$ axes (i.e. the minimum distance between $\hat{z}_{i-1}$ and $\hat{z}_i$ axis);

- the *link twist* α_i is the angle between $\hat{z}_{i-1}$ and $\hat{z}_i$ measured about $\hat{x}_i$;

- the *link length* r_i is the distance along $\hat{z}_{i-1}$ from o_{i-1} to the intersection of the $\hat{x}_i$ and the $\hat{z}_{i-1}$ axes and

- the *joint angle* θ_i is the angle between $\hat{x}_{i-1}$ and $\hat{x}_i$ measured about $\hat{z}_{i-1}$.

Figure 5.1 shows how the DH parameters are defined for the general disconnected case (i.e. link offset $d_i > 0$). This is actually a typical case, when the structure of a protein backbone chain with all its heavy atoms is considered, since the ω torsional angle around the peptide bond is a hard degree of freedom with a

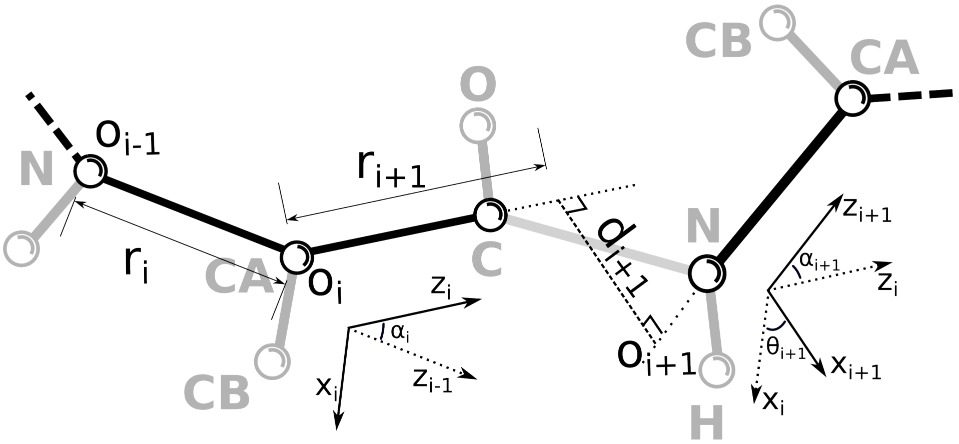


Figure 5.1: General graphical representation of a chain according to the Denavit-Hartenberg convention, as discussed in the text. Thick lines represent the physical bonds and spheres the atoms. α, θ, r and d are the DH parameters describing the chain and o is the origin of each local frame $\mathcal{O}$. The structure of a portion of a protein backbone chain with all its heavy atoms is shown superimposed.

well defined typical value. If ω is then kept strictly fixed, the DH convention allows to "spare" that degree of freedom, defining a disconnected chain as shown in Fig. 5.1. In the simplest case, when all bonds included in the DH description are connected with each other (i.e. all link offsets $d_i = 0$), the DH variables have a well defined physical meaning. Link lengths are bond lengths, link twists are supplementary of bond angles, and joint angles are torsional angles. The DH formalism is in this case equivalent to the one routinely used by software programs that reconstruct biomolecular structures subject to experimental restraints[GMW97, BAC+98] and that employ efficient internal dynamics algorithms that update only the values of torsional angles[JVR93].

By using the DH definitions the matrix $\boldsymbol{R}_j^{j-1}$ and the vector $\vec{S}_j^{j-1}$ can be explicitly expressed as

$$\boldsymbol{R}_j^{j-1} = \begin{pmatrix} \cos(\theta_j) & -\sin(\theta_j)\cos(\alpha_j) & \sin(\theta_j)\sin(\alpha_j) \\ \sin(\theta_j) & \cos(\theta_j)\cos(\alpha_j) & -\cos(\theta_j)\sin(\alpha_j) \\ 0 & \sin(\alpha_j) & \cos(\alpha_j) \end{pmatrix} \tag{5.2a}$$

$$\vec{S}_j^{j-1} = \left(d_j\cos(\theta_j),\ d_j\sin(\theta_j),\ r_j\right)^T . \tag{5.2b}$$

With a more compact notation we rewrite Eq. (5.1) with the following

$$\boldsymbol{a}^{j-1} = \mathsf{T}_j^{j-1}\boldsymbol{a}^j , \tag{5.3}$$

where T_j^{j-1} is a 4×4 matrix given by

$$\mathsf{T}_j^{j-1} = \begin{pmatrix} \boldsymbol{R}_j^{j-1} & \vec{S}_j^{j-1} \\ 0 & 1 \end{pmatrix} \tag{5.4}$$

and $\boldsymbol{a}$ is the vector $\boldsymbol{a} = (\vec{a},\ 1)^T$. With this notation it is easier to relate any $\mathcal{O}_j$ with any other $\mathcal{O}_i$ $(j > i)$; indeed the following equation holds:

$$\boldsymbol{a}^i = \mathsf{T}_j^i \boldsymbol{a}^j = \mathsf{T}_{i+1}^i \mathsf{T}_{i+2}^{i+1} \cdots \mathsf{T}_j^{j-1} \boldsymbol{a}^j. \tag{5.5}$$

It should be now clear that the chain can be described by the whole set $\{r_i, \alpha_i, d_i, \theta_i\}_{i=1,\cdots,N}$. For simplicity we denote $\{r_i\}_{i=1,\cdots,N}$ with $\boldsymbol{r}$, $\{\alpha_i\}_{i=1,\cdots,N}$ with $\boldsymbol{\alpha}$, $\{d_i\}_{i=1,\cdots,N}$ with $\boldsymbol{d}$ and $\{\theta_i\}_{i=1,\cdots,N}$ with $\boldsymbol{\theta}$; the chain configuration is then given by the set $\{\boldsymbol{r},\ \boldsymbol{\alpha},\ \boldsymbol{d},\ \boldsymbol{\theta}\}$.

5.1.2 Performing the concerted local move

It is possible to deform an initial configuration $\{\boldsymbol{r}_0,\ \boldsymbol{\alpha}_0,\ \boldsymbol{d}_0,\ \boldsymbol{\theta}_0\}$ by changing at least one of the parameters describing it.

Consider n DH parameters $\tilde{\xi}_\mu$, with $\mu = 1, \cdots, n$ and the n_b bonds to which the n parameters are related. The number of bonds to be considered is always smaller or equal to the number of DH parameters because in principle two or more parameters could be related to the same bond. As already stated we consider the general case in which these bonds can be non-consecutive. There are two particularly interesting bonds among the n_b: the first and the last i.e. the one with the lowest label and the one with the highest one. These two bonds delimit the region of the chain we are interested in modifying with an opportune change of $\tilde{\boldsymbol{\xi}}$, leaving the atoms outside this region unmodified (see Fig. 5.2). Such a change can be highly non-trivial and could not always be obtained: we will see later a condition that ensures us that this change can be performed.

For convenience, we re-label the bonds we are interested in with numbers from 1 to n_b+1, where the latter is the first bond that remains fixed subsequent to the moved portion of the chain. The condition that needs to be imposed in order to ensure the locality of the change $\{\boldsymbol{r}_0,\ \boldsymbol{\alpha}_0,\ \boldsymbol{d}_0,\ \boldsymbol{\theta}_0\} \to \{\boldsymbol{r},\ \boldsymbol{\alpha},\ \boldsymbol{d},\ \boldsymbol{\theta}\}$ is

$$\mathsf{T}_{n_b+1}^1(\boldsymbol{r},\ \boldsymbol{\alpha},\ \boldsymbol{d},\ \boldsymbol{\theta}) = \mathsf{T}_{n_b+1}^1(\boldsymbol{r}_0,\ \boldsymbol{\alpha}_0,\ \boldsymbol{d}_0,\ \boldsymbol{\theta}_0), \tag{5.6}$$

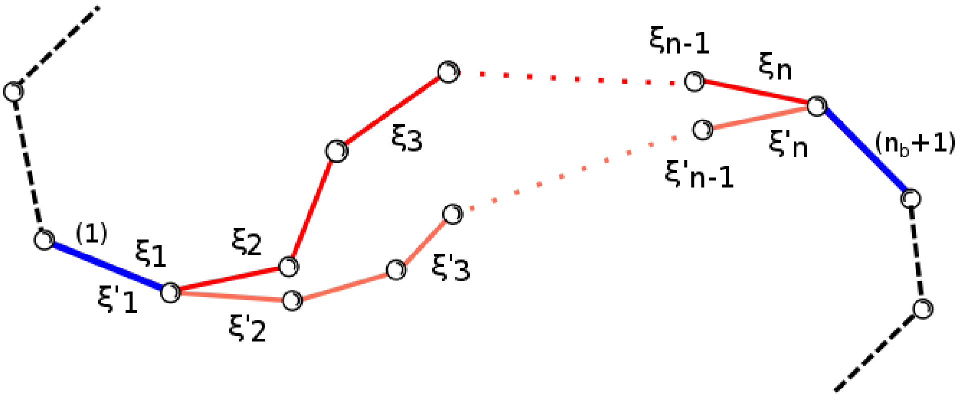


Figure 5.2: Schematic representation of the portion of a linear chain involved in a local modification. Bonds (1) and $(n_b + 1)$ are colored in blue while all the others are in red. The degrees of freedom which are varied $(\xi_1, \dots, \xi_n)$ are arbitrarily distributed inside the region. When they are concertedly changed to new values $(\xi'_1, \dots, \xi'_n)$, the new pink configuration is obtained while all the bonds outside the region (in black) remain fixed in space.

that is requiring that the local reference frame $\mathcal{O}_{n_b+1}$ does not move with respect to the first one. In order to explicit the variables $\tilde{\boldsymbol{\xi}}$ inside the relation in Eq. (5.6) we define, with abuse of notation,

$$\mathsf{T}^1_{n_b+1}(\tilde{\boldsymbol{\xi}}) \equiv \mathsf{T}^1_{n_b+1}(\boldsymbol{r},\ \boldsymbol{\alpha},\ \boldsymbol{d},\ \boldsymbol{\theta}) - \mathsf{T}^1_{n_b+1}(\boldsymbol{r}_0,\ \boldsymbol{\alpha}_0,\ \boldsymbol{d}_0,\ \boldsymbol{\theta}_0) \tag{5.7}$$

in such a way that Eq. (5.6) can be rewritten as

$$\mathsf{T}^1_{n_b+1}(\tilde{\boldsymbol{\xi}}) = 0. \tag{5.8}$$

Given the form of T^j_i (described in Eq. (5.4)), the 16 equations that are implicit in Eq. (5.8) can be reduced to 6 equations in the n variables $\tilde{\boldsymbol{\xi}}$. Three equations are needed in order to set the translational part of T^j_i and other three for the rotational part. We choose, for instance,

$$\begin{cases} f_1(\tilde{\boldsymbol{\xi}}) \equiv \left[\mathsf{T}^1_{n_b+1}(\tilde{\boldsymbol{\xi}})\right]_{01} \\ f_2(\tilde{\boldsymbol{\xi}}) \equiv \left[\mathsf{T}^1_{n_b+1}(\tilde{\boldsymbol{\xi}})\right]_{02} \\ f_3(\tilde{\boldsymbol{\xi}}) \equiv \left[\mathsf{T}^1_{n_b+1}(\tilde{\boldsymbol{\xi}})\right]_{12} \\ f_4(\tilde{\boldsymbol{\xi}}) \equiv \left[\mathsf{T}^1_{n_b+1}(\tilde{\boldsymbol{\xi}})\right]_{03} \\ f_5(\tilde{\boldsymbol{\xi}}) \equiv \left[\mathsf{T}^1_{n_b+1}(\tilde{\boldsymbol{\xi}})\right]_{13} \\ f_6(\tilde{\boldsymbol{\xi}}) \equiv \left[\mathsf{T}^1_{n_b+1}(\tilde{\boldsymbol{\xi}})\right]_{23} \end{cases} \tag{5.9}$$

or, in a more compact form, $\boldsymbol{f}(\tilde{\boldsymbol{\xi}}) = \mathbf{0}$. If n is greater than 6 and if the system of equation described in Eq. (5.9) is non degenerate then the solutions lie on a manifold with dimension $n - 6$. If, on the contrary, the system is degenerate the solutions lie on a manifold with dimension greater than $n - 6$.

Before to proceed it is important to notice that since the degrees of freedom $\tilde{\boldsymbol{\xi}}$ can describe both spatial or angular quantities the space defined by these variables is in general dimensionally non-homogeneous. For this reason we introduce a set of n scalar multipliers λ_i chosen in such a way that every $\xi_i = \lambda_i \tilde{\xi}_i$ is dimensionless. The space defined by the rescaled variables ξ_i is now homogeneous and an appropriate metric can be defined in it by means of the usual scalar product. In the case that a set of homogeneous variables are chosen as system variables (e.g. all dihedral angles) the introduction of the scalar multipliers is unnecessary but nonetheless it turns out to be useful (see Section “Tuning of fluctuations by rescaling variables”).

The problem of finding a new configuration $\boldsymbol{\xi}$ starting from an existing $\boldsymbol{\xi_0}$ in such a way that Eq. (5.8) is satisfied can now be visualized as the problem of *moving* on an $(n-6)$-dimensional manifold embedded in an n-dimensional space. The most intuitive way to perform this non-trivial task is that of generating an intermediate configuration $\boldsymbol{\xi'}$ that may not lie on the manifold but that it is not far from it. This first step is called by other authors *pre-rotation* [BBEJ+12, BFH+13], and we will adapt to this nomenclature. Starting from $\boldsymbol{\xi'}$ it is then possible to compute a true solution with numerical methods, e.g. root finding algorithms, or analytical ones [CSWD06].

In order to simplify the description of the pre-rotation step we introduce two new quantities n_m and n_v that are respectively the dimension of the tangent space M to the manifold at $\boldsymbol{\xi_0}$ and the dimension of the orthogonal space V to the manifold at $\boldsymbol{\xi_0}$. Obviously $n_m + n_v = n$ and $n_v = 6$ if the system of Eq. (5.9) is non-degenerate. In general n_v is the number of linearly independent functions in Eq. (5.9). The procedure we use to find a basis of the tangent space to the manifold takes advantage of the implicit function theorem in order to compute the derivatives $\frac{\partial \xi_i}{\partial \xi_j}$. Indeed we consider n_m among the n variables as independent and we denote them with the subscript x. The other n_v are labeled with a subscript y and will be written as a function of $\boldsymbol{\xi_x}$. With this notation we can write a set of n_m n-dimensional vectors that span the tangent space as

$$e_{x,i} = \left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\xi_{x,i}}} \right), \tag{5.10}$$

where the derivative $\left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\xi_{x,i}}}\right)$ can be performed by computing separately the contribute of the dependent variables $\boldsymbol{\xi}_y$ and that of the independent ones $\boldsymbol{\xi}_x$. The former (in the form of an n_v-dimensional vector $\frac{\partial \boldsymbol{\xi_y}}{\partial \boldsymbol{\xi_{x,i}}}$) can be easily computed by applying the implicit function theorem

$$\frac{\partial \boldsymbol{\xi_y}}{\partial \boldsymbol{\xi_{x,i}}} = -\left(\frac{\partial \boldsymbol{f}(\tilde{\boldsymbol{\xi}}_0)}{\partial \boldsymbol{\xi_y}}\right)^{-1} \cdot \frac{\partial \boldsymbol{f}(\tilde{\boldsymbol{\xi}}_0)}{\partial \boldsymbol{\xi_{x,i}}} \tag{5.11}$$

while the latter is given by the n_m relations $\frac{\partial \boldsymbol{\xi_{x,j}}}{\partial \boldsymbol{\xi_{x,i}}} = \delta_{i,j}$. In the cases in which the matrix $\left(\frac{\partial \boldsymbol{f}(\tilde{\boldsymbol{\xi}}_0)}{\partial \boldsymbol{\xi_y}}\right)$ is not invertible it is sufficient to choose a different set of $\boldsymbol{\xi_x}$ as independent variables. Vectors $e_{x,i}$ can be orthonormalized to compute a basis $\{\hat{e}_{x,i}\}_i$ for the tangent space. The intermediate configuration $\boldsymbol{\xi'}$ can finally be computed by simply summing an arbitrary linear combination of $\hat{e}_{x,i}$ to the initial configuration $\boldsymbol{\xi_0}$.

