UNIVERSITÀ DEGLI STUDI DI MILANO

Facoltà di Scienze e Tecnologie
Laurea Magistrale in Fisica

Development of a model to study the thermodynamics of large protein systems

Advisor: Prof. Guido Tiana
Co-Advisor: Riccardo Capelli

Piero Valena
Matricola n° 864788
A.A. 2015/2016

Codice PACS: 87.15.-v

Development of a model to study the thermodynamics of large protein systems

Piero Valena
Department of Physics, Università degli Studi di Milano,
Via Celoria 16, 20133 Milano, Italy

April 7, 2017

Contents

Introduction

Proteins are linear polymers built by one or more chains of amino acids; the structure of an amino acid is summarized in Fig. 1, where R denotes the sidechain, a chemical group which differs in each amino acid type (there are twenty natural amino acids). In a protein each amino acid interacts with the others and with the solvent; this gives rise to various kinds of interactions (e.g. electrostatic, Van der Waals, hydrophobic..) whose heterogeneity makes proteins complex systems. In spite of this complexity proteins usually display a unique global energy minimum, called native state, that is the three-dimensional structure in which they perform their biological functions; this means that the structure of a protein depends only on the amino acid sequence [1, 2], which would be predictable if the potential of the system composed by the protein and the solvent were known. Being the calculation of the exact potential unfeasible for experimental and computational issues, several kinds of approximated potentials have been developed [3]; they try to catch all the main features of the true potential depending only on a small number of degrees of freedom. A way to build such a potential is to make use of a statistical approach, based on the analysis of the residue contact frequencies in experimental known structures [4]. In the last decades, the rapid advances in sequencing technology made possible a rapid growth of the number of protein sequences available [5], which can be grouped into families according to structural or functional similarity. As it is explained in Chapter 1, the analysis of correlated subsitution pattern within a protein family can give quantitative information on the interaction energies among the amino acids, and can thus be used to build an effective potential for a reference amino acid sequence.

In this thesis we test the effective potential obtained from this kind of analysis by carrying out Monte Carlo simulations of some proteins. The proteins we simulate are: the Bovine Pancreatic Tripsyn Inhibitor, the Apomyoglobin, the Staphilococcal Nuclease, and the Thioredoxin. Their structure has been experimentally determined, for the first three by x-ray cristallography, while for the Thioredoxin by NMR; they can be found in the Protein Data Bank (PDB)[6] with code, respectively, 1BPI, 1BVC, 1STN and 1RQM[7, 8, 9, 10]. The aim

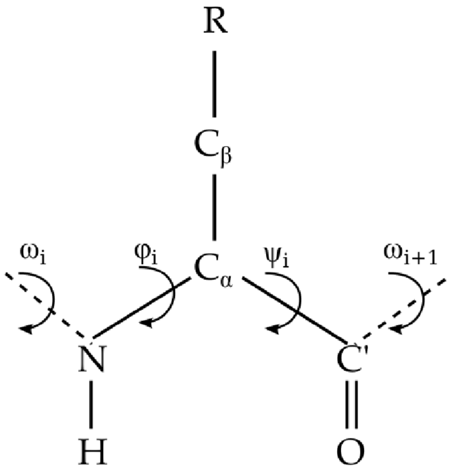


Figure 1: Schematic rapresentation of an amino acid in a protein: R indicates the sidechain, φ and ψ are the Ramachandran dihedral angles, while ω is the peptide bond dihedral.

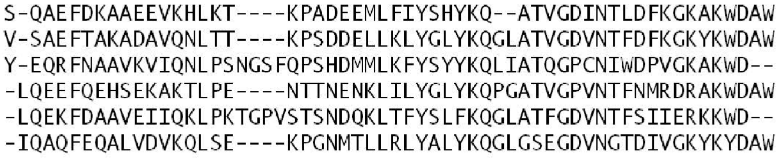
of this work is to find the best set of parameters from which our potential depends, and to verify their portability among different systems. In addition we carry out some folding simulations in order to see if the potential is able to predict the native structure of the proteins with the only knowledge of the amino acid sequence. In Chapter 1 we introduce the model underlying our potential; in Chapter 2 we present the methods adopted in the Monte Carlo simulations and we explain how we have set some quantities which appear in the potential basing on physical and statistical considerations; in Chapter 3 we describe the parametrization procedure, carried out basing on the results obtained from some simulations; in Chapter 4 we describe our preliminary study on the aggregation of the 1BPI.

Chapter 1

The model

The model that underlies the potential we use in our Monte Carlo simulations has been developed in previous works [11, 12, 13]. It is based on the analysis of residue coevolutionary data of proteins belonging to the same family.

A family is composed of proteins which show a high similarity in the amino acid sequence, that is usually translated into a sharp similarity in the three-dimensional structure and in the functionality. Moreover, proteins members of the same family are regarded as descendant of a common ancestor, so the mutations of the amino acid sequencies are the product of the evolutionary process. Tipically, proteins with more than 25% identical amino acids are evolutionary related and display the same structure. For each family, by means of a hidden Markov model[15], it is possible to build a Multiple Sequence Alignment (MSA)[14], that is a table whose rows identify the protein, and the columns are the letters corresponding to the amino acid sequence of that protein. An example of a MSA is shown in Fig. 1.1;"-" represents the lack of an amino acid, called "gap": they are added during the costruction of a MSA, in order to reproduce the deletion or the insertion of an amino acid in some sequences due to the evolutionary process. Analysing a MSA it is possibile to extract the mean frequency $f_i(\sigma)$ with which an amino acid of type σ is found in the site i, and the mean



```
S-QAEFDKAAEEVKHLKT----KPADEEMLFIYSHYKQ--ATVGDINTLDFKGKAKWDAW
V-SAEFTAKADAVQNLTT----KPSDDELLKLYGLYKQGLATVGDVNTFDFKGKYKWDAW
Y-EQRFNAAVKVIQNLPSNGSFQPSHDMMLKFYSYYKQLIATQGPCNIWDPVGKAKWD--
-LQEEFQEHSEKAKTLPE----NTTNENKLILYGLYKQPGATVGPVNTFNMRDRAKWDAW
-LQEKFDAAVEIIQKLPKTGPVSTSNDQKLTFYSLFKQGLATFGDVNTFSIIERKKWD--
-IQAQFEQALVDVKQLSE----KPGNMTLLRLYALYKQGLGSEGDVNGTDIVGKYKYDAW
```

Figure 1.1: Example of a Multiple Sequence Alignment composed by six proteins. The capital letters identify the amino acid type, while the "-" denotes the lack of an amino acid.

frequency with which the amino acid pair (σ, τ) is found in the sites (i, j), the $f_{ij}(\sigma, \tau)$.

The study of these coevolutionary data can give quantitative information about the interaction energy between residues. The basic idea is that, within a single family, there is a potential which produces the experimentally observed native structure of the proteins. This potential operates in the sequences space, and the different proteins in a family can be regarded as the result of fluctuations in sequences space. This because proteins display the lowest energy in their native configuration. If we knew the exact form of the potential, we could predict the observed single-site frequencies and the pair frequencies. Actually we deal with the inverse problem: starting from a MSA we can extract both the single-site frequencies and the pair frequencies, and we aim to find the potential which produces these observed data. Another point of view is that this kind of analysis can predict the interaction energy between the sites of a MSA because if two sites interact in favorable way, then their mutation pattern will be correlated. Starting from the assumption that proteins are energetically highly optimized systems (thanks to evolution), if two sites interact and one is modified, then there will be an increase in the energy of the system, which will be compensated by the mutation of the other site.

The input of the model is therefore a MSA of a protein family, which can be obtained from the UniProt database[16]. Starting from a MSA the aim of the model is to build a probability distribution for a reference amino acid sequence, $p(\{\sigma_i\}_{i=1}^{L})$ (where L is the number of sites and σ_i is the amino acid in position i), that reproduces the empirically observed frequencies. This means that $p(\{\sigma_i\}_{i=1}^{L})$ has to satisfy the costraints

$$\begin{gathered}\sum_{\{\sigma_k\}} p(\sigma_1, \sigma_2, ..., \sigma_L)\delta(\sigma_i, A) = f_i(A) \quad \forall i, \sigma_i \\ \sum_{\{\sigma_k\}} p(\sigma_1, \sigma_2, ..., \sigma_L)\delta(\sigma_i, A)\delta(\sigma_j, B) = f_{ij}(A, B) \quad \forall i, j, \sigma_i, \sigma_j\end{gathered} \tag{1.1}$$

where the summations are over all the sequences belonging to the family. By applying the maximum-entropy principle[17], the explicit mathematical form of $p(\{\sigma_i\}_{i=1}^{L})$ is derived: defining the effective potential

$$\mathcal{H} = \sum_{i<j}^{L} u_{ij}(\sigma_i, \sigma_j) + \sum_{i=1}^{L} \mu(\sigma_i) + \sum_{i=1}^{L}[\tilde{h}_i(\sigma_i) - \frac{1}{L}\sum_{j=1}^{L} \tilde{h}_i(\sigma_j)], \tag{1.2}$$

the probability of a specific sequence in the protein family is

$$p(\{\sigma_i\}_{i=1}^{L}) = \frac{1}{Z}e^{-\mathcal{H}}. \tag{1.3}$$

Z is the partition function, while $u_{ij}(\sigma,\tau)$, $h_i(\sigma)$ and $\mu(\sigma)$ are the Lagrange multipliers that come from the costrained maximization. Due to the analogy with the Boltzmann distribution, they can be regarded as effective energies. In particular, $e_{ij}(\sigma,\tau)$ is the two-body interaction energy between the amino acid σ in site i and the amino acid τ in site j, $\tilde{h}_i(\sigma)$ is a local one-body energy contribution (which we call h-fields), acting on the amino acid σ in site i, and $\mu(\sigma)$ is a site-indipendent energy, so we assign to it the physical meaning of the chemical potential of the amino acid σ. Being the h-fields a one-body term, they would reflect the presence of an external field; this is however hard to justify, so they must be regarded as the results of the combined effect of the sorrounding residues, meaning a many-body term. In particular, the h-fields are found to reflect the hydrophobicity of the residues [13], which is an intrinsically many-body term.