The intermediate configuration is an *open configuration* in which the position and orientation of the last reference frame do not correspond to the target ones. We therefore adjust the coordinates on the orthogonal space V by using a root-finding algorithm to obtain the final configuration. Figure 5.3 depicts an example move in case the manifold is one-dimensional (the full example is addressed in the Section “Workout example”). The solution manifold (blue) and the initial configuration are represented. The pre-rotation step corresponds to the first update (thick red arrow) while the second step along the dotted line allows to converge back to the manifold by means of a root-finding algorithm. This is in general possible only for small enough pre-rotation steps. The brown arrows show the case when a too big pre-rotation step does not allow the procedure to converge back to the manifold.

A basis for V can be efficiently computed starting from the knowledge of a basis for M and by using *ad-hoc* algorithms (e.g. QR-decomposition algorithm). With this strategy the chain closure step usually takes few iterations of the Broyden’s root finding algorithm. Notice that it is in principle possible to use a root-finding algorithm that takes advantage of the easy-to-compute gradient $\frac{\partial \boldsymbol{f}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}$ to boost the search for the solution. However our tests have shown that the time saved by the root-finding algorithm usually does not compensate for the time necessary for computing the gradient.

The full algorithm can be summarized in the following steps:

1. Compute a basis for the tangent space M at $\boldsymbol{\xi_0}$, as well as a basis for the

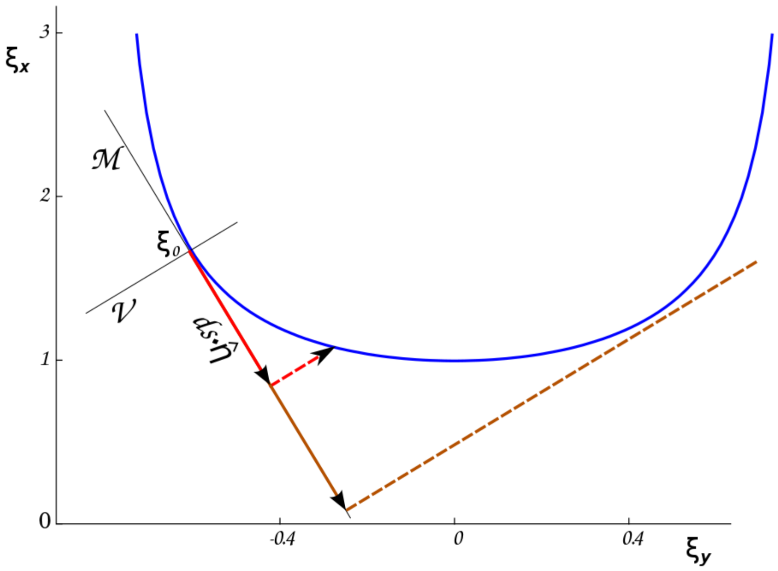


Figure 5.3: Graphical sketch of the updating of the conformations in the n-dimensional space of the degrees of freedom which are changed. The blue line represents the manifold of the solutions of Eq. (5.23), as discussed in the Section "Workout example" within "Material and Methods". The manifold is plotted within the (θ, ρ) plane, with the independent variable being $\xi_x = \rho$ and the dependent variable $\xi_y = \theta$. M and V are respectively the tangent and the orthogonal space to the manifold in the starting conformation ξ_0. The degrees of freedom are first changed along M (red continuous arrow) and a new conformation satisfying Eq. (5.9) is reached by moving along V (dashed red arrow). If the pre-rotation step along M is too large (brown continuous arrow), the post-rotation closure step along V (dashed brown arrow) fails to fall back to the manifold.

orthogonal space V;

2. Choose an arbitrary direction $\hat{\boldsymbol{\eta}}$ in M and an arbitrary step length ds;

3. Generate an intermediate configuration $\boldsymbol{\xi'} = \boldsymbol{\xi_0} + ds \cdot \hat{\boldsymbol{\eta}}$;

4. Use a root-finding algorithm in order to converge to a solution of Eq. (5.9) by moving on V;

5.1.3 Detailed balance

In this section we will denote the three-dimensional configuration of an N-atom chain at time t with the $3N$-dimensional vector $R^t = \{r_1^t, r_2^t, \cdots r_N^t\}$, where each $r_k^t = (x_k^t, y_k^t, z_k^t)^T$ represents the three-dimensional position of atom k at time t.

In order to demonstrate that the scheme we proposed for changing the backbone configuration satisfies the detailed balance condition

$$P(R^t)\Pi(R^t \to R^{t+\Delta t}) = P(R^{t+\Delta t})\Pi(R^{t+\Delta t} \to R^t\} \tag{5.12}$$

we will show separately that (1) with an appropriate change of variables, it is possible to uniformly sample the whole configurational space of the system without changes to the algorithm and (2) the probability of moving from R^t to $R^{t+\Delta t}$ can be chosen in such a way that $\Pi(R^t \to R^{t+\Delta t}) = \Pi(R^{t+\Delta t} \to R^t)$.

We start by showing that point (1) is true. If the DH's variables $\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{r}$ are used for describing the chain, the volume element in the configurational space

$$dV^t = \prod_{k=1}^{N} dV_k^t = \prod_{k=1}^{N} dx_k^t dy_k^t dz_k^t \tag{5.13}$$

can be rewritten as

$$dV^t = \prod_{k=1}^{N} J_k^t d\alpha_k^t d\theta_k^t dr_k^t, \tag{5.14}$$

where $J_k^t = \sin(\alpha_k^t)\,(r_k^t)^2$. Since the determinant $J^s = \prod_{k=1}^{N} J_k^s$ of the change of variables $R_t \to \{\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{r}\}$ is not constant, a uniform sampling of the space of the DH's variables does not result in a uniform sampling of the configurational space. In this case uniformity can be achieved by accepting the configuration change $\{\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{r}\}_1 \to \{\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{r}\}_2$ with probability $p = \min\left(1, \frac{J^{t+\Delta t}}{J^t}\right)$. Such essential choice compromises efficiency both by increasing the number of calculations per time step and by reducing the probability of obtaining a new configuration. With our approach the problem is overcome without performing any additional time-consuming calculation. Indeed it is easy to notice that the algorithm proposed in the previous section does not rely on the particular form of $\boldsymbol{\xi}$ or $\boldsymbol{f}(\tilde{\boldsymbol{\xi}})$ but rather on the possibility of computing $\boldsymbol{f}(\tilde{\boldsymbol{\xi}})$ and its derivatives. For this reason every differentiable, invertible function of $\tilde{\boldsymbol{\xi}}$, whose inverse is differentiable, can be used as degree of freedom of the system without jeopardizing the efficiency of the full scheme. If, for instance, $\boldsymbol{\psi} = g(\tilde{\boldsymbol{\xi}})$ is used as degree of freedom, the following trivial relations holds:

$$\boldsymbol{f}(\boldsymbol{\psi}) = \boldsymbol{f}(\boldsymbol{g}\left(\tilde{\boldsymbol{\xi}}\right)) \tag{5.15a}$$

$$\frac{\partial \boldsymbol{f}(\boldsymbol{\psi})}{\partial \tilde{\boldsymbol{\xi}}} = \frac{\partial \boldsymbol{f}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \frac{\partial \boldsymbol{g}(\tilde{\boldsymbol{\xi}})}{\partial \tilde{\boldsymbol{\xi}}}. \tag{5.15b}$$

If DH's parameters can be interpreted as bond lengths, bond angles and torsional angles (θ is always a torsional angle, α is a bond angle for bonds connected to the previous one, r is a bond length for bonds connected to both the previous and the subsequent one) it is sufficient to use the variables $\tilde{\boldsymbol{\xi}}_{\boldsymbol{\alpha}} = cos(\alpha)$ and $\tilde{\boldsymbol{\xi}}_{\boldsymbol{r}} = \frac{r^3}{3}$ as new degrees of freedom in order to guarantee the determinant of the Jacobian of the change of variables to be constant: this implies that the new configuration can always be accepted. This result relies only on the fact that the Jacobian does not depend on $\boldsymbol{\xi}$; we derived it independently from the values of the rescaling factors.

Notice that even if the inverse of the cosine function is not differentiable in 0 and π, these points are at the boundary of the domain in which the bond angle α is defined. Hence the analysis is valid in the open domain $(0, \pi)$ but not in its closure. The same is true for $\frac{r^3}{3}$, whose inverse is not differentiable in $r = 0$. Of course these drawbacks are not relevant for most applications, since $\alpha = 0, \pi$ and $r = 0$ are far from physical values. Therefore a uniform sampling of the space of the new variables $\left\{\tilde{\boldsymbol{\xi}}_{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\xi}}_{\boldsymbol{r}}, \theta\right\}$ is sufficient to ensure a uniform sampling of the configurational space.

In order to ensure the detailed balance condition to be valid it is now necessary to check whether the step size ds that has been used in the forward update $\xi_1 \to \xi_2$ is the same that would be necessary for the reverse transformation $\xi_2 \to \xi_1$. Indeed the step size is typically chosen from a fixed probability distribution and therefore if the backward step size ds' is different from the forward one, it is necessary to take into account the different probability of proposing such a step. The backward step can be easily computed as the norm of the projection of $\xi_2 - \xi_1$ onto the tangent space to the manifold at ξ_2. If the step size is chosen from a normal distribution with zero average and variance σ^2 it is sufficient to accept the update with probability

$$P = \min\left(1, \exp\left(\frac{ds^2 - ds'^2}{2\sigma^2}\right)\right) \tag{5.16}$$

to ensure the equivalence between forward and backward probability. In many practical cases $\frac{ds'}{ds}$ is close to one and, as a consequence, the probability of rejecting a Monte Carlo move is negligible. This calculation can be done for whatever choice of the factors λ_i. The detailed balance can therefore be satisfied independently from the choice of each λ_i.

Tuning fluctuations by rescaling variables

The role of rescaling factors λ_i previously introduced can be further exploited. As already briefly discussed, their introduction has been necessary to map the *original* non-homogeneous space of the variables $\tilde{\boldsymbol{\xi}}$ in the homogeneous one of the variables $\boldsymbol{\xi} = \lambda \cdot \tilde{\boldsymbol{\xi}}$ where a metric can be defined. But on top of that they can be used in order to tune the fluctuations of the corresponding degrees of freedom.

Given a starting configuration $\tilde{\boldsymbol{\xi}}_0$ suppose that some of the $\tilde{\boldsymbol{\xi}}_{0,i}$ are hard degrees of freedom ($\tilde{\boldsymbol{\xi}}_{0,H}$) while the others are soft ($\tilde{\boldsymbol{\xi}}_{0,S}$). We can use two different values of λ_i, depending on the class of the corresponding degrees of freedom, thus mapping

$$\tilde{\boldsymbol{\xi}}_0 = (\tilde{\boldsymbol{\xi}}_{0,S}, \tilde{\boldsymbol{\xi}}_{0,H})^T \to (\lambda_S \cdot \tilde{\boldsymbol{\xi}}_{0,S}, \lambda_H \cdot \tilde{\boldsymbol{\xi}}_{0,H})^T = \boldsymbol{\xi}_0, \tag{5.17}$$

where $\lambda_H > \lambda_S$. As described in previous sections the algorithm is applied to the initial conformation in the homogeneous, deformed space. For simplicity we consider in the following discussion the particular case in which the soft degrees of freedom correspond to the independent variables, in such a way that the new configuration ξ can be written as a function of the variation vectors $\Delta\boldsymbol{\xi}_S$ and $\Delta\boldsymbol{\xi}_H = \nabla_{\boldsymbol{\xi}_S}\boldsymbol{\xi}_H \cdot \Delta\boldsymbol{\xi}_S$ as

$$\xi = \boldsymbol{\xi}_0 + \Delta\boldsymbol{\xi} = \left(\lambda_S \cdot \tilde{\boldsymbol{\xi}}_{0,S} + \Delta\boldsymbol{\xi}_S,\ \lambda_H \cdot \tilde{\boldsymbol{\xi}}_{0,H} + \Delta\boldsymbol{\xi}_H\right)^T, \tag{5.18}$$

where $\nabla_{\boldsymbol{\xi}_S}\boldsymbol{\xi}_H$ is the matrix of the derivatives of the hard degrees of freedom with respect to the soft ones.

It is possible to express the norm of the variation vector for independent variables $\Delta\tilde{\boldsymbol{\xi}}_S$ as a function of the size of the actual move performed in our algorithm, the step size ds in the tangent space: $ds = \left(\lambda_S^2||\Delta\tilde{\boldsymbol{\xi}}_S||^2 + \lambda_H^2||\Delta\tilde{\boldsymbol{\xi}}_H||^2 \right)^{\frac{1}{2}}$. The previous equation allows to define the function $g_{\lambda_s,\lambda_H}\left(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H\right)$ through $||\Delta\tilde{\boldsymbol{\xi}}_S|| = \frac{g_{\lambda_s,\lambda_H}\left(\tilde{\boldsymbol{\xi}}_S,\tilde{\boldsymbol{\xi}}_H\right)}{\lambda_S}\ ds$. The function $g_{\lambda_s,\lambda_H}\left(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H\right)$ describes the reduction of the variation of the soft independent variables, for a fixed step size ds, due to the orientation of the tangent space with respect to the space of independent variables. It depends, in general, both on the position of the initial point in the manifold and on the chosen rescaling factors.

Each point of the deformed manifold can be remapped to the original manifold with the inverse transformation $\tilde{\boldsymbol{\xi}} = \frac{1}{\lambda_i}\boldsymbol{\xi}$. Therefore the new configuration found in Eq. (5.18) corresponds to a final configuration in the original space equal to

$$\tilde{\boldsymbol{\xi}} = \left(\tilde{\boldsymbol{\xi}}_{0,S} + \tfrac{1}{\lambda_S}\left[g_{\lambda_s,\lambda_H}\left(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H\right)\ ds\right]\ \hat{\Delta\tilde{\boldsymbol{\xi}}}_S,\ \tilde{\boldsymbol{\xi}}_{0,H} + \tfrac{1}{\lambda_H}\left[g_{\lambda_s,\lambda_H}\left(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H\right)\ ds\right]\ \nabla_{\boldsymbol{\xi}_S}\boldsymbol{\xi}_H \cdot \hat{\Delta\tilde{\boldsymbol{\xi}}}_S\right)^T, \tag{5.19}$$

where $\hat{\Delta\tilde{\boldsymbol{\xi}}}_S$ is a normalized vector.

The previous equation holds also in the $\lambda_S = \lambda_H = 1$ case, when no rescaling occurs. The rescaling factor K_S for the variation of the independent variables $\tilde{\boldsymbol{\xi}}_S$ as a consequence of the introduction of $\lambda_S, \lambda_H \neq 1$ is then easily computed:

$$K_S(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H) = \frac{1}{\lambda_S}\frac{g_{\lambda_s,\lambda_H}\left(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H\right)}{g_{1,1}\left(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H\right)} \tag{5.20}$$

while the rescaling factor K_H for the variation of the hard degrees of freedom $\tilde{\boldsymbol{\xi}}_H$ reads

$$K_H(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H) = \frac{\lambda_S}{\lambda_H}\ K_S(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H)\frac{||\nabla_{\boldsymbol{\xi}_S}\boldsymbol{\xi}_H \cdot \hat{\Delta\tilde{\boldsymbol{\xi}}}_S||}{||\nabla_{\tilde{\boldsymbol{\xi}}_S}\tilde{\boldsymbol{\xi}}_H \cdot \hat{\Delta\tilde{\boldsymbol{\xi}}}_S||}. \tag{5.21}$$

This is the central equation of this section, as it shows to what extent the variation vector for hard degrees of freedom is modified differently than for soft degrees of freedom, due to the rescaling of the original variables. Besides the global tuning factor $\frac{\lambda_S}{\lambda_H}$ a second factor $\frac{||\nabla_{\boldsymbol{\xi}_S}\boldsymbol{\xi}_H \cdot \hat{\Delta\tilde{\boldsymbol{\xi}}}_S||}{||\nabla_{\tilde{\boldsymbol{\xi}}_S}\tilde{\boldsymbol{\xi}}_H \cdot \hat{\Delta\tilde{\boldsymbol{\xi}}}_S||}$ appears, related to the local geometrical properties of the considered manifold.

For one-dimensional manifolds, it is easy to see that $\nabla_{\boldsymbol{\xi}_S}\boldsymbol{\xi}_H = \frac{\lambda_H}{\lambda_S}\nabla_{\tilde{\boldsymbol{\xi}}_S}\tilde{\boldsymbol{\xi}}_H$, so that $K_S(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H) = K_H(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H)$; the variations of both the hard and the soft degrees of freedom are rescaled in the same way, as if a new effective step size $ds(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H) = K_S(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H)\, ds$ were used.

For manifold of higher dimension, if $\nabla_{\boldsymbol{\xi}_S}\boldsymbol{\xi}_H \neq \frac{\lambda_H}{\lambda_S}\nabla_{\tilde{\boldsymbol{\xi}}_S}\tilde{\boldsymbol{\xi}}_H$ then $K_S(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H) \neq K_H(\tilde{\boldsymbol{\xi}}_S, \tilde{\boldsymbol{\xi}}_H)$ and the λ factors can effectively be used in order to tune the amplitude of the fluctuations of hard degrees of freedom with respect to the soft ones. It is obviously not easy to quantify 'a priori' the local geometric factor $\frac{||\nabla_{\boldsymbol{\xi}_S}\boldsymbol{\xi}_H \cdot \hat{\Delta\tilde{\boldsymbol{\xi}}_S}||}{||\nabla_{\tilde{\boldsymbol{\xi}}_S}\tilde{\boldsymbol{\xi}}_H \cdot \hat{\Delta\tilde{\boldsymbol{\xi}}_S}||}$.

In the most general case, with no restriction on the soft degrees of freedom being either independent or dependent variables, we expect to find qualitatively similar results both for high-dimensional and one-dimensional manifolds.

Workout example

In order to better explain each step of the algorithm we provide a simple example that can be solved exactly. Beyond the context of concerted local movements in chain molecules, the simplest possible case for our algorithm corresponds to the motion along a one-dimensional manifold defined by a single constraint within a two-dimensional space. In this spirit, we consider a physical system in which a single particle is constrained to move within the (x, y) plane on the right branch $(x \geq 1)$ of a hyperbola of equation

$$x^2 - y^2 - 1 = 0. \tag{5.22}$$

We introduce new polar coordinates $\left(\tilde{\rho} = \sqrt{x^2 + y^2}, \tilde{\theta} = \text{atan}\left(\frac{y}{x}\right)\right)$ that, in this example, will have the same role the DH parameters have for the description of a chain molecule. Notice that $\tilde{\theta}$ is defined in the open interval $(-\frac{\pi}{4}, \frac{\pi}{4})$. We also introduce two scalar quantities λ_ρ and λ_θ defined in such a way that $\rho = \lambda_\rho\tilde{\rho}$ and $\theta = \lambda_\theta\tilde{\theta}$ are both dimensionless quantities. Constraints are defined by the analogous of Eq. (5.9)

$$f_1(\tilde{\rho}, \tilde{\theta}) : \tilde{\rho}^2 \cos(2\tilde{\theta}) - 1 = 0 \tag{5.23}$$

that, as a function of the rescaled variables becomes

$$f_1(\rho, \theta) : \frac{\rho^2}{\lambda_\rho^2} \cos(2\frac{\theta}{\lambda_\theta}) - 1 = 0. \tag{5.24}$$

The tangent space M and the orthogonal space V in the initial configuration (ρ_0, θ_0) can be easily computed starting from the expression of the derivatives of $f_1(\rho, \theta)$

$$\begin{cases} \frac{\partial f_1(\rho,\theta)}{\partial \rho} = 2\frac{\rho}{\lambda_\rho^2}\cos(2\frac{\theta}{\lambda_\theta}) \\ \frac{\partial f_1(\rho,\theta)}{\partial \theta} = -2\frac{\rho^2}{\lambda_\rho^2\lambda_\theta}\sin(2\frac{\theta}{\lambda_\theta}) \end{cases} \tag{5.25}$$

by means of the implicit function theorem. Indeed, if the initial configuration is such that $\frac{\partial f_1(\rho,\theta)}{\partial \theta} \neq 0$ (i.e. $\theta \neq 0$) it is possible to choose ρ as independent variable and to compute the derivative $\frac{\partial \theta}{\partial \rho} = \frac{\lambda_\theta}{\rho}\,\mathrm{cotan}\left(2\frac{\theta}{\lambda_\theta}\right)$. The tangent space M is therefore generated by the vector $\eta(\rho,\theta) = \left(1, \frac{\partial \theta}{\partial \rho}\right)^T$ and any intermediate configuration can be selected as $(\rho', \theta') = (\rho_0, \theta_0) + ds \cdot \hat{\eta}(\rho_0, \theta_0)$, where ds is an arbitrary step-size and $\hat{\eta}$ is normalized. Once η is computed, the QR algorithm is used in order to generate a basis for the orthogonal space V. In this simple example V is one-dimensional and the generating vector can be computed explicitly as $\eta^\perp(\rho,\theta) = \left(-\frac{\partial \theta}{\partial \rho}, 1\right)^T$. A root finding algorithm is finally used in order to find a solution to the equation $f_1(\rho,\theta) = 0$ with $(\rho,\theta) = (\rho', \theta') + k \cdot \hat{\eta}^\perp$ and with varying k. In case $\theta = 0$, it is not possible to use ρ as the independent variable, and θ has to be chosen instead. Notice that in order to ensure numerical stability it is safer to use θ as independent variable in a suitable interval centered in $\theta = 0$. Also, the existence of a solution is in general ensured only for small enough step sizes, as depicted in Fig. 5.3.

We now consider within this example the effect of introducing the rescaling factors λ_ρ and λ_θ, as discussed in the previous section. We assume $\tilde{\theta} \neq 0$, so that ρ can be used as the independent variable. In this case we see that

$$\frac{\partial \theta}{\partial \rho} = \frac{\lambda_\theta}{\lambda_\rho}\frac{\partial \tilde{\theta}}{\partial \tilde{\rho}} \tag{5.26}$$

and therefore any rescaling induced for $d\rho$ is applied to $d\theta$ as well, as expected for a one-dimensional manifold. The relation between the step size ds tangent to the manifold and the variation $d\tilde{\rho}$ of the independent unrescaled variable is $ds = d\tilde{\rho}\left[\lambda_\rho^2 + \lambda_\theta^2\left(\frac{\partial \tilde{\theta}}{\partial \tilde{\rho}}\right)^2\right]^{\frac{1}{2}}$, that allows to recover the function $g_{\lambda_\rho,\lambda_\theta}\left(\tilde{\rho}, \tilde{\theta}\right) = \frac{\lambda_\rho}{ds/d\tilde{\rho}} = \left[1 + \frac{\lambda_\theta^2}{\lambda_\rho^2}\frac{\mathrm{cotan}^2(2\tilde{\theta})}{\tilde{\rho}^2}\right]^{-\frac{1}{2}}$. We finally obtain the rescaling function from

Eq. (5.20) as

$$K(\tilde{\rho},\tilde{\theta}) = \frac{1}{\lambda_\rho}\frac{\left[1+\frac{1}{\tilde{\rho}^2}\operatorname{cotan}^2\left(2\tilde{\theta}\right)\right]^{\frac{1}{2}}}{\left[1+\frac{\lambda_\theta^2}{\lambda_\rho^2\tilde{\rho}^2}\operatorname{cotan}^2\left(2\tilde{\theta}\right)\right]^{\frac{1}{2}}}\,.$$

Results

In this section we present some applications of our method for the study of polypetide molecular systems, preceded by a test of detailed balance and of how the fluctuations of different degrees of freedom can be tuned by using rescaled variables. In polypeptide chains the rules of quantum chemistry constrain bond lengths and bond angles to fluctuate slightly around known values and double bonds to be approximately planar. The flexibility of the chain is therefore mainly due to the variation of ϕ, ψ Ramachandran's angles. In order to mimic this behavior, hard degrees of freedom are typically strongly constrained by using stiff quadratic potentials. This is not necessary with our approach because the hard degrees of freedom can be kept frozen in their minimum energy value thus reducing the number of degrees of freedom.

For this reason, in all applications discussed below we consider only the Ramachandran's ϕ and ψ torsional angles as degrees of freedom of the system; since ω is kept fixed, the set of bonds considered in the DH description is disconnected, as shown in Fig. 5.1. In practice, in our simulations Eq. (5.9) is always not degenerate, in such a way that the solution manifold has co-dimension 6.

Detailed balance

In order to verify that the detailed balance is satisfied we performed a Monte Carlo simulation on a 63-residue long protein fragment (pdb code 1CTF) by using our algorithm. We allowed only the Ramachandran's ϕ and ψ angles to vary during the simulation; these modifications were obtained by randomly selecting either a Pivot move around a randomly selected bond or our locally concerted move on a set of randomly selected angles. The introduction of the Pivot move was necessary to move the last bond of the chain and thus ensure simulation ergodicity. Figure 5.4 shows the distribution of the values of different torsional angles of a protein chain that has been simulated using the proposed schema. As expected when detailed balance holds, the distribution is

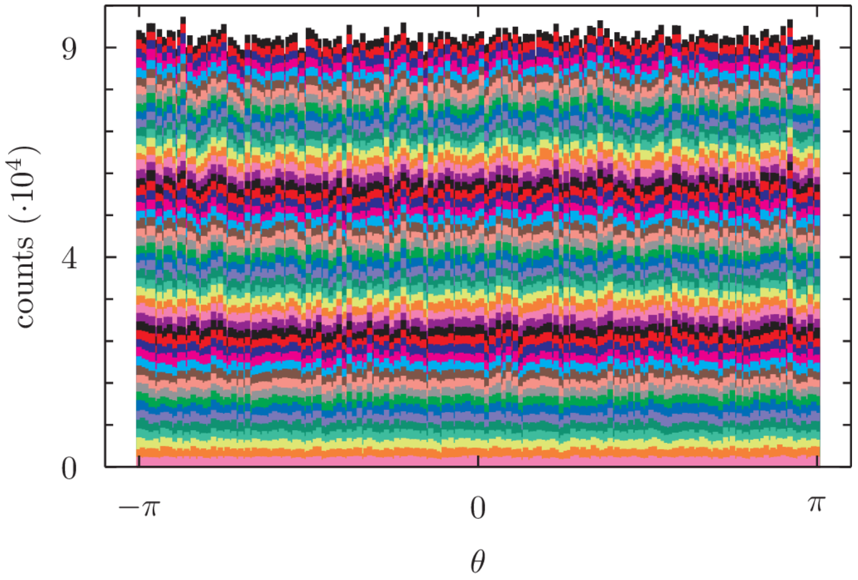


Figure 5.4: Distribution of the values of torsional angles of a 67 residues long protein (1CTF) during a $3.5 \cdot 10^5$ time-steps Monte-Carlo simulation. Each of the stacked barchart is relative to a different torsional angle. At each simulation step the structure is deformed by applying our algorithm to a randomly chosen portion of the chain (with probability 0.7) or by a pivot move (with probability 0.3). In the former case the step size ds is chosen from a normal distribution with mean 0 rad and variance 0.08 rad^2. In the latter case a randomly chosen ϕ or ψ angle is perturbed by adding to it a quantity that is chosen from a normal distribution with mean 0 rad and variance 0.4 rad^2.

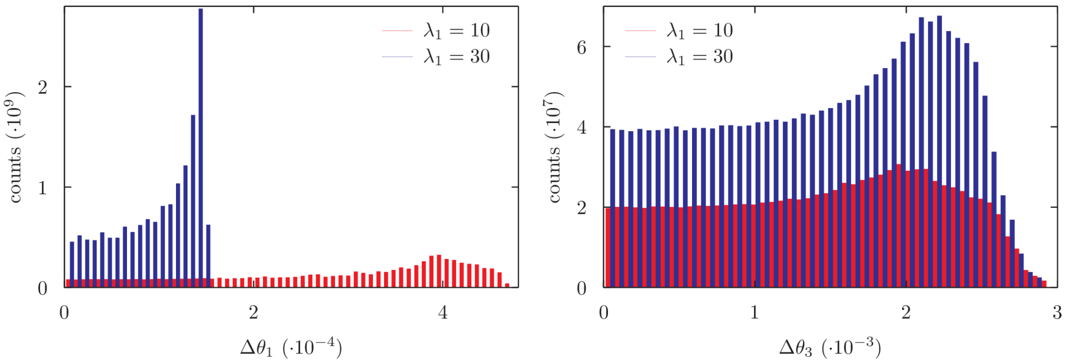


Figure 5.5: Distribution of variability of θ_1 and θ_3 angles for different values of λ_1 for a 9 degrees of freedom chain. The distribution of θ_3 is largely not affected by the change of λ_1 while the distribution of θ_1 is rescaled according toEq. (5.19).

flat, within natural stochastic fluctuations.

Fluctuation tuning by rigidity rescaling

We also studied how the dynamics of the exploration of the solution manifold is affected by the rigidity factors λ_i used to rescale the original variables (see Materials and Methods). This has been done by comparing the distribution of the fluctuations $\Delta\theta$ of two different torsional angles upon changing the rigidity of one of the two. The data have been obtained by performing short simulations with fixed step-size $ds = 0.01$ along a randomly chosen direction onto the tangent space to the manifold. First a nine degrees of freedom fragment, corresponding to a three-dimensional manifold, was explored for different values of λ_1. Figure 5.5 shows the distribution of the fluctuations of the corresponding torsional angle θ_1 and of another torsional angle (θ_3) used as negative control. A rescaling of λ_1 by a factor of 3 does not globally affect the distribution of $\Delta\theta_3$, but rescales the corresponding distribution of $\Delta\theta_1$ by a factor roughly 3. This shows that the main role in Eq. (5.19) is played by the global rescaling factor $1/\lambda$, wheres the role of local manifold geometry is in practice less relevant, at least in the considered example. On the other hand, the effect of local manifold geometry may explain some reshaping that can be appreciated in the plotted distributions apart from the global rescaling.