Operatively, the model is implemented in CoCaInE [18] (a python written code), with which, given an amico acid sequence and the corresponding MSA, we can extract both the two-body energies and the h-fields. We are not interested in the chemical potentials, because they represent just an energy shift, which does not contribute to the protein structure stability.

1.1 The potential

In order to carry out Monte Carlo simulations of the systems under investigation, we have to translate the effective energies obtained from CoCaInE (two-body and h-fields) to a potential that rules the dynamics. To do this we define a potential of the form

$$U = U_{2b} + U_{dih} + U_{hf}, \tag{1.4}$$

where U_{2b} is the two-body interaction term between the amino acids, U_{dih} is the potential acting on the Ramachandran dihedrals, obtained indipendently of the alignment, and U_{hf} is a one-body potential that reflects the presence of the h-fields.

1.1.1 Two-body potential

The two-body term of the potential (1.4) is

$$U_{2b} = \sum_{i=1}^{L} \sum_{j=i+1}^{L} N_i N_j e_{ij}, \tag{1.5}$$

where i and j are amino acid indexes, L is the protein length in residues, N_i and N_j are the number of atoms (not belonging to the backbone) in the respective

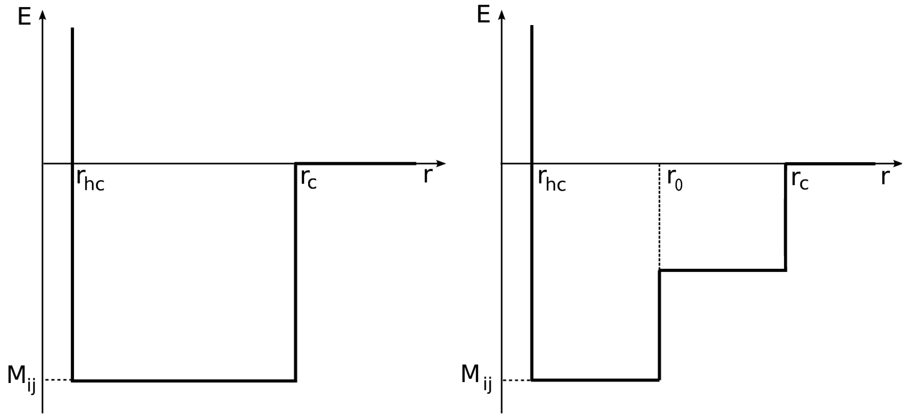


Figure 1.2: Schematic representation of two possible choice of thes two-body potential for a generic amino acid pairs (i, j). On the left the single square well, on the right the double square well.

amino acids, and e_{ij} is the matrix element of this pair, which can be either a square well or a double square well; specifically

$$e_{ij}(r) = \begin{cases} +\infty & \text{if } r < r_{hc} \\ M_{ij} & \text{if } r_{hc} < r < r_c \\ 0 & \text{if } r > r_c \end{cases} \tag{1.6}$$

or

$$e_{ij}(r) = \begin{cases} +\infty & \text{if } r < r_{hc} \\ M_{ij} & \text{if } r_{hc} < r < r_0 \\ M_{ij}/2 & \text{if } r_0 < r < r_c \\ 0 & \text{if } r > r_c \end{cases} \tag{1.7}$$

where r_{hc} is the hardcore distance, set to 2 Å, r_0 is set to 2.5 Å, and r_c is the contact radius. A schematic representation of the two-body potential (fixed an amino acid pair (i, j)) is shown in Fig 1.2. This term of the potential acts on each pair of atoms not belonging to the backbone, but the interaction energies are the same for each atom pair belonging to the same amino acid pair. The effective energies obtained by means of a coevolutionary analysis are in fact defined over the amino acid pairs, but we want to build an all atom model; the simplest way to do this is defining the interactions as described above. The choice between the two alternative forms of the well is one of the goals of the parametrization of the potential (1.4), as it is the choice of r_c value. The well depths M_{ij} are obtained starting from the u_{ij} that come from CoCaInE, and then applying the method described in Chapter 2.

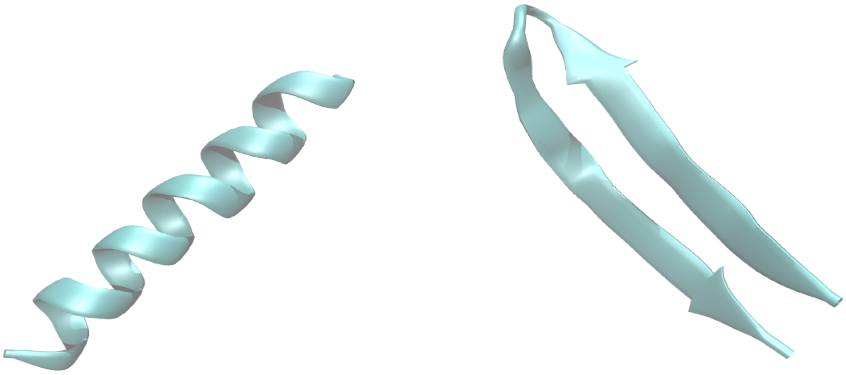


Figure 1.3: Example of an $\alpha-$helix (left) and of a $\beta-$sheet (right).

1.1.2 Dihedrals potential

The potential acting on Ramachandran dihedrals is

$$\begin{aligned} U_{dih} &= e_\alpha \sum_{i=1}^{L} U_\alpha^i + e_\beta \sum_{i=1}^{L} U_\beta^i \\ U_\alpha^i &= p_{i\alpha}(\mathcal{N}(\phi_{0\alpha}, \sigma_{\phi\alpha}) + \mathcal{N}(\psi_{0\alpha}, \sigma_{\psi\alpha})) \\ U_\beta^i &= p_{i\beta}(\mathcal{N}(\phi_{0\beta}, \sigma_{\phi\beta}) + \mathcal{N}(\psi_{0\beta}, \sigma_{\psi\beta})) \end{aligned} \tag{1.8}$$

$\mathcal{N}(x, \sigma_x)$ is the normal distribution centered in x with standard deviation σ_x, e_α and e_β are the depth of the gaussians which stabilize respectively the $\alpha-$helixes and the $\beta-$sheets (an example of these secondary structures is shown in Fig. 1.3), while $p_{i\alpha}$ and $p_{i\beta}$ are the probabilities to find the $i-$th amino acid in an $\alpha-$helix or $\beta-$sheet conformation[1]. A schematic representation of the dihedral potential acting on a generic amino acid is shown in Fig. 1.4. This potential acts to stabilize the secondary structures of a protein ($\alpha-$helixes and $\beta-$sheets), whose presence depends on the dihedrals values. Therefore we have a term for each dihedral angle (φ and ψ), and for each kind of secondary structure. The values of the quantities that appear in Eq.(1.8) are chosen following the methods described in Chapter 2.

1.1.3 h-fields potential

The potential which takes into account the h-fields is

$$U_{hf} = \frac{1}{\alpha} \sum_{i=1}^{L} N_i h_i \cdot n_i(\mathbf{x}), \tag{1.9}$$

[1] The meaning of these probabilities is explained in Section 2.1.2.

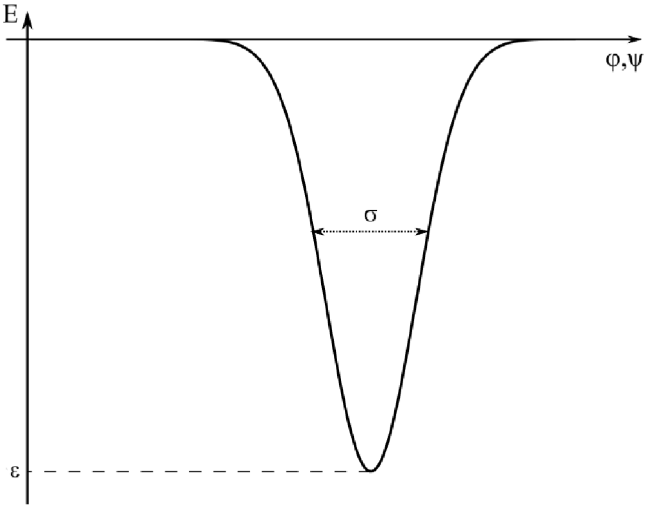


Figure 1.4: Schematic representation of the dihedral potential (energy versus Ramachandran dihedral angle) for a generic amino acid, ε indicates the depth of the Gaussian.

where α is a number that sets the relative importance of this term with respect to the total potential, N_i is the number of atoms in the amino acid i, $n_i(\mathbf{x})$ is the number of atomic contacts made by the $i-$th amino acid in the protein conformation $\mathbf{x}$, and h_i is the effective energy associate to the $i-$th amino acid. As for the two-body term, h_i is the same for each atom in the amino acid i. The value of the constant α will be set during the optimization of the potential, while h_i is obtained starting from the CoCaInE value $\tilde{h}_i$, and then applying the method described in Chapter 2. Being the h-fields related to the hydrophobicity of the amino acids, the functional form of this potential has been chosen in such a way that, if an amino acid has a negative value of h_i, then it will try to maximize its atomic contacts. The global effect is therefore to bury the hydrophobic residues and to expose the hydrophilic ones, as the hydrophobic interaction does.