Similarly, Fig. 5.6 shows the distribution of the fluctuations of the angles θ_1 and θ_3 obtained while exploring the solution manifold of a seven degrees of freedom fragment with a fixed step-size ds for different values of λ_1. In this case the solution manifold is one-dimensional and therefore the fluctuations of θ_1 and θ_3 are always related. Indeed, after the rigidity λ_1 is increased,

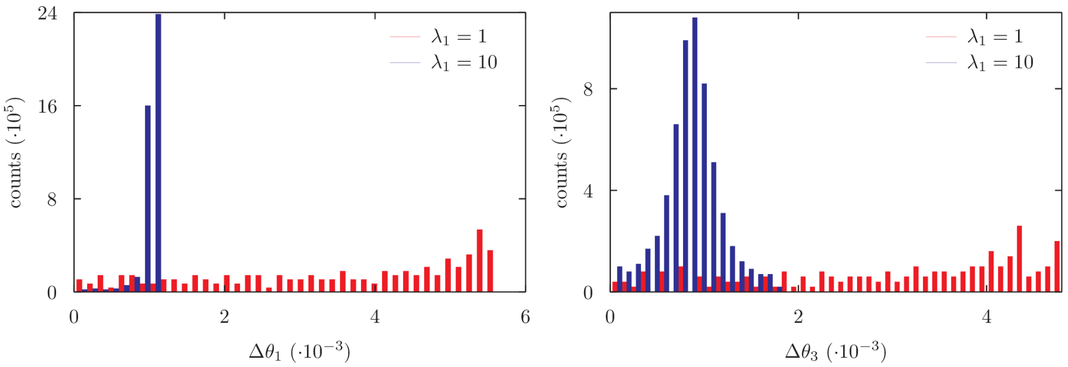


Figure 5.6: Distribution of variability of θ_1 and θ_3 angles for different values of λ_1 for a 7 degrees of freedom chain. In this case it is not possible to change a single dof without altering the others: as a consequence a rescaling of λ_1 affects the distributions of θ_1 as well as of the other angles (shown: θ_3). In this case the effect of changing the rigidity of one or more parameters is the same as rescaling the step size used during the simulation.

both distributions are globally rescaled by the same quantity. The dynamics on the deformed manifold corresponds, in this case, to a dynamics on the original manifold with a rescaled step size. For the one-dimensional example as well, some reshaping can be seen in the plotted distributions apart from the global rescaling. Again, this observation may be explained as an effect of local manifold geometry, consistently with Eq. (5.19).

Exploring the conformational space

The first application we describe is the exploration of the whole conformational space of a small portion of a polypeptide, i.e. finding all configurations of that fragment that are compatible with the locality constraints. The problem can be resolved by finding a method to compute all the solutions of Eq. (5.9). This is not a simple task when the number of degrees of freedom is large but, on the contrary, if the number of variables is 7 the exploration of the manifold is simple. In this case the manifold has dimension 1 and so it is sufficient to move on the manifold always along the same direction. This can be done by choosing at each step k an initial direction $\hat{\boldsymbol{\eta}}_k$ in such a way that $\hat{\boldsymbol{\eta}}_{k-1} \cdot \hat{\boldsymbol{\eta}}_k > 0$. In practice, we choose $\hat{\boldsymbol{\eta}}_k$ to be parallel to the projection of $\hat{\boldsymbol{\eta}}_{k-1}$ onto the tangent space to the manifold in the actual configuration.

Figure 5.7 shows a three-dimensional projection of a one-dimensional manifold obtained by changing 7 consecutive ϕ and ψ Ramachandran's angles of a portion of a protein. Some configurations of the polypeptide that have been gen-

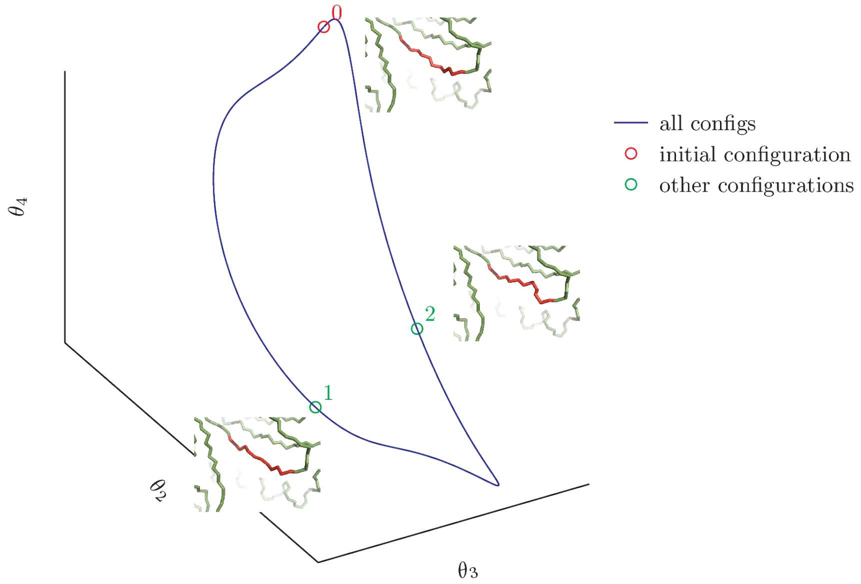


Figure 5.7: Three-dimensional projection of the solution manifold. The whole manifold lies in a seven-dimensional space. Some configurations are highlighted: the starting configuration is emphasized with a red circle while two other possible solutions are emphasized in green. The red part of the structure is the portion which has been modified with our moves.

erated during the exploration are also plotted. There are no other configurations of the selected region that are compatible with the constraints imposed by the locality requirement and that can be generated with a continuous modification of the original configuration.

Higher dimensional manifolds can be explored as well but, in these cases, there is not a general strategy that allows an efficient exploration of the whole space. This exploration can be achieved with a Monte-Carlo simulation or with *ad hoc* procedure as in the next example. As a proof of concept, with the only purpose of showing the viability of the method in higher dimensions, we consider a small cyclic molecule (cyclooctatetraene) and we assume that all its 8 torsional angles are soft degrees of freedom that can be modified. Figure 5.8 shows four conformations of cyclooctatetraene obtained with our procedure and Fig. 5.9 shows the whole bidimensional solution manifold.

We highlight the efficiency of the method. The full exploration of cyclooctatetraene configurational space has been performed in about 5 minutes on a single core 2.0 Ghz computer, collecting more than 10^5 different structures with a sample-step of 10^{-2} radians.

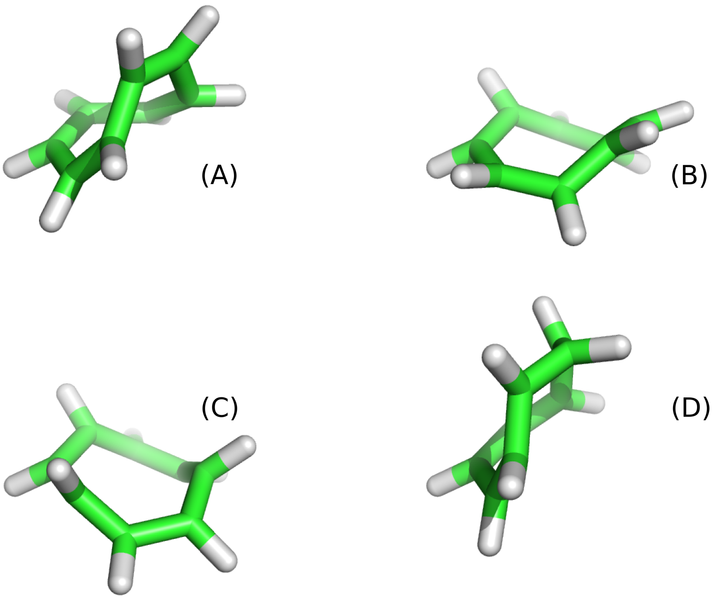


Figure 5.8: Four conformations of the ciclic molecule cyclooctatetraene obtained while exploring its whole conformational space by changing the 8 torsional angles degrees of freedom. Bond length and bond angles are kept fixed. This is a toy model to illustrate the efficiency of the method: the molecule alternates double and single bonds and therefore half of the torsional angles are constrained and only four are completely free. Images of molecular structures have been generated with PyMol [Sch10].

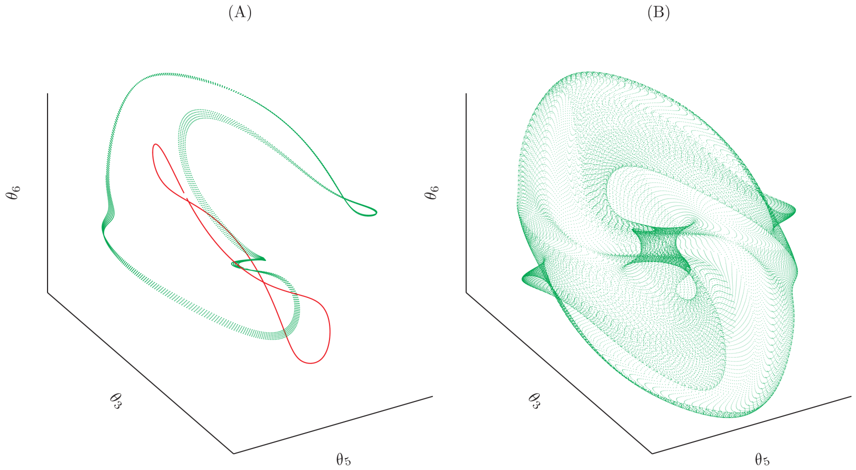


Figure 5.9: A three-dimensional projection of the solution manifold for our model of cyclooctotetraene molecule (B) and a schematics of how the manifold has been explored (A). First a set of seven angles are chosen and the relative one-dimensional manifold is visited (red line). Then the one-dimensional manifold relative to a different choice of degrees of freedom is explored using each of the generated structure as a starting point (green dots).

5.1.4 Protein backbone mobility

In this section we describe how it is possible to estimate the *mobility* of a portion of protein backbone (*local backbone mobility*) by using a simple schema based on the algorithm proposed here. The hypothesis we use is that the local mobility is proportional to the number of configurations that can be explored locally without modifying the rest of the chain, the *local backbone volume*. In principle, the number of configurations taken into account in this counting could be reduced by eliminating those conformations that exhibit steric clashes. Also, it would be possible to introduce a pair-wise potential in order to consider interaction effects. Here we limit the study of the mobility to non-interacting chains.

Consider the $3N$-dimensional configurational space describing the position of each atom composing the system. In analogy with the usual definition of entropy, we define the *local backbone entropy* as the logarithm of the local backbone volume, that is the volume of the solution manifold measured in the $3N$-dimensional configurational space. This volume can be computed by integrating the Gramian of the transformation $\{\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{r}\} \to R^t$ on the solution manifold in the DH's variable space. The Gramian function specifies how the $(n-6)$-dimensional volume element of the manifold in the space of DH parameters is mapped into the $3N$-dimensional configurational space. If $\boldsymbol{s}$ are the coordinates on the manifold, determined within an orthonormal system of basis vectors in the tangent space, the Jacobian of the transformation can be written as $\nabla_{\boldsymbol{s}} R^t$, and the Gramian is computed as

$$G(\boldsymbol{s}) = \{det\left[(\nabla_{\boldsymbol{s}} R^t)^T \, \nabla_{\boldsymbol{s}} R^t\right]\}^{\frac{1}{2}}, \tag{5.27}$$

(see [Ros80, Rud64]).

Values of entropy can be assigned to different protein fragments, i.e. to different subsets of degrees of freedom, by considering the corresponding manifolds defined by their concerted variations. We first estimate the entropy $S(i)$ associated to a single residue i as the sum of the entropies computed for all the different fragments comprising that residue. For simplicity, we restrict our calculation to fragments that comprise seven consecutive ϕ and ψ Ramachandran's angles. We then call the map $i \mapsto A \times \exp\left(S(i)\right)$ the *mobility profile* of the structure. The factor A is a proportionality constant which has the dimension of an Angstrom squared.

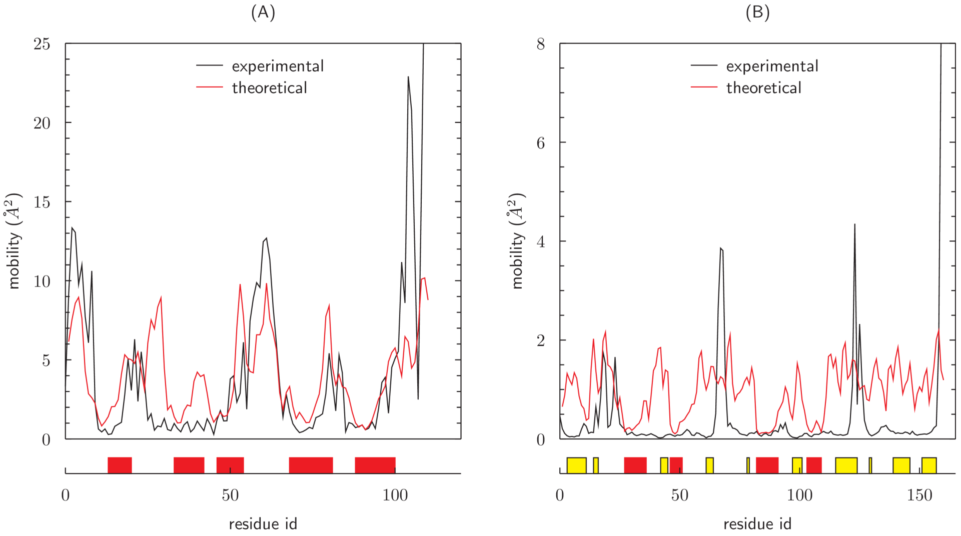


Figure 5.10: Comparison between the experimental mobility and the theoretical *mobility profile* of different structures. Data in (A) are relative to an all-α protein (1YGM) while (B) shows the mobility of mixed β and α protein (2ITH). The experimental profile has been computed as the variance in the position of α-carbons in different *NMR* models of the same protein. Red lines represent theoretical calculations, black lines experimental values. Boxes below the horizontal axis locate the position of secondary structures along the chain: red for α-helices and yellow for β-sheet. The matching between the theorethical and experimental profiles has been obtained by optimizing the proportionality constant A with a least square fit. In (A) the fit has been computed over the whole length of the chain, while in (B) the fit concerned only the α regions.

In Fig. 5.10, we compare the *mobility profile* with the corresponding experimental data, i.e. the variance of the positions of the α carbon atoms in different NMR models of the same structure. Note that the predicted mobility profile is matched to the experimental one by fitting A, so that only the ratio between the mobilities of different regions of the same protein structure can be considered as a real prediction of our model.

Remarkably, the mobility of 1YGM, an *all-α structure*, is described with a good degree of accuracy by the theoretical estimations. This fact, confirmed by similar analysis on other *all-α structures* (data not shown) suggests that the local geometrical constraints of the protein backbone taken into account by our method are enough to predict the relative mobility of helical and non helical regions. The surprising conclusion is that the presence of both steric and energetic effects reduces the available phase space, and hence the mobility, by the same amount for both regions. On the contrary, β-sheet mobility is not captured at all, probably because we do not consider in our analysis the non-local inter-strand interactions that are crucial for their stability.

Structure refinement

The ability of exploring completely the conformation space of a limited fragment of a protein can be exploited for reconstructing or refining a small portion of a polypeptide structure. Let us suppose to have a putative fragment of a protein which has been roughly reconstructed by experimental methods: the hard degrees of freedom (bond angles and bond lengths) of this portion have correct values, whereas some of the torsional angles may not all be consistent with the allowed values of the Ramachandran's plot. Given such initial configuration Γ_0, our approach allows to modify exhaustively the soft degrees of freedom and to check if it is feasible to obtain a solution which is compatible with the standard Ramachandran's plot.