Chapter 2

Methods and implementation

2.1 A priori chosen quantities

A very large part of the present thesis is devoted to the selection of the values of the variables that appear in the potential (1.4). Some of these parameters are chosen a priori, with a choice based on physical considerations and on the information that comes from the structures of some known proteins. Other variables have to be chosen by carrying out simulations and inspecting useful observables. Moreover, some of these parameters (the two-body interaction matrix M, the propensities p_α and p_β, and the h-fields h) are of course system dependent, while the others have in principle universal validity. In this section we describe how we have set the values of all the quantities which are chosen a priori. Moreover, since one of the goals of this work is to predict the native structure of a protein knowing only the amino acid sequence, all the system-dependent quantities have to be derived from this only knowledge.

2.1.1 Two-body interaction matrix M

Given the protein under investigation, we can obtain the MSA of the family to which it belongs, and its amino acid sequence. With these inputs we are able to run CoCaInE. If L is the length of the alignment (i.e. the number of columns in the MSA) and q is the number of types of amino acids[1], we obtain the four-dimensional two-body energy tensor $\mathbf{U}$, whose elements are $u_{ij}(\sigma, \tau)$ (with $i, j = 1..L$ and $\sigma, \tau = 1..q$). As described in Chapter 1, they represent the effective two-body interaction energies between the amino acids of the type σ and τ found

[1] In a MSA we have $q = 21$, because there are the natural 20 residues types plus the gap, that represents the absence of an amino acid in a specific site (column) of a MSA. The presence of the gap is also the reason why L does not correspond with the length of the protein.

respectively in the site i and j of the MSA. However, it is not enough to select the elements $u_{ij}(\sigma_i, \sigma_j)$ corresponding to the reference protein to obtain the two-body energies of interest to us; before of that we have to modify them, for the reasons and by applying the methods described below.

Contact filtering

The energy values $u_{ij}(\sigma, \tau)$ are obtained from the inversion of a correlation matrix[11]; in particular

$$u_{ij}(\sigma, \tau) = C_{ij}^{-1}(\sigma, \tau), \tag{2.1}$$

where

$$C_{ij}(\sigma, \tau) = f_{ij}(\sigma, \tau) - f_i(\sigma) f_j(\tau), \tag{2.2}$$

where f_{ij} and f_k are the frequencies defined in Chapter 1. Being the statistics limited, the u_{ij} are subject to errors; specifically some energies are large while they would be small if calculated with an infinite statistics. We therefore have to implement a filter on the contacts energies to minimize the statistical error.

To do this we follow the approach developed by Morcos and co-workers [11]. A measure of the correlation between two sites i and j of a given MSA is the mutual information (MI) between i and j, defined as

$$MI_{ij} = \sum_{\sigma,\tau} f_{ij}(\sigma, \tau) \ln \frac{f_{ij}(\sigma, \tau)}{f_i(\sigma) f_j(\tau)}, \tag{2.3}$$

which equals zero if and only if $i-$th and $j-$th sites are uncorrelated and it is positive otherwise. To retain only the contribute to the MI which comes from the direct coupling alone, we introduce for each column pair (i, j) an *isolated* two-site model, and we define the direct information (DI) between i and j [19] as

$$DI_{ij} = \sum_{\sigma,\tau} P_{ij}^{(\mathrm{dir})}(\sigma, \tau) \ln \frac{P_{ij}^{(\mathrm{dir})}(\sigma, \tau)}{f_i(\sigma) f_j(\tau)}, \tag{2.4}$$

where $P_{ij}^{(\mathrm{dir})}(\sigma, \tau)$ is defined by the equations

$$\begin{aligned} P_{ij}^{(\mathrm{dir})}(\sigma, \tau) &= \frac{1}{Z_{ij}} \exp(u_{ij}(\sigma, \tau) + \hat{h}_i(\sigma) + \hat{h}_j(\tau)) \\ f_i(\sigma) &= \sum_{\tau} P_{ij}^{(\mathrm{dir})}(\sigma, \tau) \\ f_j(\tau) &= \sum_{\sigma} P_{ij}^{(\mathrm{dir})}(\sigma, \tau) \end{aligned} \tag{2.5}$$

In Eqs. (2.5) $u_{ij}(\sigma, \tau)$ are the elements of the tensor $\mathbf{U}$ which comes from CoCaInE, while the $\hat{h}_i$ are implicitly defined here as auxiliary fields by compatibility with the observed single-site frequencies[2]. The principle under this filter is that if the mutation pattern of two sites i and j is correlated even in a toy model in which we do not consider all the other sites (big DI_{ij}), then we can assert that i and j truly interact; conversely, if they present a small DI_{ij} value it means that the two sites are weakly interacting, and if the statistical had been ideal, we would have found $u_{ij} \simeq 0$.

Operatively, we wrote a python code that, for each pair (i, j), solves the Eqs. (2.5) (in which the variables are the arrays $\hat{h}_i$ and $\hat{h}_j$) and calculates DI_{ij}. This means solving a system of $2q$ paired equations for each pair of sites. In doing this we have to take into account that not all of the equations are independent, because all the variables have to satisfy the normalization conditions (the second and the third of Eqs. (2.5)), so there are $2q-2$ indipendent equations; this means that we have to fix the values of two variables. We choose

$$\hat{h}_i(q) = \hat{h}_j(q) = 0 \quad \forall i, j = 1..L, \tag{2.6}$$

where $q = 21$ (that corresponds to the residue gap). Once the equations are solved we can compute the DI matrix by means of Eq. (2.4), and we redefine the interaction matrix $u_{ij}(\sigma_i, \sigma_j)$ as

$$u_{ij}(\sigma_i, \sigma_j) = \begin{cases} u_{ij}(\sigma_i, \sigma_j) & \text{if } DI_{ij} > DI_0 \\ 0 & \text{if } DI_{ij} < DI_0 \end{cases} \tag{2.7}$$

where the threshold value of the direct information DI_0 is one of the model's parameters, and it is chosen by carrying out some simulation (see Chapter 3).

Normalization of the residue-residue energies

Once the contacts are filtered, we have to take into account that the effective energies which come from CoCaInE are multiplied by an immaterial constant, which is different from system to system[3] We therefore normalize the energy matrix over the standard deviation of its elements values; this is done to fix the magnitude of the energies that come from CoCaInE to 1 indipendently from the system under investigation. In this way we make comparable the parameters of the potential (1.4) which fix the relative weight of the various terms (e_α, e_β and α) among the different systems. We choose to divide over the standard deviation and not over the mean because the latter fluctuates around zero, so the division would be a very noisy operation.

[2] Note that the $\hat{h}$ are not related to the $\tilde{h}$ that come from CoCaInE.

[3] This because the effective energies basically come from a count in the MSA.

Finally we have to face the problem that we want to simulate an all-atom system, but the two-body energies are defined over the residues. This means that we have to normalize every matrix element in some way in order to split the total residue-residue interaction among all the atoms of the two amino acids. We choose to normalize every matrix element over the maximum number of atom contacts that occurs in the amino acid pair (i, j) in any protein[4]. Operatively, to calculate these normalizing factors, we wrote a C++ code which scans 20000 protein structures taken from the PDB; for each one we consider only atom pairs that do not belong to the backbone (those atoms do not interact in our simulations) and that are separated by at least two residues (otherwise they do not interact in our simulations), and, during the scanning procedure, we update the maximum value of atomic contacts for each amino acid pair. In doing this we define a contact radius r_c (which is the same of the one in Eqs. (1.6) and (1.7)), and we consider two atoms in contact if their distance is less than r_c.

Once all these operations are performed, we can finally define the interaction matrix M, whose elements are

$$M_{ij} = \begin{cases} \dfrac{u_{ij}(\sigma_i, \sigma_j)}{s \cdot c_{ij}} & \text{if } DI_{ij} > DI_0 \\ 0 & \text{if } DI_{ij} < DI_0 \end{cases} \tag{2.8}$$

where s is the standard deviation described above and c is the maximum contacts matrix.

2.1.2 Secondary structures propensities

In a protein each amino acid can be classified from a secondary structure point of view: it can be in a coil, helix or sheet state. If it belongs to a coil state, it means that the amino acid does not form any secondary structure, while helix and sheet mean that it is part of, respectively, an $\alpha-$helix or a $\beta-$sheet. Each amino acid has a propensity to be in one of these three states, which can be quantified as a probability to find the amino acid in the corresponding one; this probability depends only on the amino acid sequence. For each protein we can predict the propensities of all the amino acids starting from the only knowledge of the amino acid sequence by means of PSIPRED[20, 21], a tool that, by means of a two-stage neural network algorithm, quantifies the propensities of a given amino acid sequence.

[4] The "correct" normalization factors, calculated over the number of contacts that occur between the pair (i, j) in the protein under investigation, cannot be obtained without the knowledge of the native structure.

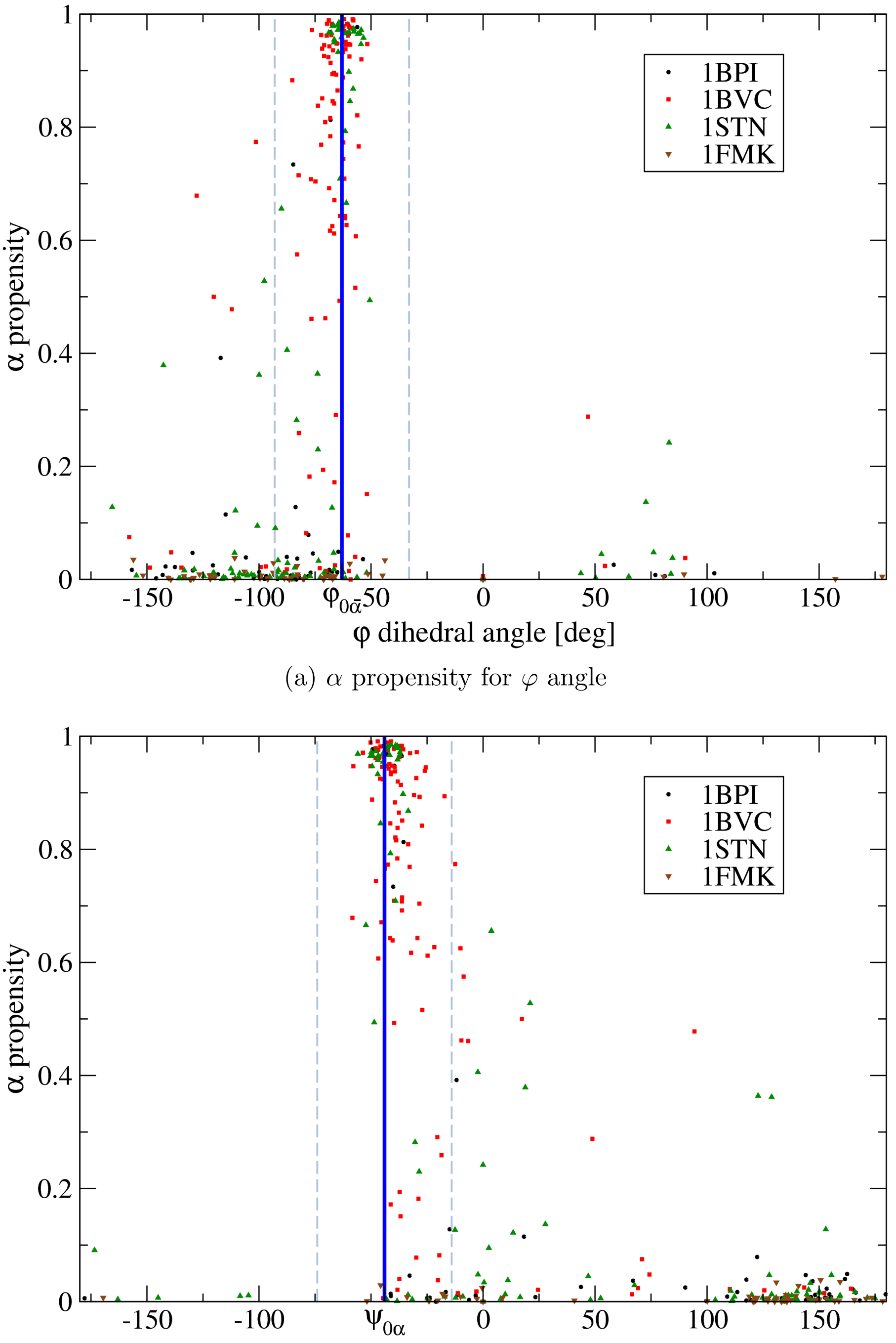
Dihedral type	Optimal angle [deg]	Standard deviation [deg]
φ_α	-63	30
φ_β	-105	40
ψ_α	-44	30
ψ_β	140	40

Table 2.1: Optimal dihedral angles and standard deviations for each dihedral type that appears in the potential (1.8)

The α and β propensities obtained from PSIPRED are directly the p_α and the p_β that appear in the dihedral potential (1.8), while we do not make use of the coil propensity, because no potential acts to stabilize a coil state.

2.1.3 Dihedrals angles and standard deviations

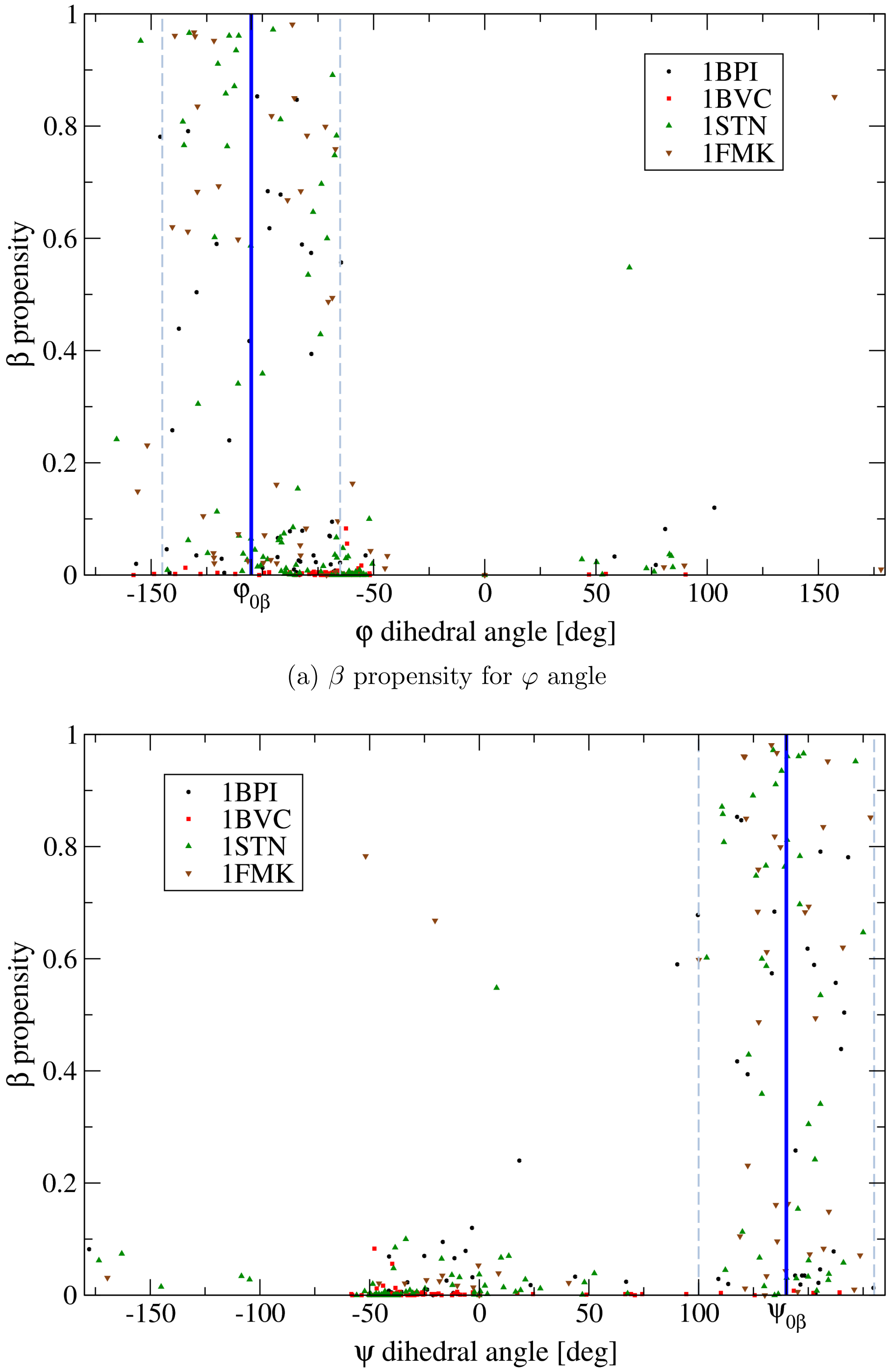
The potential acting on the Ramachandran dihedrals is basically the same for every amino acid, it consists of a sum of four normal distributions, each centered in a particular angle and with a particular variance. These angles should be universal, meaning that each amino acid that belongs to a specific secondary structure should have approximately the same dihedral angles, independently of the protein to which it belongs. This because the secondary structures are local arrangements of the amino acids in a protein, so their presence will depend from some properties of the involved amino acids, and not from all the protein. In other words, an $\alpha-$helix conformation, for example, has basically the same properties in every protein, so what characterizes its presence does not depend from the system in which it is. To verify this assertion, and to find the optimal dihedrals angles $(\psi_{0\alpha}, \psi_{0\beta}, \varphi_{0\alpha}, \varphi_{0\beta})$, we adopt the following method. Starting from the PDB files of some known protein (1BPI, 1BVC, 1STN, 1FMK[22]) we calculate both the Ramachandran dihedral angles of each amino acid and the secondary structures propensities. The propensities are calculated as described above, while the dihedral angles are computed by means of gmx angle, a GROMACS[23] tool. At this stage we can associate to each amino acid of the considered proteins the propensities $(p_{i\alpha}, p_{i\beta})$ and the dihedral values (ψ_i, ϕ_i); then we make a scatterplot of the propensities versus the dihedrals for each of the four possible combinations (as shown in Fig. 2.1 and 2.2), and therefore choose the best values for the parameters $\psi_{0\alpha}, \psi_{0\beta}, \varphi_{0\alpha}, \varphi_{0\beta}$ and their standard deviations. The basic idea underlying the procedure is that our potential has to facilitate the presence of a secondary structure for which an amino acid has a high propensity; for each



(a) α propensity for φ angle

(b) α propensity for ψ angle

Figure 2.1: α propensities versus the φ dihedral angles (a) and ψ dihedral angles (b) for the residues of 1BPI, 1BVC, 1STN, 1FMK. The solid blue line indicates the optimal dihedral value ($\varphi_{0\alpha}$ and $\psi_{0\alpha}$), while the dashed grey lines define the region within one σ away from the optimal angle.