To test this possibility, we first partition the Ramachandran's plot in a grid of squares with size $2^\circ \times 2^\circ$. By analyzing the databank Top500 formed by a non-redundant, specially refined set of 500 high resolution X-ray crystallographic structures of globular proteins [LDA+03], we consider as *good*, those bins which have a fractional occupancy higher than 0.04. We investigate if the condition of having *good* Ramachandran's angles in a small fragment of a protein is a condition sufficient to reconstruct/refine the protein backbone. Therefore we randomly explored the conformational space of different portions of a protein. For each portion we store only the *acceptable* configurations, i.e. those configurations with good Ramachandran's angles.

Different 5 residues long fragments were analyzed, from α, β, or coil structures (that are neither in α nor in β). Figures 5.11 and 5.12 show all the acceptable configurations that have been generated during the exploration of an α-helix and a β-strand portion, respectively. Each subplot shows the values sampled by a different pair of Ramachandran's angles within the considered fragment. From Fig. 5.11, we can notice that all acceptable solutions generated from an α-helix structure lie in a very narrow region of the Ramachandran's plot. The exploration of different α-helix regions confirms this finding and shows that usually about the 90% of the configurations that are compatible with the imposed constraints are acceptable, and hence shown in Fig. 5.11.

Going back to the original problem of refining a rough initial estimation of a protein portion, our results imply that, for an α-helix, it is easy to reconstruct a solution that is essentially unique, within very small changes of Ramachandran's

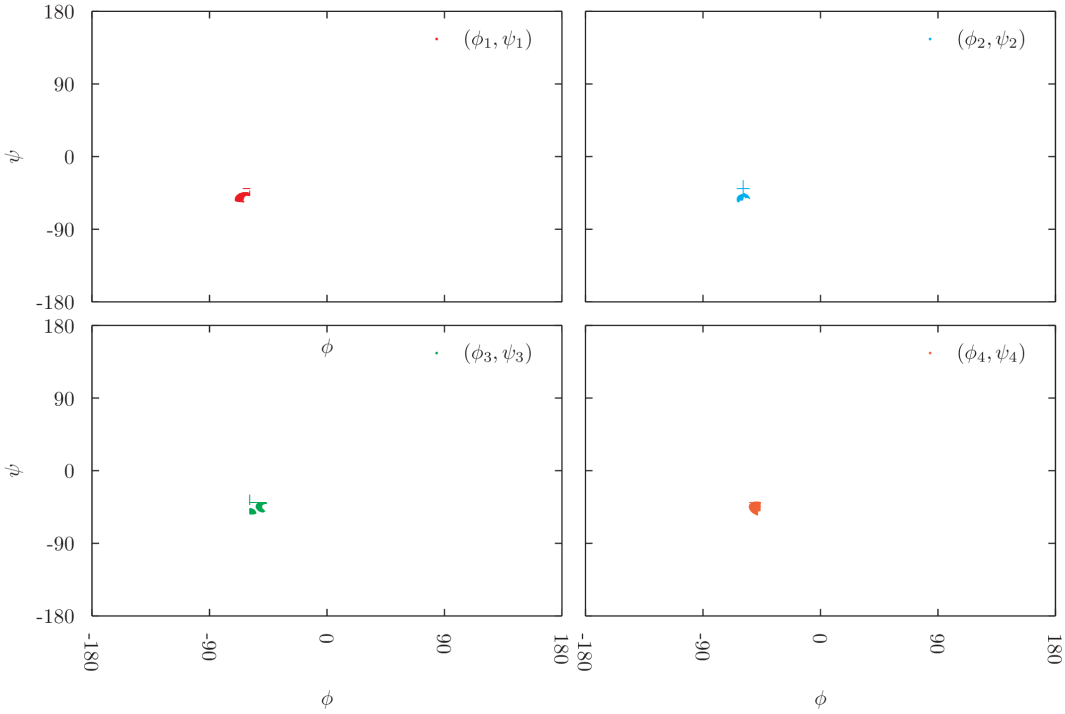


Figure 5.11: All acceptable Ramachandran angles obtained during the exploration of the solution manifold of a 5-residue-long α fragment. All points form a unique cluster, meaning that there is only one acceptable conformation of the helix that is compatible with the initial configuration. Thus, having fixed the first and the last bond, there is only one possible way to reconstruct the missing helix.

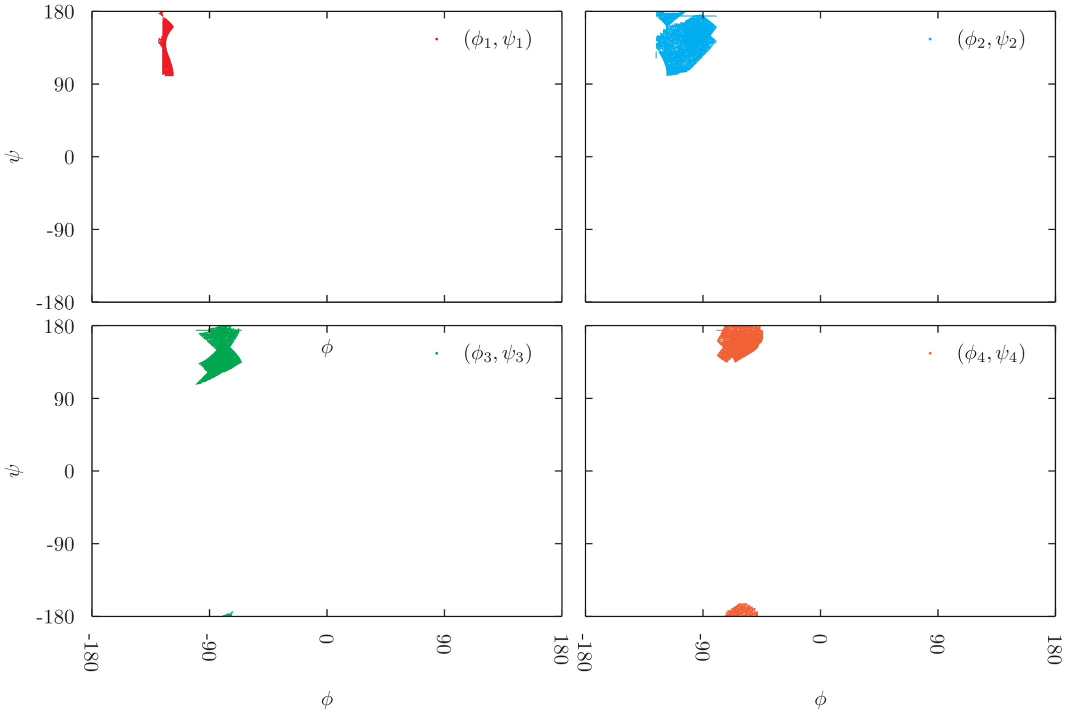


Figure 5.12: All acceptable Ramachandran angles obtained during the exploration of the solution manifold of a 5-residue-long β fragment. Points form different clusters, meaning that there are more than one acceptable conformation of the strand that are compatible with the initial configuration. It is not possible to safely reconstruct or refine a β fragment by only requiring the final configuration to have acceptable Ramachandran's angles.

angles. Results are different in the case of β-structure; in such situation the algorithm works smoothly to find solutions but most of them are not in good regions of the Ramachandran's plot (more than 90%) and those satisfying such constraints, the ones shown in Fig. 5.12, are more widely distributed in the good region. Therefore, a reconstruction starting by a β-like structure is feasible within our approach, but produces a broader range of possible solutions with respect to α-helical fragments. For both the α and the β fragments, in order to evaluate how exhaustive is our exploration strategy, the stored configurations were analyzed with a sophisticated cluster algorithm [RL14], checking that the conformational clusters found in each half of the search trajectories do not change.

Repeating the same analysis for coil fragments, it is found that only a very low percentage (usually below 1%) of configurations have good Ramachandran's angles, whereas the acceptable configurations span all possible allowed values (Figure not shown). Under such circumstances, an exhaustive search of all possible acceptable solutions can be quite time consuming.

5.2 Discussion and Conclusions

We introduced a novel technique that can be used to locally deform linear or cyclic polymer chain structures. Different kinds of degrees of freedom (torsional angles, bond lengths, bond angles or any arbitrary combination of these) can be used. There are no general requirements on the choice of the degrees of freedom: in the case of protein chains they can belong to the same residue as well as to residues that are far away from each other. In the general case, for the algorithm to work, it is necessary that at least seven different degrees of freedom are used.

The three-dimensional configuration of a linear chain is commonly described by using the cartesian coordinates of each monomer with respect to the same fixed frame of reference. Instead, in our algorithm we use the Denavit-Hartenberg (DH) convention [HD55, HD64], that is very popular in robotics and has already been used by different authors [GS70, BK85, MZW95, CSJD04, NOS05] in order to describe a polypeptide chain. One advantage of the DH convention is its ability to readily describe a general disconnected subset of the bonds of the physical chain. This is useful if one is interested in varying in a concerted way degrees of freedom from disconnected chain segments, while keeping fixed the degrees of freedom in between. This occurs, for instance, when consecutive ϕ, ψ Ramachandran's angles are chosen to be varied, whereas the torsional angle ω around the peptide bond is kept fixed [Din00]; in this case it is not necessary to include the peptide bond in the DH description (see Fig. 5.1).

In the simplest case, when all bonds included in the DH description are connected with each other, the DH variables have a well defined physical meaning. Since two consecutive bonds always share an atom, link offsets $\boldsymbol{d}$ are zero, and therefore we can interpret link lengths $\boldsymbol{r}$ as *bond lengths*, link twists $\boldsymbol{\pi} - \boldsymbol{\alpha}$ as *bond angles*, and joint angles $\boldsymbol{\theta}$ as *torsional angles*. In the general case, some physical bonds may not be considered so that two consecutive bonds included in the DH description do not share an atom. When a disconnected bond is added, the link offset is different from zero, the link length and the link twist do not have a physical interpretation anymore, whereas the joint angle can still be interpreted as a torsional angle, albeit with an offset in its definition. As a consequence, torsional angles can always be included within the DH description as such, bond angles only if the previous bond in the DH description is not disconnected, and bond lengths if both the previous and the subsequent bond are not disconnected.

In general, the change of a single DH parameter is responsible for a global modification of the structure, i.e. a modification that, on average, affects a number of atoms proportional to the number of atoms of the chain. For instance the change $\theta_i \to \theta_i + \Delta\theta_i$ is responsible for a rotation around $\hat{z}_{i-1}$ of all the atoms of the chain that belong to the bonds labeled with $j > i$; this configuration change is known as a pivot move. In order to locally update a configuration $\{\boldsymbol{r}_0,\ \boldsymbol{\alpha}_0,\ \boldsymbol{d}_0,\ \boldsymbol{\theta}_0\}$ it is necessary to simultaneously change more than one DH parameter with the costraint that the remaining part of the chain is kept fixed. By using the DH description we were able to write a set of equations (see Eq. (5.9)) that describe the constraints and whose solutions correspond to the locally deformed configurations we want to compute. This set of equations is the central mathematical framework on which several other methods are based on [GS70, BK85, MZW95, CSJD04, NOS05]. The solutions derived by existing approaches are however limited to considering only torsional angles. Moreover, in order to compute a solution of the equations, existing techniques usually restrict their application to the study of particular geometries, such as the ideal Pauling-Corey geometry, or need to slightly modify a number of other degrees of freedom of the chain.

Our method generalizes existing algorithms by proposing a strategy that allows for a concerted modification of any arbitrary set of degrees of freedom of the chain while keeping *all* the other strictly fixed. The only requirement is that the number of selected degrees of freedom is greater than the number of linearly independent equations that define the constraints in Eq. (5.9), i.e. is greater than 6 in the non degenerate case.

All algorithms performing concerted local structural changes for a polymer chain roughly follow the same general strategy. First, a *pre-rotation* step is proposed, that is an update of a selected subset of *driver* pre-rotation angles among all the ones that will be eventually involved in the local move. If actually performed, the pre-rotation would generate an intermediate configuration $\boldsymbol{\xi}'$ corresponding to a global structural change. Then, the *post-rotation* step is performed, by explicitly finding the remaining post-rotation angles that satisfy the locality constraints.

The algorithm here proposed introduces a novel way to perform the pre-rotation step (see Fig. 5.3). Indeed, while other methods arbitrarily selects the driver angles and then generate the intermediate configuration $\boldsymbol{\xi}'$ by randomly perturbing them, our algorithm generates $\boldsymbol{\xi}'$ by moving from the initial configuration along a random direction in the tangent space to the manifold of the configurations compatible with the locality constraints. Thus, the pre-rotation step is already a change of all the degrees of freedom involved in the local move concerted in a way that is intrinsically driven by the geometrical properties of the manifold of explorable configurations.

The post-rotation step is then performed by using a root-finding algorithm to converge again to the manifold of correct configurations. Despite its simplicity, the root-finding approach is effective since it takes advantage of the fact that the intermediate configuration $\boldsymbol{\xi}'$ is already a good approximation to the correct solution as well as of the fact we can restrict the root-finding algorithm to search for a solution by moving within the space orthogonal to the manifold at the initial configuration. The orthogonal space can be efficiently computed based on the knowledge of the tangent space already needed in the pre-rotation step. Restricting the search of the solution within the orthogonal space also ensures that the solution searched for in the post-rotation step is unique, for small enough pre-rotation moves, providing at the same time a simple way to compute the probability of the backward transition and thus to enforce detailed balance in a Monte Carlo simulation.

The possibility to numerically compute the derivatives as in Eq. (5.11) is not only useful to determine the basis vectors for the tangent space to the manifold of chain configurations compatible with the locality constraints, but also to obtain any directional derivative on it of scalar functions that depend on chain configuration such as, for instance, potential energy functions. Notably, 300-400 concerted moves per second can be performed on a single core 2.0 Ghz processor in the present implementation. This makes the efficiency of our general numerical methodology not so distant from the one reported in [MCK09], 2000 loop closure solutions per second, with an analytic-based strategy that relies on a specific choice of the torsional angles to be modified.

Importantly, once the difference between forward and backward probabilities is taken into account, the usage of orthonormal basis vectors in both the tangent and orthogonal spaces ensures that the space of DH parameters involved in the local move is sampled uniformly in a Monte Carlo simulation, if no other reweighting is employed in the acceptance/rejection test. This is shown in Fig. 5.4 for the simple case of DH parameters that correspond to torsional angles, that are indeed expected to display a uniform distribution at equilibrium in the absence of any interaction. Thus, at variance with existing algorithms, we do not need to perform, for further reweighting, the time-consuming calculation of the Jacobian factor due to the solution of the post-rotation step [UJ03].

For DH parameters that correspond to bond lengths and bond angles, that are not expected to have uniform distributions at equilibrium, no reweighting is again needed, provided that the locality constraints and the corresponding manifold of possible chain configurations are defined in terms of simply modified variables that are instead expected to be uniformly sampled.