(a) β propensity for φ angle

(b) β propensity for ψ angle

Figure 2.2: β propensities versus the φ dihedral angles (a) and ψ dihedral angles (b) for the residues of 1BPI, 1BVC, 1STN, 1FMK. The solid blue line indicates the optimal dihedral value ($\varphi_{0\beta}$ and $\psi_{0\beta}$), while the dashed grey lines define the region within one σ away from the optimal angle.

angle we then choose a value that fits the dihedral values of the amino acids with high propensity, and a standard deviation that allows all the amino acids with high propensity to have a significant value of the dihedral potential. By looking at Fig. 2.1 and 2.2 we can choose the values beared in Table 2.1.

2.1.4 h-fields

The values of the h-fields are obtained starting from the CoCaInE h-fields output (the arrays $\tilde{h}_i$), and then performing a normalization similar to that carried out for the two-body matrix. First we normalize over the two-body energy standard deviation s (in this way we measure the h-fields effective energies with respect to the two-body energies); then we rescale every element of the h-fields array using the maximum number of contacts done by the corresponding amino acid in any protein. To perform this normalization we follow the same procedure used for the two-body term, and the code is a variation of the previous one. This operation has a double effect: first it takes into account that we are dealing with an all-atom model (as in the two-body term); secondly it normalizes the energies in such a way that when we multiply $\tilde{h}_i$ by the number of contacts n_i in the potential (1.9), we obtain the original effective energy reweighted by the fraction of contacts that the $i-$th amino acid can make. Summing up we have

$$h_i = \frac{\tilde{h}_i}{s \cdot c_i}, \tag{2.9}$$

where c_i is the maximum contact array.

2.2 Overview of the Monte Carlo method

All the simulations performed in this thesis are carried out by means of a Monte Carlo (MC) sampling. In a MC simulation the system under investigation follows a fictitius dynamics, where the conformational changes are not driven by physical forces, but by a set of implemented moves, whose only aim is to make the system explore the conformational space according to some probability distribution, in our case a Boltzmann distribution. At each step of the simulation the system tries to jump from an initial configuration μ to a final one ν, and a move can be either accepted or rejected. The transition rate from μ to ν is what actually rules the dynamics, and is chosen as

$$w(\mu \to \nu) = w_0 P_{ap}(\mu \to \nu) A(\mu \to \nu), \tag{2.10}$$

where w_0 is a constant that sets the time scale of the jumps, $P_{ap}(\mu \to \nu)$ is the *a priori* probability to jump from the conformation μ to ν and $A(\mu \to \nu)$ is the

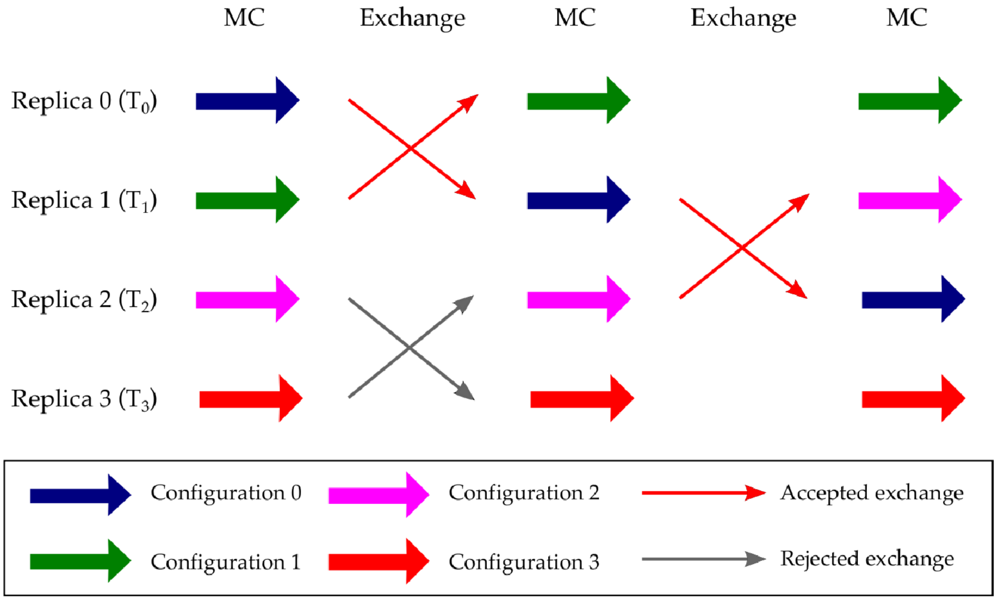


Figure 2.3: Schematic representation of the parallel tempering behaviour for a system composed of four replicas.

acceptance rate of the move which leads from μ to ν. To reproduce a Boltzmann distribution we have to satisfy several conditions. First, the *a priori* probabilities have to be chosen in such a way that the rates w satisfy the detailed balance principle, which states

$$w(\mu \rightarrow \nu)w(\nu \rightarrow \eta)w(\eta \rightarrow \mu) = w(\mu \rightarrow \eta)w(\eta \rightarrow \nu)w(\nu \rightarrow \mu)\ \forall \mu, \nu, \eta. \quad (2.11)$$

There are several possible choices of P_{ap} that satisfy the condition (2.11), for example one is to choose $P_{ap}(\mu \rightarrow \nu) = f(|\mu - \nu|)$, where f is a function which depends only on some distance between the two conformations μ and ν. Secondly, the acceptance rate has to be [24]

$$A(\mu \rightarrow \nu) = \min[1, e^{-\frac{U_\mu - U_\nu}{T}}], \quad (2.12)$$

where U_α is the energy of the configuration α and T is the temperature of the system [5]. If conditions (2.11) and (2.12) are satisfied, and if the sampling is ergodic, then, at equilibrium, the probability to find the system in a conformation $\vec{x}$ is given by the Boltzmann distribution

$$p(\vec{x}) = \frac{1}{Z}e^{-\frac{U(\vec{x})}{T}}, \quad (2.13)$$

[5] We fix $k_b = 1$, that means we use the same (arbitrary) measurements units for the energies and the temperatures.

where Z is the (unknown) partition function of the system. This means that, at equilibrium, according to the ergodic theorem, the thermal average of an observable $\langle O \rangle$ can be calculated as an aritmetic "time" average over all the conformations visited during a simulation, so

$$\langle O \rangle = \sum_{\{\vec{x}\}} O(\vec{x})p(\vec{x}) = \frac{1}{N}\sum_{i=1}^{N} O(\vec{x}_i), \tag{2.14}$$

where $\{\vec{x}\}$ is the ensamble of every possible conformation $\vec{x}$, N is the number of Monte Carlo steps and $\vec{x}_i$ is the conformation visited at the $i-$th step. The power of the Monte Carlo method relies on the fact that, if one can find a wise way to sample the phase space of the system, then it is possible to obtain all the equilibrium properties of the system without any knowledge of its partition function Z (whose calculation is unfeasible for every non-trivial system).

An effective sampling of the phase space is the key point for a good MC simulation, which is translated into a wise choice of the allowed moves that lead the system from an initial conformation μ to a final one ν. As said before, the sampling has to be ergodic, meaning that the system can explore the entire conformational space without any dependence on the initial state. Defining two conformations μ and ν in contact if there is a single move which leads from μ to ν, the set of moves defines a contact matrix $M(\mu, \nu)$ between each possible configuration pair (μ, ν), whose elements are 1 if there is a contact, 0 otherwise. The ergodicity is translated in the requirement that the matrix M cannot be decomposed in blocks, otherwise the system cannot escape the block from where it starts, thus cannot reach the equilibrium state. Moreover, the set of moves has to make the sampling efficient: on one hand these moves have to drive the system to not too different consecutive conformations in order to keep the acceptance rate (2.12) high, on the other two consecutive conformation cannot be too similar, because the exploration of entire configuration space would require a huge computational time. Moreover, even if the set of moves is optimal, the system will tend to get trapped in local minima of the free energy, reducing the efficiency of the sampling.

A way to overcome this problem is to adopt a Parallel Tempering technique, in which N_r identical replicas of the system under investigation are simulated, each at a different temperature T_i. The simulations are carried out indipendently, except for the fact that every n_s step an exchange between two replicas is attempted: this exchange can be regarded as a Monte Carlo move, and it is driven by the rate

$$w(m \leftrightarrow n) = \min[1, e^{-(\beta_m - \beta_n)(U_n - U_m)}]. \tag{2.15}$$

The idea underlying this procedure is that high replicas (which are simulated at high temperatures) can diffuse freely in the phase space, because their free energy profile is basically flat, so if a low replica gets trapped in a local minimum it can escape by raising its temperature. As before, the set of temperatures and the pairs involved in an exchange have to be chosen in such a way to keep high the rate (2.15): to this aim the exchange has to be attempted between replicas in which U_n is a typical energy at T_m and vice versa. Therefore the exchange is usually attempted between subsequent replicas, and the set of temperatures is chosen uneven: low temperatures are very close to each other, because here the energy fluctuations are small, while higher ones are more separated, being the fluctuations bigger. A schematic representation of the Parallel Tempering procedure is shown in Fig. 2.3.

Chapter 3

Results

3.1 Parameters optimization

In this section we describe how we have set the values of the parameters that are not a priori fixed. We set their values by carrying out Monte Carlo simulations of known proteins and selecting the values that best reproduce the observables that well-describe the experimental native structures of the proteins. These parameters are:

- the constant which sets the h-fields relative weight α (Eq. 1.9);
- the dihedral potential depths e_α and e_β (Eq. 1.8);
- the contact radius r_c (Eq. 1.5);
- the presence of a splice in the two-body potential well;
- the direct information threshold DI_0 (Eq. 2.7).

An ideal parametrization should be done by binning in an adeguately way each parameter, and by carrying out simulations using all the possible combinations of the parameters. Alternatively, if one could define a particular order parameter, with which quantify the fitness of a simulation, it would be possible to apply an automatic optimization algorithm that, moving in the parameters space, finds the set that maximizes this order parameter. Our parametrization should be made for at least two proteins (one which presents only α-helix and one which presents only β-sheets as secondary structures) to verify the universality of the best set of parameters. Unfortunately we do not have the computational power to apply the first method, nor a single quantity that could be associated to an order parameter to apply the second one. We choose then to perform a complete parametrization on one protein (the 1BPI, which is the smaller one in our set) and then verify the fitness of the best parameter set on the others.

The variable we monitor to quantify the fitness of a simulation is the root mean square deviation (RMSD) of the atomic position of the C_α of the protein with respect to their experimental native positions. The RMSD is always calculated after having performed a rototranslation of the conformation under investigation onto its reference structure. The RMSD of the vector $\mathbf{v}$ with respect to $\mathbf{w}$ is thus defined as

$$\mathrm{RMSD}(\mathbf{v}, \mathbf{w}) = \min_{\{\mathrm{rt}\}} \sqrt{\frac{1}{N} \sum_{i=1}^{N} ||v_i - w_i||^2}, \tag{3.1}$$

where $\{\mathrm{rt}\}$ is the set of rototranslations performed to align the structures, and N is the length of the vector (in our case the protein length in residues). Starting from the RMSD of a single simulation frame, calculated by means of the GROMACS tool g_rms[23],we can build the mean RMSD ($\overline{\mathrm{RMSD}}$) as a function of the temperature and we can also make a scatterplot in which we draw the RMSD of a single frame versus the energy of the corresponding conformation; these are the quantities we actually consider. The former shows the global trend of the RMSD and allows the identification of a folding temperature as well as the quality of the folded conformations, while the latter gives some information about the shape of the energy profile. In the ideal case we would see a very low value of the $\overline{\mathrm{RMSD}}$ in correspondence of the lowest temperatures, and a scatterplot in which the RMSD minimum corresponds to the energy minimum, and this point (which represents the native structure) would be separated by the others[1]. In addition to the RMSD we always look at the Monte Carlo trajectory to have a global idea of the structure of the simulated proteins, which cannot be completely caught by the RMSD.

The technical parameters (e.g. the number of Monte Carlo steps) used in all the simulations are chosen as described in Appendix A. Except where otherwise specified the starting conformation is the experimental native state of the protein under investigation, taken from the PDB, and the starting point for the calculation of all the quantities is the Monte Carlo step $2 \cdot 10^7$. This is a point at which we have verified that the systems seem to have equilibrated: by inspecting the energy shape of some systems (see Fig. A.2) we can see how the relative energy change between the step $2 \cdot 10^7$ and the end of the simulations is very small compared to the total energy change. Moreover, we calculate some quantities varying

[1] The proteins we simulate experimentally show a unique native conformation, which is energetically separated by the others by a free energy barrier

the ending point of the trajectory, and we see how the results are basically the same (see Fig. A.3).

3.1.1 Parametrization of 1BPI

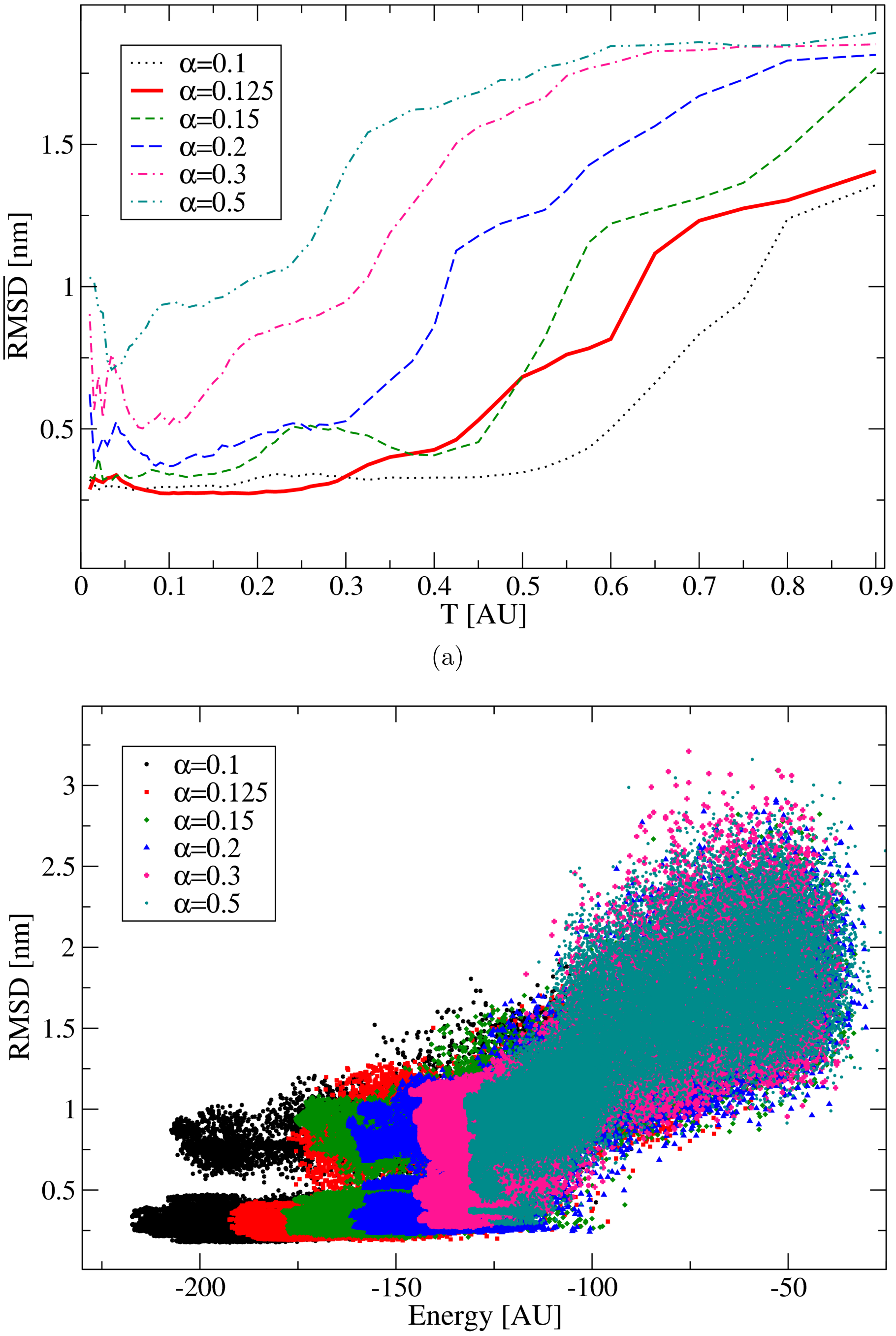
The parametrization of the 1BPI is carried out under the assumption that each parameter is independent from the others, so the scheme we adopt is the following: we scan a parameter at time, keeping fixed the others, and we find its best value. This is used in the scan of the next parameter, and so on until we have found all the values. The parameters are scanned in the same order in which they are listed above. Finally we run a longer simulation ($n_s = 5 \cdot 10^8$) in which the starting conformation is an unfolded one, in order to verify if it is true and within which limits our model can predict the native structure without any knowledge on it. The DI threshold is found to be a critical parameter, in the sense that varying sligthly its value changes significantly the simulations results. For this reason, before the complete parametrization we carry out some preliminary simulations in order to find a reasonable value of DI_0, but the real scan of this parameter is made at the end. We have chosen this approach instead of carrying out the DI scan directy at the beginning mainly because in this way we can have a double check of the threshold best value, which partially allows to verify the real independence of the parameters, and makes us sure of the choice of this critical parameter.

Scan of α

The first parameter we optimize is the h-fields constant α. Since we have to use reliable values also for the other parameters, we carry out some preliminar simulations, and, by the analysis based on the RMSD described above, we find that a plausible set of parameters is:

- $0.1 < \alpha < 0.5$;
- $e_\alpha = e_\beta = 80$;
- $r_c = 3.0\,\text{Å}$ without splice;
- $DI_0 = 0.007$.

We then run some simulations using these parameters and varying α between 0.1, 0.125, 0.15, 0.2, 0.3, 0.5. For each one we carry out the analysis based on the RMSD calculation, and we obtain the results shown in Fig. 3.1. By inspecting Fig. 3.1, we can see how $\alpha = 0.125$ is the value which shows the best $\overline{\text{RMSD}}$ at low temperatures, and best describes the minimum energy conformation as the native one. We therefore choose as the best $\alpha = 0.125$.



(a)

(b)

Figure 3.1: 1BPI α parametrization results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for $\alpha = 0.1$, 0.125,0.15, 0.2, 0.3, 0.5, while the other parameters are fixed to $e = 80$, $r_c = 3.0\,\text{Å}$, $DI_0 = 0.007$. The red continue line indicates the best α value, which is $\alpha = 0.125$.