Moreover, a simple rescaling of uniformly sampled variables used in the local move maintains their sampling properties while allowing to tune the relative fluctuations of the non rescaled variables. This could be useful in dealing with polypeptide chains, when variables originally related to bond lengths and bond angles are expected to fluctuate much less than torsional angles. The need of further reweigthing is again avoided due to the proper exploitation of the intrinsic geometrical properties of the manifold of correct configurations, as defined in terms of the rescaled variables. However, as shown in Fig. 5.5, the tuning of relative fluctuations by means of variable rescaling is possible only for manifolds with dimension at least two. Moreover, Fig. 5.5 shows that the effect the rescaling of one variable induces on the other variables, through the coupling with local manifold geometry, is in practice negligible. As a consequence, we may expect that the number of variables whose fluctuations can be independently tuned by rescaling is given by the manifold dimension minus one. Consistently, as shown in Fig. 5.6, for unidimensional manifolds the relative fluctuations of the different variables involved in the local move are in itself dictated by manifold geometry and are not affected by rescaling.

Instead, if DH parameters that do not have a physical interpretation are involved in the local move, it is not possible to go simply back to the case of uniform sampling. It would be necessary to take into account the expected non uniform sampling of all bond lengths and bond angles related to the unphysical DH parameters by proper reweighting factors.

We showed different possible applications of the technique that we introduced.

A first application is an efficient scheme to explore the whole configurational space of small fragments of a polypeptide backbone or of other chain structures that are compatible with locality constraints. We demonstrated the concept, first, in the simplest case of a protein fragment where the degrees of freedom involved in the local move are 7 consecutive ϕ, ψ Ramachandran's angles along the polypeptide backbone. In that case, all possible configurations to be explored lie on a one-dimensional manifold embedded in a periodic 7-dimensional space, whose projection in a 3-dimensional space is shown in Fig. 5.7. Since our technique relies on the computation of the tangent space to the manifold, in order to define the pre-rotation step, it is both straightforward and efficient to stride along the manifold following the same direction until the starting configuration is revisited.

For higher-dimensional manifolds the systematic exploration of all possible configurations is a difficult task. We employed the cyclo-octotetraene cyclic molecule as a toy model, by considering all its 8 torsional angles as degrees of freedom. The manifold of possible configurations (in the special case of a cyclic closed chain these are *all* possible configurations) is a bidimensional one in an 8-dimensional space. We used the strategy of dividing the exploration in separate one-dimensional trajectories that are tracked along the same direction until the initial configuration is recovered, as in the previous case. The whole manifold can be recovered in this way, by changing the initial configuration and the subset of 7 torsional angles to be varied along a given one-dimensional trajectory. The resulting manifold is shown as a projection in a 3-dimensional space in Fig. 5.9 and some representative conformations are shown in Fig. 5.8. For higher-dimensional manifolds, the usage of more sophisticated sampling techniques, such as generalized ensemble Monte Carlo methods [WL01] or metadynamics [LP02], may well be more efficient.

It is important to observe that our technique relies on the previous knowledge of a configuration already compatible with the locality constraints, and the loop closure problem is solved only at a local level, in the post-rotation stage (see Fig. 5.3). In the classic loop closure problem, instead, one is given the task of reconstructing 'ab initio' a missing portion of a linear chain. It is then easy to do it by using 'wrong' values for, say, one bond length and one bond angle. The hard problem of finding a configuration with the 'right' values, then, could be in principle recast, within the framework of our technique, as the problem of the searching for a subset of configurations with the 'right' values within a manifold suitably chosen where the bond length and the bond angle to be fixed are among the degrees of freedom that are allowed to change. While the actual implementation of the above sketched strategy is beyond the scope of the present thesis, it provides a context where the exploration efficiency demonstrated by our technique could prove extremely useful.

The other applications of our technique that we investigated are more directly related to the local distortion properties of protein chain structures, when only backbone heavy atoms are considered.

First, we introduced the notion of *local backbone volume*, as the volume spanned in the $3N$-dimensional Cartesian space by all configurations that can be adopted by a protein segment, compatibly with the locality constraints, as the degrees of freedom involved are changed on the corresponding manifold. As a consequence, the local backbone volume may strongly depend on the choices made for both the constraints and the degrees of freedom. Locality constraints can differ a lot depending, for example, on the secondary structure content of the considered protein chain segment.

It is natural to relate the volume at disposal for local concerted movements to the mobility of residues, so that the higher the local volume the more mobile the residues. Indeed, the local backbone volume computed for different protein chains, in the simplest case of 7 consecutive ϕ, ψ Ramachandran's angles, allows to easily recognize α-helices as the most locally rigid portions of proteins. In the case of an all-α helical protein, a quantitatively good matching can be performed between the local volume profile and the residue mobility profile resulting by different NMR structural models of the same protein (see Fig. 5.10). This is a non trivial result, since the local backbone volume is a geometrical feature that does not take into account any interaction or excluded volume effect. Consistently, β-strands are not identified as locally rigid segments in our approach (see Fig. 5.10), since, at variance with α-helices, they need to be stabilized by hydrogen bonding to a nearby strand.

Second, we tried to assess the following point: How many realistic protein structures do exist that are compatible with a given locality constraint? This is a central issue in structure refinement, when the task is often to improve over an existing non realistic configuration of a protein segment. In practice, we start from a real protein segment configuration and we simply use our technique to perform a thorough exploration of the manifold of possible solutions compatible with the locality constraints. We look for standard values of the Ramachandran's angles ϕ and ψ to filter realistic structures. Again, the locality constraint, and thus the answer to the raised question, crucially depend on the secondary structure of the chosen protein segment.

We use a state-of-the-art cluster analysis to make sure that the exploration of the manifold was completed, when no new clusters are observed. Consistently with the local backbone volume analysis, we observe that if the initial segment has an α-helical structure, most of generated configurations are realistic, the latter are all α-helical ones and span a very narrow region in the Ramachandran's plot (see Fig. 5.11). If the initial segment has a β-strand structure, the fraction of generated configurations decreases, whereas all realistic configurations found are in the β-strand region of the Ramachandran's plot, spanning a wider region (see Fig. 5.12). If the initial segment has no secondary structure, the generated configurations essentially span the whole Ramachandran's plot and the fraction of realistic conformations is very small. Based on pure geometrical properties, a helical structure is quite easily refined, a strand segment is less easily refined, a loop coil region is not quite easily refined.

Our aim here is to show how the efficiency of our local exploration technique can be easily employed in the context of protein structure refinement; systematic results could be obtained by testing, within the same approach, initial segments as helix or strand ends, or hairpin turns. More importantly, a bias can be easily incorporated in the sampling of the manifold of possible solutions, according to a potential energy function, or to a general scoring function, or to a measure of consistency with known experimental data, such as electron density maps for X-rays diffraction experiment on protein crystals.

Chapter 6

Conclusion

Proteins are the most important biological macro-molecules together with nucleic acids. They perform a vast array of functions inside the cell, spanning from enzymatic activity to signal transduction. Moreover therapeutic proteins have recently gained a predominant importance in the field of medical treatments and the ever increasing number of protein-based medical products suggests that this role is still only in its infancy.
In this perspective research and development of physical-based methodologies to be applied to the study, modeling and understanding of protein structure and protein-protein interactions can successfully boost the advancement in the field of medicine and pharmaceutics.

Aim of this thesis is to evaluate the feasibility of new computational methodologies to be applied to the study of binding processes between two proteins or between a protein and a small ligand.
Three main aspects have been considered of particular importance for this problem and have been examined:

- monomer-solvent interactions
- entropic contributions due to the change of accessible configurational space upon binding
- conformational modifications upon-binding.

The focus has been put on the estimation of binding affinities between two monomers (Chapter 3) and on the development of a technique which allows to explore 'local' conformations of polypeptide chains and that could be most useful when considering conformational modifications upon binding (Chapter 5).

In Chapter 4 we summarized some general considerations on statistical methods in protein physics (KBPs) and tried to treasure them for the development of two novel KBP, presented in the same chapter.

The thesis can therefore be divided into two macro-parts, the first focused on KBPs (Chapters 2 to 4) and the second focused on a new method for the local deformation of polypeptide chains (Chapter 5).

As testing ground for benchmarking new methods in the field of KBPs we employed BACH, a simple KBP presented in [CGL+12] that was already proved to be able to recognize the protein native state. We introduced BACH in Section 2.2.

The method we devised to locally modify polypeptide chains is instead inspired by the world of robotic hand manipulators and constitutes a generalization of many existing methods, briefly described in the introduction to Chapter 5.

6.1 Summary of the thesis

Protein-solvent interactions: determining the surface of a macromolecule BACH classify each residue depending on its status of either solvent-exposed or solvent-buried. We devised two new methodologies that allow this classification to be more reliable and efficient with respect to the one originally implemented in [CGL+12].

The first algorithm implemented is a modified version of the LCPO algorithm [WSS99] and has been published in [SZC+13]. The working principle on which it is based is to compute the SASA of the residue by removing from the sum of the whole surface contribution of each atom the estimated overlap of the surfaces of nearby atoms: residues with values of SASA greater than zero are considered exposed, or buried otherwise. This new implementation is faster and more flexible than the original one; moreover the performances of this upgraded version of BACH show an improvement over the standard version.

The second method we implemented is based on the computation of the α-shape of macromolecules. The α-shape of a set of points (in this context identified by the coordinates of all heavy atoms of the system) describes the piece-wise linear surface which enclose those points that cannot be *touched* by a sphere of radius $\alpha^{\frac{1}{2}}$ without superposing it to some other atoms. This algorithm shows three major improvement over the modified LCPO one. First of all, by comparing the solvent-exposure classification obtained by an analytic calculation (provided by the *getArea* server) with the ones obtained with our algorithm, we conclude that the classification suggested by the α-shape algorithm is twice more accurate than the modified LCPO one, and therefore more reliable. Moreover by renouncing to explicitly compute SASA, alpha-shape computation is faster and also has a better scaling behavior on the molecule dimension ($N \log N$ in the case of α-shape vs. N^2 in the case of LCPO).

The improved accuracy and the faster execution time can have great impact in applications in which the surface of large molecules has to be computed many times, e.g. the case in which a KBP has to be applied to all frames of a molecular dynamics.

Protein-protein interactions: the role of entropy in the estimation of binding affinities In the context of protein-protein interaction, devising numerical procedures for estimating binding affinities between associating protein sub-units is of great importance: such techniques could indeed guide biologists in the development of new protein drugs specifically designed to bind a target protein (or molecule) without the need of performing a vast number of long and onerous experiments.

In Chapter 3 we tried to estimate the binding affinity of different pairs of proteins for which an experimental value is available from the literature. But rather than using an *ad-hoc* KBP or a machine-learn approach (both very popular) we employed a BACH-like scoring function [SGS$^+$15] that estimates the free-energy contribution due to the formation of new inter-residue contacts in the interface between proteins. This first term has been combined with the entropic contribution, in turn constituted by three terms that account for translational, rotational and vibrational contributions. The first two entropic contributions have been computed exactly from the three-dimensional structures of the monomers and of the complex. The vibrational term instead has been estimated by employing an ANM whose elastic constant has been determined by matching the mobility profile obtained with the network model to the one computed from a short MD trajectory. We compared the experimental binding affinities of the studied complexes with a linear combination of the interface and entropic terms. The coefficient of the linear combination have been computed by minimizing the errors between experimental and theoretical results on a set of 12 complexes (train set) while the final benchmark has been done on a different set of 15 complexes.

Results indicate the role of entropic contribution to be fundamental to correctly describe experimental data. Indeed, even if the performance of the proposed method are slightly worse than that of other state-of-the-art methods, the improvement we obtained by correctly accounting for the entropic contributions is significant.

Although our method for the estimation of the binding affinity of two proteins is not of easy practical use, due to the need of running expensive MD simulations, we conclude that rotational, translational and vibrational entropic contributions are fundamental for predicting binding affinities. We also deduced that the vibrational term can be estimated with a sufficient accuracy by employing an ANM.

Inside KBPs: some considerations An important step in the development of new KBPs was that of understanding the limitation of existing methods and, in particular, the limitations of the approach used in BACH.
We therefore proceeded in deriving with a maximum-likelihood approach the inverse-Boltzmann-like formula commonly employed in the definition of KBPs. This (trivial) derivation had the benefit of highlighting all the hypothesis that are necessary to justify the usage of the formula. Probably the most evident hypothesis is to consider the chain as an ensamble of pointless and free particles (residues); correlations naturally due to the connectivity of the chain are indeed not taken into account. With the aid of a lattice HP model we showed that this fact can result in a considerable error in the estimation of the parameters involved in the KBP.

Development and testing of two new KBPs We developed two different KBPs with the aim of improving the performances of BACH in the recognition of protein native states and evaluating the importance of correlations in real proteins.
The first KBP we implemented is basically a modified version of BACH in which the interacting units are defined as small functional groups, with a well-defined chemical behavior, rather than as residues. The benefit of this improvement is that interactions are described in a more specific fashion. Results are encouraging: the performance is the same as BACH even if the involved parameters are roughly half those of BACH. This algorithm, referred as mBACH in the main text, employed the α-shape technique for classifying each functional group as buried or exposed to the solvent. As in BACH, correlations are not taken into account.

The second KBP we proposed is specifically designed to consider the correlations along the chain, but renounce to describe inter-residue interaction in any specific way. We indeed consider the probability distribution of distances between α carbons of the same protein; if the separation along the chain of the corresponding residues is inside a specific range (described in [BHM05]) we expect this probability to be approximately gaussian. The basic idea of the KBP proposed, and named Gaussian Chain Potential (GCP), is that of considering the fluctuations from the expected gaussian distribution as due to the effect of physical pairwise potentials, that we were able to deduce. The obtained potentials exhibit distance-dependent behaviors that are residue-specific.

Locally constrained backbone conformations We analyzed the problem of the local movements of a chain molecule where a small subset of degrees of freedom, e.g. dihedral angles, bonds angles or bond lengths, are concertedly modified inside a specific portion of the chain, in such a way that *only* the atoms in the selected region are moved while all the others are fixed. The possibility of determining such locally constrained conformations can be fundamental in describing the conformational changes of the interface between two interacting proteins but has also proved to be extremely valuable in Monte Carlo (MC) simulations. The method we proposed is very general and is devised in such a way that it can be implemented in a MC simulation without the need of a *Jacobian reweighting*, which is usually necessary to ensure the detailed balance condition.

We tested the algorithm on different applications and we showed that:

- the configurational space of backbone fragments that comprises 7 free degrees of freedom can be efficiently and completely explored; higher dimensional configurational space can be explored as well with MC techniques;

- the configurational space of a small ring with 8 free torsional angles can be efficiently and completely explored. The technique used in this case can be implemented in principle for exploring the configurational space of an 8 degrees of freedom fragment, but in this case the completeness of the exploration is not guaranteed;

- the mobility profile of *all-α* proteins is well described by the volume (in the configurational space) that is accessible to each residue;

- the fluctuation of each free degree-of-freedom can be tuned by a simple rescaling of the variable involved in the algorithm;
- detailed balance is satisfied by the algorithm without the need of additional (and usually time consuming) calculations;

Moreover we suggested the possibility of employing the proposed method for loop modeling problems.
The work described in Chapter 5 has been accepted for publication after peer-reviewing process in PlosOne [ZRST].