Scan of e_α and e_β

The dihedral energy parametrization is performed using:

- $\alpha = 0.125$;
- $40 < e_\alpha, e_\beta < 120$;
- $r_c = 3.0\,\text{Å}$ without splice;
- $DI_0 = 0.007$.

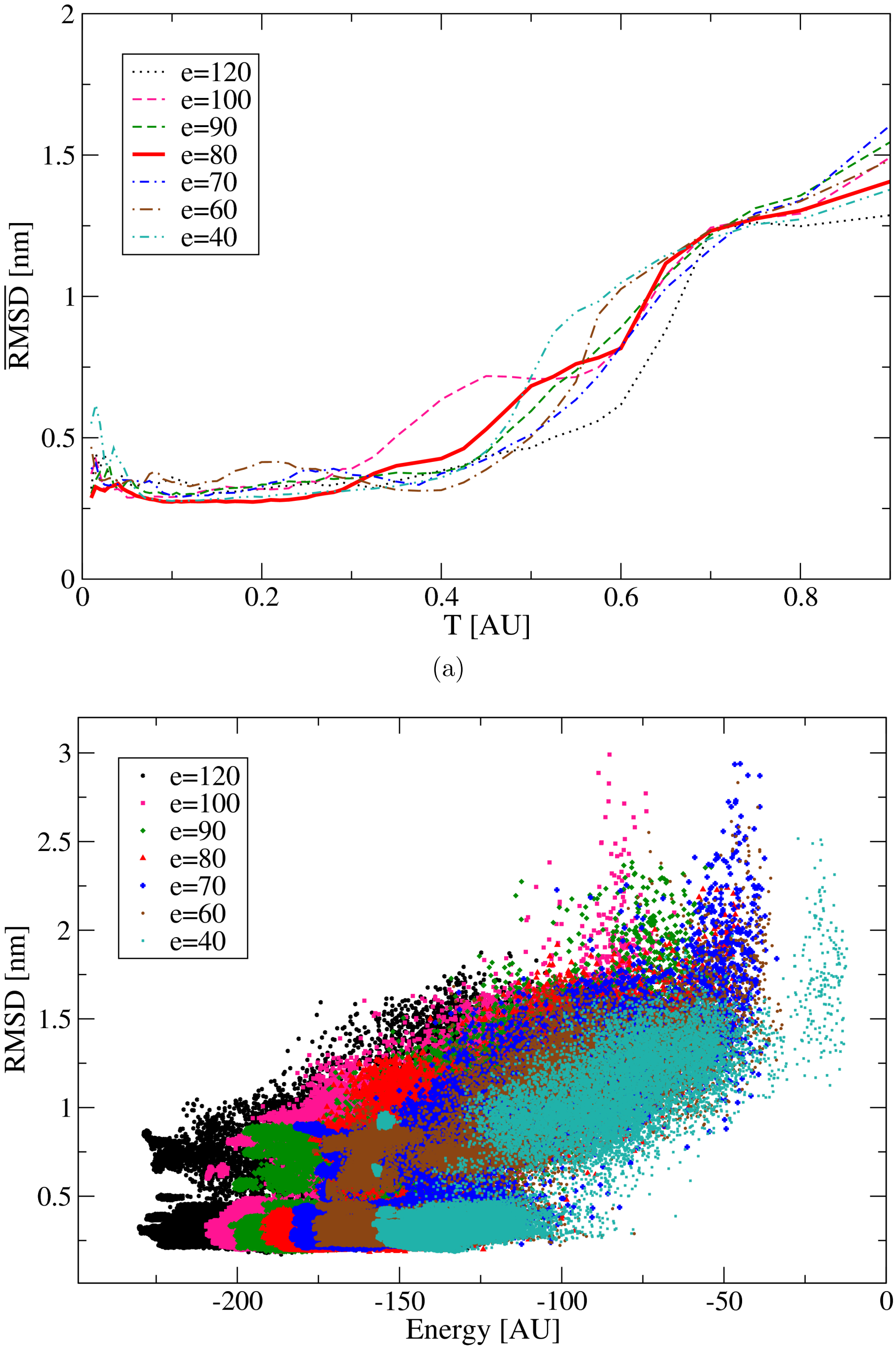
We vary e_α and e_β between 40, 60, 70, 80, 90, 100, 120; from the results shown in Fig. 3.2 we can see how $e_\alpha = e_\beta = 80$ best reproduce the experimental native configuration at low temperatures and low energies. Note that in principle e_α and e_β could be different, but the global structure of this protein is due to the $\beta-$sheets, so this parametrization is mainly devoted to e_β (we do a complementary parametrization of e_α for the 1BVC, which contains only $\alpha-$helixes).

Scan of r_c

The contact radius scan is performed using:

- $\alpha = 0.125$;
- $e_\alpha = e_\beta = 80$;
- $3.0\,\text{Å} < r_c < 4.5\,\text{Å}$;
- $DI_0 = 0.007$.

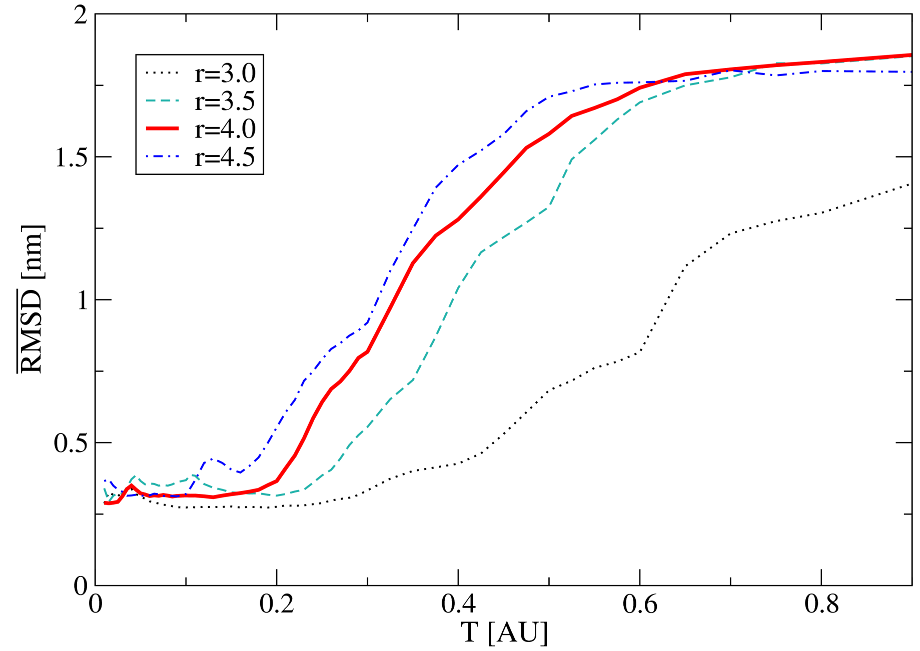
We vary r_c between 3.0, 3.5, 4.0, 4.5 Å; since from preliminary simulations it is found that the presence of a splice does not modify in a significant way the results, we scan this parameter without it, and at the end, chosen the best value, we perform an additional simulation in which we add a splice in the two-body potential well. From the results shown in Fig. 3.3 we see how $r_c = 4.0\,\text{Å}$, despite the presence of a metastable state with low energy and high RMSD, is the radius for which the minimum energy values is truly associated to the minimum RMSD. Furthermore, by looking at Fig. 3.4, we see that the presence of a splice is substantially inconsequential, so we demand the definitive choice to a simulation of the 1STN.



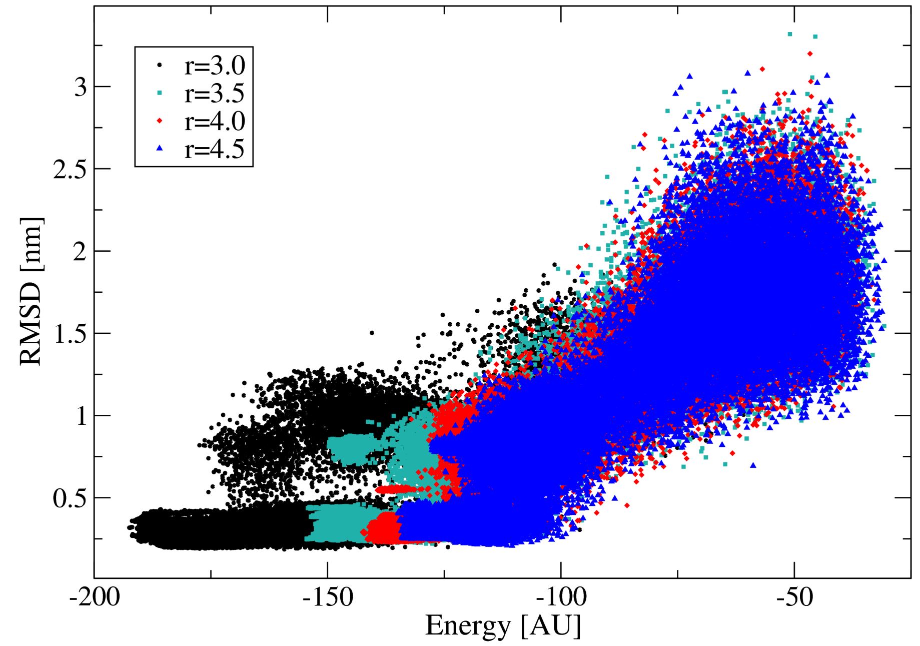
(a)

(b)

Figure 3.2: 1BPI dihedral energy parametrization results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for the energies 120, 100, 90, 80, 70, 60, 40, while the other parameters are fixed to $\alpha = 0.125$, $r_c = 3.0\,\text{Å}$, $DI_0 = 0.007$. The red continue line indicates the best e value, which is $e = 80$.