6.2 Future perspectives

The role of entropy for the estimation of the binding affinity (described in Chapter 3) is promising and should be examined more in depth. In particular the estimation of the vibrational term is not applicable in real-life scenarios because it requires three MD simulations to be run (one for each monomer and one for the complex): a new method that do not depend on simulations is therefore needed, for correctly estimating the rescaled elastic constant to be used in the elastic network model. Moreover, one could think of more refined elastic network models, e.g. by introducing different kinds of elastic constants. In this way, one could get a better matching with the MD mobility profile and, as a consequence, a better estimate of the vibrational term.
The new KBPs proposed in Chapter 4 are still in their embryonic form and need more tests. But mBACH has already proved to be at least as valuable as BACH despite the highly reduced number of parameters it employs. Moreover the usage of mBACH, being based on functional groups rather than amino acid residues, is not limited to proteins but could be extended to other molecules, thus paving the way to many applications in pharmaceutics (for drug design) and more generally in medicine.
The ideas underlying the development of the GCP can potentially lead to the development of a KBP in which parameters are not biased by correlations but, at the moment, it fails in capturing the specific interactions occurring between pair of residues. The enhancements necessary for KBP improvement should therefore be focused on describing with more specificity pairwise interactions. For instance the potential could be parameterized not only by the $C_\alpha - C_\alpha$ distance but also by the mutual orientation of the chain at the corresponding

residues; or a functional-group-based version of the GCP could also be devised. The algorithm we devised for generating local perturbations of polypeptide chains has a vast array of possible applications, from its implementation in efficient MC moves to the estimation of protein chain mobility. Moreover techniques employed in Section 5.1.4 can be extended to the more general case of loop modeling, in which the problem of reconstructing a delimited portion of a polymeric chain is considered. But the application that could be more relevant in the field of protein-binding is that of binding-site modeling and prediction. Indeed binding sites usually comprise a low number (~ 10) of residues and are therefore the ideal target for the application of the algorithm if backbone torsional angles are the only free degrees of freedom considered.

Bibliography

[AAO12] Ramu Anandakrishnan, Boris Aguilar, and Alexey V Onufriev. H++ 3.0: automating pk prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids research*, 40(W1):W537–W541, 2012.

[ADJ$^+$01] AR Atilgan, SR Durell, RL Jernigan, MC Demirel, O Keskin, and I Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

[AGM$^+$90] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[AR06] Patrick Aloy and Robert B Russell. Structural systems biology: modelling protein interactions. *Nature Reviews Molecular Cell Biology*, 7(3):188–197, 2006.

[AS07] Joseph Audie and Suzanne Scarlata. A novel empirical free energy function that explains and predicts protein–protein binding affinities. *Biophysical chemistry*, 129(2):198–211, 2007.

[AW04] Michelle R Arkin and James A Wells. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature reviews Drug discovery*, 3(4):301–317, 2004.

[BAC$^+$98] Axel T Brünger, Paul D Adams, Marius G Clore, Warren L DeLano, Piet Gros, Ralf W Grosse-Kunstleve, Jian-Sheng Jiang, John Kuszewski, Michael Nilges, Navraj S Pannu, Randy J Read, Luke M Rice, Thomas Simonson, and Gregory L Warren. Crystallography nmr system: A new software suite for macromolecular structure determination. *Acta Crystallographica D*, 54:905–921, 1998.

[Bak00] David Baker. A surprising simplicity to protein folding. *Nature*, 405(6782):39–42, 2000.

[BB97] Fausto Bernardini and Chandrajit L Bajaj. Sampling and reconstructing manifolds using alpha-shapes. 1997.

[BB01] DW Bolen and Ilia V Baskakov. The osmophobic effect: natural selection of a thermodynamic force in protein folding. *Journal of molecular biology*, 310(5):955–963, 2001.

[BBEJ$^+$12] Sandro Bottaro, Wouter Boomsma, Kristoffer E. Johansson, Christian Andreetta, Thomas Hamelryck, and Jesper Ferkinghoff-Borg. Subtle monte carlo updates in dense molecular systems. *Journal of Chemical Theory and Computation*, 8(2):695–702, 2012.

[BBL$^+$95] James W Bryson, Stephen F Betz, Helen S Lu, Daniel J Suich, Hongxing X Zhou, Karyn T O'Neil, and William F DeGrado. Protein design: a hierarchic approach. *Science*, 270(5238):935–941, 1995.

[BD94] Sarina Bromberg and Ken A Dill. Side-chain entropy and packing in proteins. *protein Science*, 3(7):997–1009, 1994.

[BEJ$^+$99] Ivet Bahar, Burak Erman, Robert L Jernigan, Ali Rana Atilgan, and David G Covell. Collective motions in hiv-1 reverse transcriptase: examination of flexibility and enzyme function. *Journal of molecular biology*, 285(3):1023–1037, 1999.

[Bet05] Marcos R Betancourt. Efficient monte carlo trial moves for polypeptide simulations. *The Journal of chemical physics*, 123(17):174905–174905, 2005.

[BFH$^+$13] Wouter Boomsma, Jes Frellsen, Tim Harder, Sandro Bottaro, Kristoffer E Johansson, Pengfei Tian, Kasper Stovgaard, Christian Andreetta, Simon Olsson, Jan B Valentin, et al. Phaistos: A framework for markov chain monte carlo simulation and inference of protein structure. *Journal of computational chemistry*, 34(19):1697–1705, 2013.

[BGM13] Somendra M Bhattacharjee, Achille Giacometti, and Amos Maritan. Flory theory for polymers. *arXiv preprint arXiv:1308.2414*, 2013.

[BGT00] Igor N Berezovsky, Alexander Y Grosberg, and Edward N Trifonov. Closed loops of nearly standard size: common basic element of protein structure. *Febs Letters*, 466(2):283–286, 2000.

[BHM05] Jayanth R Banavar, Trinh Xuan Hoang, and Amos Maritan. Proteins and polymers. *The Journal of chemical physics*, 122(23):234910, 2005.

[BHN03] Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. *Nature Structural* M*olecular Biology*, 10(12):980–980, 2003.

[BHPS61] Yehoshua Bar-Hillel, Micha Perles, and Eliahu Shamir. On formal properties of simple phrase-structure grammars. *STUF-Language Typology and Universals*, 14(1-4):143–172, 1961.

[BJ99] Ivet Bahar and Robert L Jernigan. Cooperative fluctuations and subunit communication in tryptophan synthase. *Biochemistry*, 38(12):3478–3490, 1999.

[BK85] Robert E. Bruccoleri and Martin Karplus. Chain closure with bond angle variations. *Macromolecules*, 18(12):2767–2773, 1985.

[BK05] Luciano Brocchieri and Samuel Karlin. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic acids research*, 33(10):3390–3400, 2005.

[BNL03] T Andrew Binkowski, Shapor Naghibzadeh, and Jie Liang. Castp: computed atlas of surface topography of proteins. *Nucleic Acids Research*, 31(13):3352–3355, 2003.

[BR03] Matthew J Betts and Robert B Russell. Amino acid properties and consequences of substitutions. *Bioinformatics for geneticists*, 317:289, 2003.

[BTS08] Pascal Benkert, Silvio CE Tosatto, and Dietmar Schomburg. Qmean: A comprehensive scoring function for model quality assessment. *Proteins: Structure, Function, and Bioinformatics*, 71(1):261–277, 2008.

[BWF$^+$00] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[BYY$^+$11] Hongjun Bai, Kun Yang, Daqi Yu, Changsheng Zhang, Fangjin Chen, and Luhua Lai. Predicting kinetic constants of protein–protein interactions based on structural properties. *Proteins: Structure, Function, and Bioinformatics*, 79(3):720–734, 2011.

[C$^+$70] Francis Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[CAH$^+$09] Vincent B Chen, W Bryan Arendall, Jeffrey J Headd, Daniel A Keedy, Robert M Immormino, Gary J Kapral, Laura W Murray, Jane S Richardson, and David C Richardson. Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 66(1):12–21, 2009.

[CD06] Fabrizio Chiti and Christopher M Dobson. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, 75:333–366, 2006.

[cga] CGAL, Computational Geometry Algorithms Library. http://www.cgal.org.

[CGL$^+$12] Pilar Cossio, Daniele Granata, Alessandro Laio, Flavio Seno, and Antonio Trovato. A simple and efficient statistical potential for scoring ensembles of protein structures. *Scientific Reports*, 2, 2012.

[Cho56] Noam Chomsky. Three models for the description of language. *Information Theory, IRE Transactions on*, 2(3):113–124, 1956.

[CL97] Robin W Carrell and David A Lomas. Conformational disease. *The Lancet*, 350(9071):134–138, 1997.

[Con83] Michael L Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612):709–713, 1983.

[CSJD04] Evangelos A. Coutsias, Chaok Seok, Matthew P. Jacobson, and Ken A. Dill. A kinematic view of loop closure. *Journal of Computational Chemistry*, 25(4):510–528, 2004.

[CSWD06] Evangelos A Coutsias, Chaok Seok, Michael J Wester, and Ken A Dill. Resultants and loop closure. *International Journal of Quantum Chemistry*, 106(1):176–189, 2006.

[CT93] Carlos J Camacho and D Thirumalai. Kinetics and thermodynamics of folding in model proteins. *Proceedings of the National Academy of Sciences*, 90(13):6369–6372, 1993.

[CWB06] Asher D Cutter, James D Wasmuth, and Mark L Blaxter. The evolution of biased codon and amino acid usage in nematode genomes. *Molecular biology and evolution*, 23(12):2303–2315, 2006.

[DBT93] L.R. Dodd, T.D. Boone, and D.N. Theodorou. A concerted rotation algorithm for atomistic monte carlo simulation of polymer melts and glasses. *Molecular Physics*, 78(4):961–996, 1993.

[DCJ90] Jacques Des Cloizeaux and Gérard Jannink. *Polymers in solution: their modelling and structure*, volume 4. Clarendon Press Oxford, 1990.

[Del09] Daniele Dell'Orco. Fast predictions of thermodynamics and kinetics of protein–protein recognition from structures: from molecular design to systems biology. *Molecular BioSystems*, 5(4):323–334, 2009.

[Dil85] Ken A Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6):1501–1509, 1985.

[Dil90] Ken A Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.

[Din00] AAron R Dinner. Local deformations of polymers with nonplanar rigid main-chain internal coordinates. *Journal of Computational Chemistry*, 21(13):1132–1144, 2000.

[DM97] Bassil I Dahiyat and Stephen L Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–87, 1997.

[DWL89] William F DeGrado, Zelda R Wasserman, and James D Lear. Protein design, a minimalist approach. *Science*, 243(4891):622–628, 1989.

[DWL06] Qiwen Dong, Xiaolong Wang, and Lei Lin. Novel knowledge-based mean force potential at the profile level. *BMC bioinformatics*, 7(1):324, 2006.

[Edg10] Robert C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[EKS83] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *Information Theory, IEEE Transactions on*, 29(4):551–559, 1983.

[FB98] Robert Fraczkiewicz and Werner Braun. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of Computational Chemistry*, 19(3):319–333, 1998.

[FIS01] Giorgio Favrin, Anders Irbäck, and Fredrik Sjunnesson. Monte carlo update for chain molecules: Biased gaussian steps in torsional space. *The Journal of Chemical Physics*, 114(18):8154–8158, 2001.

[FM09] Marwan Fayed and Hussein T Mouftah. Localised alpha-shape computations for boundary recognition in sensor networks. *Ad Hoc Networks*, 7(6):1259–1269, 2009.

[FS06] Qiaojun Fang and David Shortle. Protein refolding< i> in silico</i> with atom-based statistical potentials and conformational search using a simple genetic algorithm. *Journal of molecular biology*, 359(5):1456–1467, 2006.

[FSW91] Hans Frauenfelder, Stephen G Sligar, and Peter G Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[FV+69] Paul Flory, M Volkenstein, et al. *Statistical mechanics of chain molecules.* Wiley Online Library, 1969.

[FVM07] Evandro Ferrada, Ismael A Vergara, and Francisco Melo. A knowledge-based potential with an accurate description of local interactions improves discrimination between native and near-native protein conformations. *Cell biochemistry and biophysics*, 49(2):111–124, 2007.

[GG07] Steven Goodman and Sander Greenland. Assessing the unreliability of the medical literature: a response to "why most published research findings are false". *Johns Hopkins University, Dept. of Biostatistics Working Papers.*, 2007.

[GMF+05] John C Gordon, Jonathan B Myers, Timothy Folta, Valia Shoja, Lenwood S Heath, and Alexey Onufriev. H++: a server for estimating pkas and adding missing hydrogens to macromolecules. *Nucleic acids research*, 33(suppl 2):W368–W371, 2005.

[GMW97] P Güntert, C Mumenthaler, and K Wütrich. Torsion angle dynamics for nmr structure calculation with the new program dyana. *Journal of Molecular Biology*, 273:283–298, 1997.

[GS70] Nobuhiro Go and Harold A Scheraga. Ring closure and local conformational deformations of chain molecules. *Macromolecules*, 3(2):178–187, 1970.

[GZ07] Michael K Gilson and Huan-Xiang Zhou. Calculation of protein-ligand binding affinities*. *Annual review of biophysics and biomolecular structure*, 36(1):21, 2007.

[HD55] Richard Scheunemann Hartenberg and Jacques Denavit. A kinematic notation for lower-pair mechanisms based on matrices in trans asme. *J. Appl. Mech*, pages 215–221, 1955.

[HD64] Richard Scheunemann Hartenberg and Jacques Denavit. *Kinematic synthesis of linkages.* McGraw-Hill, 1964.

[HH92] Wen-Miin Hwang and Yii-Wen Hwang. Computer-aided structural synthesis of planar kinematic chains with simple joints. *Mechanism and Machine Theory*, 27(2):189–199, 1992.

[HHL+08] Y. J. Huang, D. Hang, L. J. Lu, L. Tong, M. B. Gerstein, and G. T. Montelione. Targeting the human cancer pathway protein interaction network by structural genomics. *Mol. Cell Proteomics*, 7(10):2048–2060, Oct 2008.

[HHS88] Winnfried Hasel, Thomas F Hendrickson, and W Clark Still. A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Computer Methodology*, 1(2):103–116, 1988.

[HHS$^+$05] Shura Hayryan, Chin-Kun Hu, Jaroslav Skřivánek, Edik Hayryane, and Imrich Pokornŷ. A new analytical method for computing solvent-accessible surface area of macromolecules and its gradients. *Journal of computational chemistry*, 26(4):334–343, 2005.

[HK96] Daniel Hoffmann and Ernst-Walter Knapp. Polypeptide folding with off-lattice monte carlo dynamics: the method. *European Biophysics Journal*, 24(6):387–403, 1996.

[HKL09] Julia Handl, Joshua Knowles, and Simon C Lovell. Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. *Bioinformatics*, 25(10):1271–1279, 2009.

[HKVDSL08] Berk Hess, Carsten Kutzner, David Van Der Spoel, and Erik Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation*, 4(3):435–447, 2008.

[HL92] Nancy Horton and Mitchell Lewis. Calculation of the free energy of association for protein complexes. *Protein Science*, 1(1):169–181, 1992.

[HPT$^+$98] Pehr B Harbury, Joseph J Plecs, Bruce Tidor, Tom Alber, and Peter S Kim. High-resolution protein design with backbone freedom. *Science*, 282(5393):1462–1467, 1998.

[HS96] Liisa Holm and Chris Sander. Mapping the protein universe. *Science*, 273(5275):595–602, 1996.

[HTB03] Bosco K Ho, Annick Thomas, and Robert Brasseur. Revisiting the ramachandran plot: Hard-sphere repulsion, electrostatics, and h-bonding in the α-helix. *Protein Science*, 12(11):2508–2522, 2003.

[Ioa05] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[Jac06] Meyer B Jackson. *Molecular and cellular biophysics*. Cambridge University Press, 2006.

[JBC08] Joël Janin, Ranjit P Bahadur, and Pinak Chakrabarti. Protein–protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(02):133–180, 2008.

[JGM+02] Lin Jiang, Ying Gao, Fenglou Mao, Zhijie Liu, and Luhua Lai. Potential of mean force for protein–protein interaction studies. *Proteins: Structure, Function, and Bioinformatics*, 46(2):190–196, 2002.

[JH03] Gareth M Jenkins and Edward C Holmes. The extent of codon usage bias in human rna viruses and its evolutionary origin. *Virus research*, 92(1):1–7, 2003.

[JT96] Susan Jones and Janet M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 1996.

[JTT92] David T Jones, WR Taylort, and Janet M Thornton. A new approach to protein fold recognition. 1992.

[JVR93] A Jain, N Vaidehi, and G Rodriguez. A fast recursive algorithm for molecular dynamics simulation. *Journal of Computational Physics*, 106(2):258–268, 1993.

[KDI+03] Brian Kuhlman, Gautam Dantas, Gregory C Ireton, Gabriele Varani, Barry L Stoddard, and David Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, 2003.

[KG99] Akio Kitao and Nobuhiro Go. Investigating protein dynamics in collective coordinate space. *Current opinion in structural biology*, 9(2):164–169, 1999.

[KMH+11] Panagiotis L Kastritis, Iain H Moal, Howook Hwang, Zhiping Weng, Paul A Bates, Alexandre MJJ Bonvin, and Joël Janin. A structure-based benchmark for protein–protein binding affinity. *Protein Science*, 20(3):482–491, 2011.

[Kol93] Peter Kollman. Free energy calculations: applications to chemical and biochemical phenomena. *Chemical reviews*, 93(7):2395–2417, 1993.

[KPK+09] Carol Kilkenny, Nick Parsons, Ed Kadyszewski, Michael FW Festing, Innes C Cuthill, Derek Fry, Jane Hutton, and Douglas G Altman. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PloS one*, 4(11):e7824, 2009.

[KS83] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[KVS+10] Manoj Kumar, Shikha Verma, Sujata Sharma, Alagiri Srinivasan, Tej P Singh, and Punit Kaur. Structure-based in silico design of ahigh-affinity dipeptide inhibitor for novel protein drug target shikimate kinase of mycobacterium tuberculosis. *Chemical biology* d*rug design*, 76(3):277–284, 2010.

[KZV10] Petras J Kundrotas, Zhengwei Zhu, and Ilya A Vakser. Gwidd: genome-wide protein docking database. *Nucleic acids research*, 38(suppl 1):D513–D517, 2010.

[Lan98] Rae Langton. Kantian humility: Our ignorance of things in themselves. 1998.

[LBG04] Rhonald Lua, Alexander L Borovinskiy, and Alexander Yu Grosberg. Fractal and statistical properties of large compact polymers: a computational study. *Polymer*, 45(2):717–731, 2004.

[LBG08] Benjamin Leader, Quentin J Baca, and David E Golan. Protein therapeutics: a summary and pharmacological classification. *Nature Reviews Drug Discovery*, 7(1):21–39, 2008.

[LCOS08] Adam Liwo, Cezary Czaplewski, Stanisław Ołdziej, and Harold A Scheraga. Computational techniques for efficient conformational sampling of proteins. *Current Opinion in Structural Biology*, 18:134–139, 2008.

[LDA+03] Simon C. Lovell, Ian W. Davis, W. Bryan Arendall, Paul I. W. de Bakker, J. Michael Word, Michael G. Prisant, Jane S. Richardson, and David C. Richardson. Structure validation by cα geometry: ϕ,ψ and cβ deviation. *Proteins: Structure, Function, and Bioinformatics*, 50(3):437–450, 2003.

[LK00] Themis Lazaridis and Martin Karplus. Effective energy functions for protein structure prediction. *Current opinion in structural biology*, 10(2):139–145, 2000.

[LNC08] Albert Lehninger, DL Nelson, and MM Cox. *Lehninger Principles of Biochemistry.* WH Freeman: New York, NY, USA, 2008.

[LP02] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[LZZZ04] Song Liu, Chi Zhang, Hongyi Zhou, and Yaoqi Zhou. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins: Structure, Function, and Bioinformatics*, 56(1):93–101, 2004.

[MAB11] Iain H Moal, Rudi Agius, and Paul A Bates. Protein–protein binding affinity prediction on a diverse set of structures. *Bioinformatics*, 27(21):3002–3009, 2011.

[MCK09] Daniel J Mandell, Evangelos A Coutsias, and Tanja Kortemme. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature methods*, 6(8):551–552, 2009.

[Mer03] Rainer Merkl. A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. *Journal of molecular evolution*, 57(4):453–466, 2003.

[MFK$^+$09] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Burkhard Rost, and Anna Tramontano. Critical assessment of methods of protein structure prediction—round viii. *Proteins: Structure, Function, and Bioinformatics*, 77(S9):1–4, 2009.

[MJ85] Sanzo Miyazawa and Robert L Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552, 1985.

[MLM02] Cristian Micheletti, Gianluca Lattanzi, and Amos Maritan. Elastic properties of proteins: insight on the folding process and evolutionary selection of native structures. *Journal of molecular biology*, 321(5):909–921, 2002.

[MMHT92] Anne Louise Morris, Malcolm W MacArthur, E Gail Hutchinson, and Janet M Thornton. Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics*, 12(4):345–364, 1992.

[MS94] Victor Munoz and Luis Serrano. Intrinsic secondary structure propensities of the amino acids, using statistical ϕ–ψ matrices: Comparison with experimental scales. *Proteins: Structure, Function, and Bioinformatics*, 20(4):301–311, 1994.

[MSE03] Brendan J McConkey, Vladimir Sobolev, and Marvin Edelman. Discrimination of native protein structures using atom–atom contact scoring. *Proceedings of the National Academy of Sciences*, 100(6):3215–3220, 2003.

[Mue02] Ingo Muegge. A knowledge-based scoring function for protein-ligand interactions: probing the reference state. In *Virtual Screening: An Alternative or Complement to High Throughput Screening?*, pages 99–114. Springer, 2002.

[MWLZC02] Xiao Hui Ma, Cun Xin Wang, Chun Hua Li, and Wei Zu Chen. A fast empirical approach to binding free energy calculations based on protein interface information. *Protein engineering*, 15(8):677–681, 2002.

[MZW95] Dinesh Manocha, Yunshan Zhu, and William Wright. Conformational analysis of molecular chains using nano-kinematics. *Computer applications in the biosciences: CABIOS*, 11(1):71–86, 1995.

[NFW11] Sander Nieuwenhuis, Birte U Forstmann, and Eric-Jan Wagenmakers. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, 14(9):1105–1107, 2011.

[NOS05] Kimberly Noonan, David O'Brien, and Jack Snoeyink. Probik: Protein backbone motion by inverse kinematics. *The International Journal of Robotics Research*, 24(11):971–982, 2005.

[Not02] Cédric Notredame. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, 3(1):131–144, 2002.

[NT03] Irene Nooren and Janet M Thornton. Diversity of protein–protein interactions. *The EMBO journal*, 22(14):3486–3492, 2003.

[Pai97] Paul Painter. *Fundamentals of polymer science : an introductory text*. Technomic Pub. Co, Lancaster, Pa, 1997.

[PMD06] Géraldine Pascal, Claudine Médigue, and Antoine Danchin. Persistent biases in the amino acid composition of prokaryotic proteins. *Bioessays*, 28(7):726–738, 2006.

[PPS$^+$13] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David van der Spoel, et al. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, page btt055, 2013.

[Ram68] GN Ramachandran. Conformation of polypeptides and proteins'wl. *Advances in protein chemistry*, 23:283, 1968.

[RF10] Dmitry Rykunov and Andras Fiser. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC bioinformatics*, 11(1):128, 2010.

[RL14] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[RM76] GN Ramachandran and Alok K Mitra. An explanation for the rare occurrence of cis peptide units in proteins and polypeptides. *Journal of molecular biology*, 107(1):85–92, 1976.

[RMF08] R Rajgaria, SR McAllister, and CA Floudas. Distance dependent centroid to centroid force fields using high resolution decoys. *Proteins: Structure, Function, and Bioinformatics*, 70(3):950–970, 2008.

[Roc04] Eduardo PC Rocha. Codon usage bias from trna's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome research*, 14(11):2279–2286, 2004.

[Ros80] Kenneth A Ross. *Elementary analysis*. Springer, 1980.

[RR89] Janes S Richardson and David C Richardson. The de novo design of protein structures. *Trends in biochemical sciences*, 14(7):304–309, 1989.

[RRS63] GN Ramachandran, C t Ramakrishnan, and V Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*, 7(1):95–99, 1963.

[RSMB04] Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. Protein structure prediction using rosetta. *Methods in enzymology*, 383:66–93, 2004.

[Rud64] Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-Hill New York, 1964.

[SBG$^+$05] Paul M Sharp, Elizabeth Bailes, Russell J Grocock, John F Peden, and R Elizabeth Sockett. Variation in the strength of selected codon usage bias among bacteria. *Nucleic acids research*, 33(4):1141–1153, 2005.

[Sch10] L L C Schrödinger. The PyMOL molecular graphics system, version 1.3r1, August 2010.

[SGS$^+$15] E. Sarti, D. Granata, F. Seno, A. Trovato, and A. Laio. Native fold and docking pose discrimination by the same residue-based scoring function. *Proteins: Structure, Function and Bioinformatics (in print)*, 2015.

[Sip90] Manfred J Sippl. Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. *Journal of molecular biology*, 213(4):859–883, 1990.

[SK13] Amelia Stein and Tanja Kortemme. Improvements to robotics-inspired conformational sampling in rosetta. *PLOS ONE*, 8(5):e63090, 2013.

[SM98] Arthur G Street and Stephen L Mayo. Pairwise calculation of protein solvent-accessible surface areas. *Folding and Design*, 3(4):253–258, 1998.

[SR08] Armando D Solis and S Rackovsky. Information and discrimination in pairwise contact potentials. *Proteins: structure, function, and bioinformatics*, 71(3):1071–1087, 2008.

[SRK$^+$99] Kim T Simons, Ingo Ruczinski, Charles Kooperberg, Brian A Fox, Chris Bystroff, and David Baker. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Structure, Function, and Bioinformatics*, 34(1):82–95, 1999.

[SYDS11] OZ Sharabi, Chen Yanover, Ayelet Dekel, and Julia M Shifman. Optimizing energy functions for protein–protein interface design. *Journal of computational chemistry*, 32(1):23–32, 2011.

[SZC$^+$13] E Sarti, S Zamuner, Pilar Cossio, Alessandro Laio, Flavio Seno, and Antonio Trovato. Bachscore. a tool for evaluating efficiently and reliably the quality of large sets of protein structures. *Computer Physics Communications*, 184(12):2860–2865, 2013.

[SZP$^+$07] Alexander M Strasak, Qamruz Zaman, Karl P Pfeiffer, G Gobel, and Hanno Ulmer. Statistical errors in medical research-a review of common pitfalls. *Swiss medical weekly*, 137(3/4):44, 2007.

[SZS$^+$] T. Skrbic, S. Zamuner, E. Sarti, R. Hong, F. Seno, A. Laio, and A. Trovato. Entropic contributions determine binding affinity of protein complexes. *In preparation.*

[SZX$^+$09] Yu Su, Ao Zhou, Xuefeng Xia, Wen Li, and Zhirong Sun. Quantitative prediction of protein–protein binding affinity with a potential of mean force considering volume correction. *Protein Science*, 18(12):2550–2558, 2009.

[TAD06] Kristin Tøndel, Endre Anderssen, and Finn Drabløs. Protein alpha shape (pas) dock: A new gaussian-based score function suitable for docking in homology modelled protein structures. *Journal of computer-aided molecular design*, 20(3):131–144, 2006.

[TB03] Edward N Trifonov and Igor N Berezovsky. Evolutionary aspects of protein structure and folding. *Current opinion in structural biology*, 13(1):110–114, 2003.

[TBM$^+$03] Jerry Tsai, Richard Bonneau, Alexandre V Morozov, Brian Kuhlman, Carol A Rohl, and David Baker. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 53(1):76–87, 2003.

[TCPB04] G Tiana, M Colombo, D Provasi, and RA Broglia. Deriving amino acid contact potentials from their frequencies of occurrence in proteins: a lattice model study. *Journal of Physics: Condensed Matter*, 16(15):2551, 2004.

[TKKB01] Edward N Trifonov, Alla Kirzhner, Valery M Kirzhner, and Igor N Berezovsky. Distinct stages of protein evolution as suggested by protein sequence analysis. *Journal of molecular evolution*, 53(4-5):394–401, 2001.

[UJ03] Jakob P. Ulmschneider and William L. Jorgensen. Monte carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a gaussian bias. *The Journal of Chemical Physics*, 118(9):4261–4271, 2003.

[UJ04] Jakob P. Ulmschneider and William L. Jorgensen. Monte carlo backbone sampling for nucleic acids using concerted rotations including variable bond angles. *The Journal of Physical Chemistry B*, 108(43):16883–16892, 2004.

[VHPW12] Thom Vreven, Howook Hwang, Brian G Pierce, and Zhiping Weng. Prediction of protein–protein binding free energies. *Protein Science*, 21(3):396–404, 2012.

[VKMT10] Jari Vauhkonen, Ilkka Korpela, Matti Maltamo, and Timo Tokola. Imputation of single-tree attributes using airborne laser scanning-based height, intensity, and alpha shape metrics. *Remote Sensing of Environment*, 114(6):1263–1276, 2010.

[VMDL$^+$08] Marco Vassura, Luciano Margara, Pietro Di Lena, Filippo Medri, Piero Fariselli, and Rita Casadio. Reconstruction of 3d structures from protein contact maps. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 5(3):357–367, 2008.

[Whi05] David Whitford. *Proteins: structure and function.* John Wiley Sons, 2005.

[WL00] Lary C Walker and Harry LeVine. The cerebral proteopathies. *Molecular neurobiology*, 21(1-2):83–95, 2000.

[WL01] Fugao Wang and DP Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Physical Review E*, 64(5):056101, 2001.

[WSS99] Jörg Weiser, Peter S Shenkin, and W Clark Still. Approximate atomic surfaces from linear combinations of pairwise overlaps (lcpo). *Journal of Computational Chemistry*, 20(2):217–230, 1999.

[WWS02] William J Wedemeyer, Ervin Welker, and Harold A Scheraga. Proline cis-trans isomerization and protein folding. *Biochemistry*, 41(50):14637–14644, 2002.

[Yam71] Hiromi Yamakawa. Modern theory of polymer solutions. 1971.

[Zac03] Martin Zacharias. Protein–protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Science*, 12(6):1271–1282, 2003.

[ZC05] Lei Zhao and Jean Chmielewski. Inhibiting protein–protein interactions using designed molecules. *Current opinion in structural biology*, 15(1):31–34, 2005.

[ZL08] Zong-Hao Zeng and Yong C Li. Empirical parameters for estimating protein-protein binding energies: number of short-and long-distance atom-atom contacts. *Protein and peptide letters*, 15(2):223–231, 2008.

[ZPNH10] Qiangfeng Cliff Zhang, Donald Petrey, Raquel Norel, and Barry H Honig. Protein interface conservation across structure space. *Proceedings of the National Academy of Sciences*, 107(24):10896–10901, 2010.

[ZRST] S. Zamuner, A. Rodriguez, F. Seno, and A. Trovato. An efficient algorithm to perform local concerted movements of a chain molecule. *Submitted to PlosOne.*

[ZZ02] Hongyi Zhou and Yaoqi Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein science*, 11(11):2714–2726, 2002.