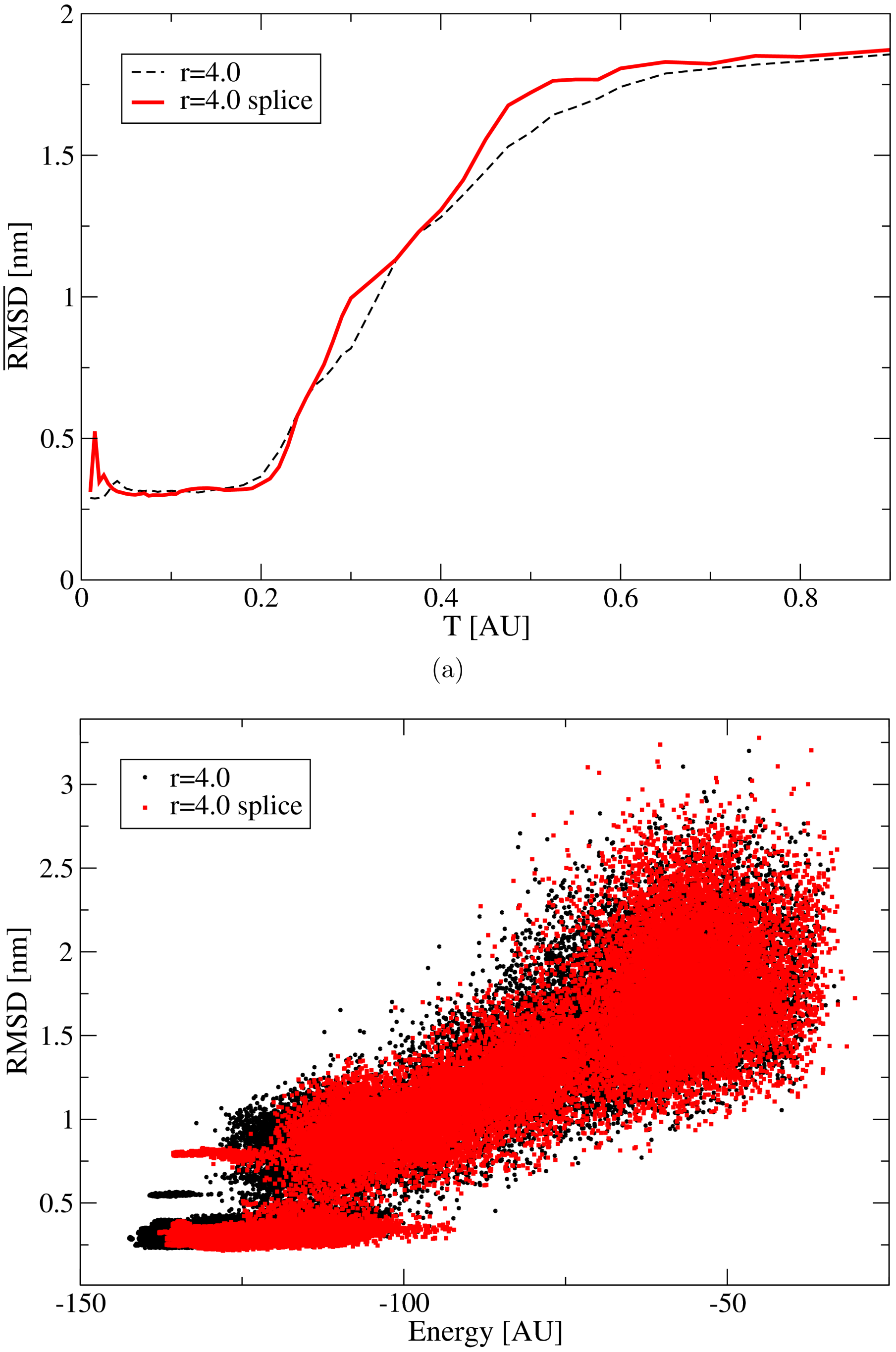


(a)



(b)

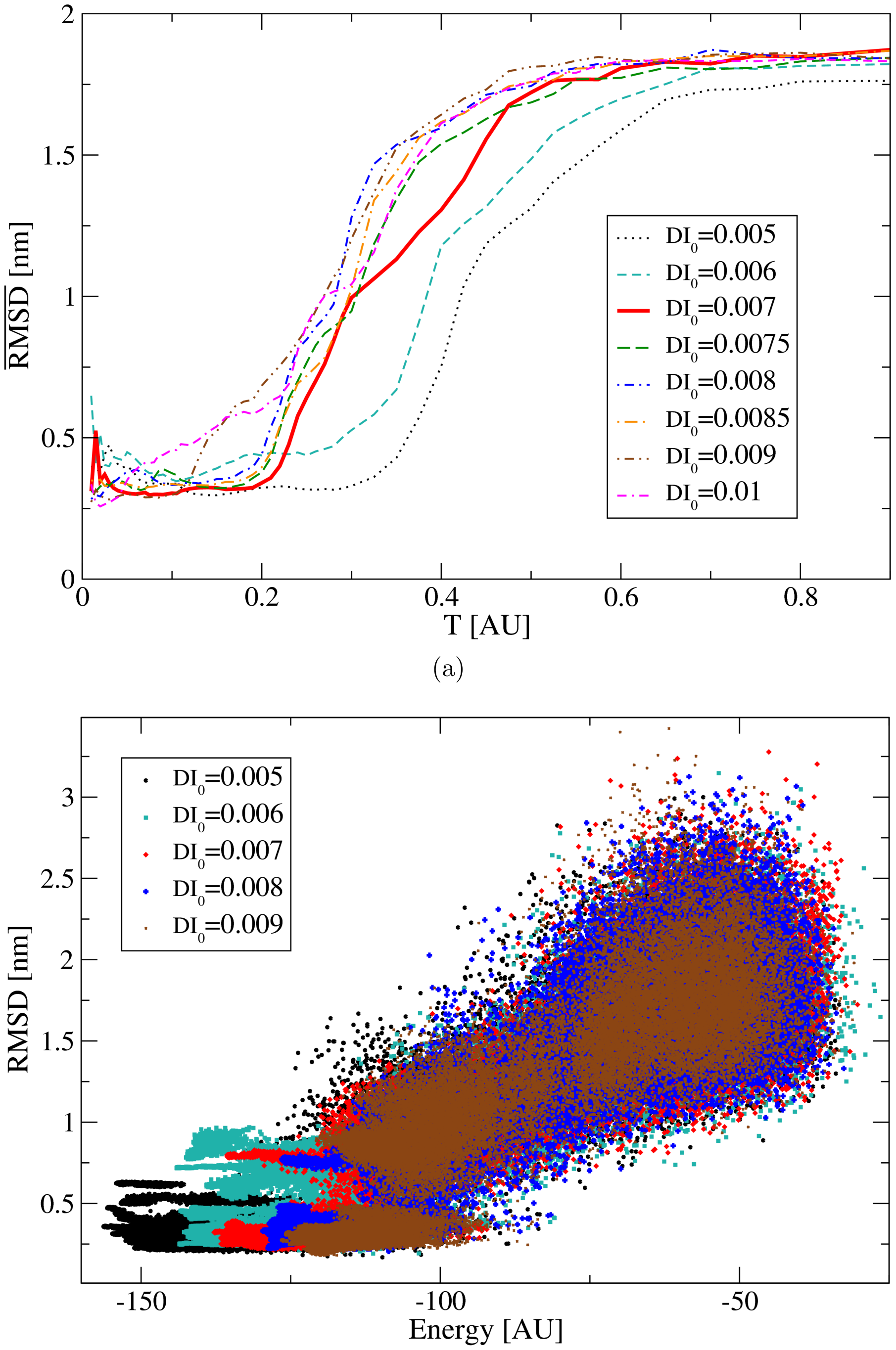
Figure 3.3: 1BPI contact radius parametrization results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for r_c =3.0, 3.5, 4.0, 4.5 Å, while the other parameters are fixed to $\alpha = 0.125$, $e = 80$, $DI_0 = 0.007$. The red continue line indicates the best r_c value, which is $r_c = 4.0$ Å.



(a)

(b)

Figure 3.4: 1BPI splice parametrization results: (a) $\overline{\mathrm{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The simulations are performed using $\alpha = 0.125$, $e = 80$, $DI_0 = 0.007$ and $r_c = 4.0\,\text{Å}$ with and without a splice. The results are subtantially equivalent, so we do not choose a best simulation.



(a)

(b)

Figure 3.5: 1BPI DI threshold parametrization results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for $DI_0 = 0.005$, 0.006, 0.007, 0.0075, 0.008, 0.0085, 0.009, 0.01, while the other parameters are fixed to $\alpha = 0.125$, $e = 80$, $r_c = 4.0\,\text{Å}$ with splice. The red continue line indicates the best DI_0 value, which is $DI_0 = 0.007$.

Scan of DI_0

Finally, the DI_0 scan is carried out with the set:

- $\alpha = 0.125$;
- $e_\alpha = e_\beta = 80$;
- $r_c = 4.0\,\text{Å}$ with splice;
- $0.005 < DI_0 < 0.01$.

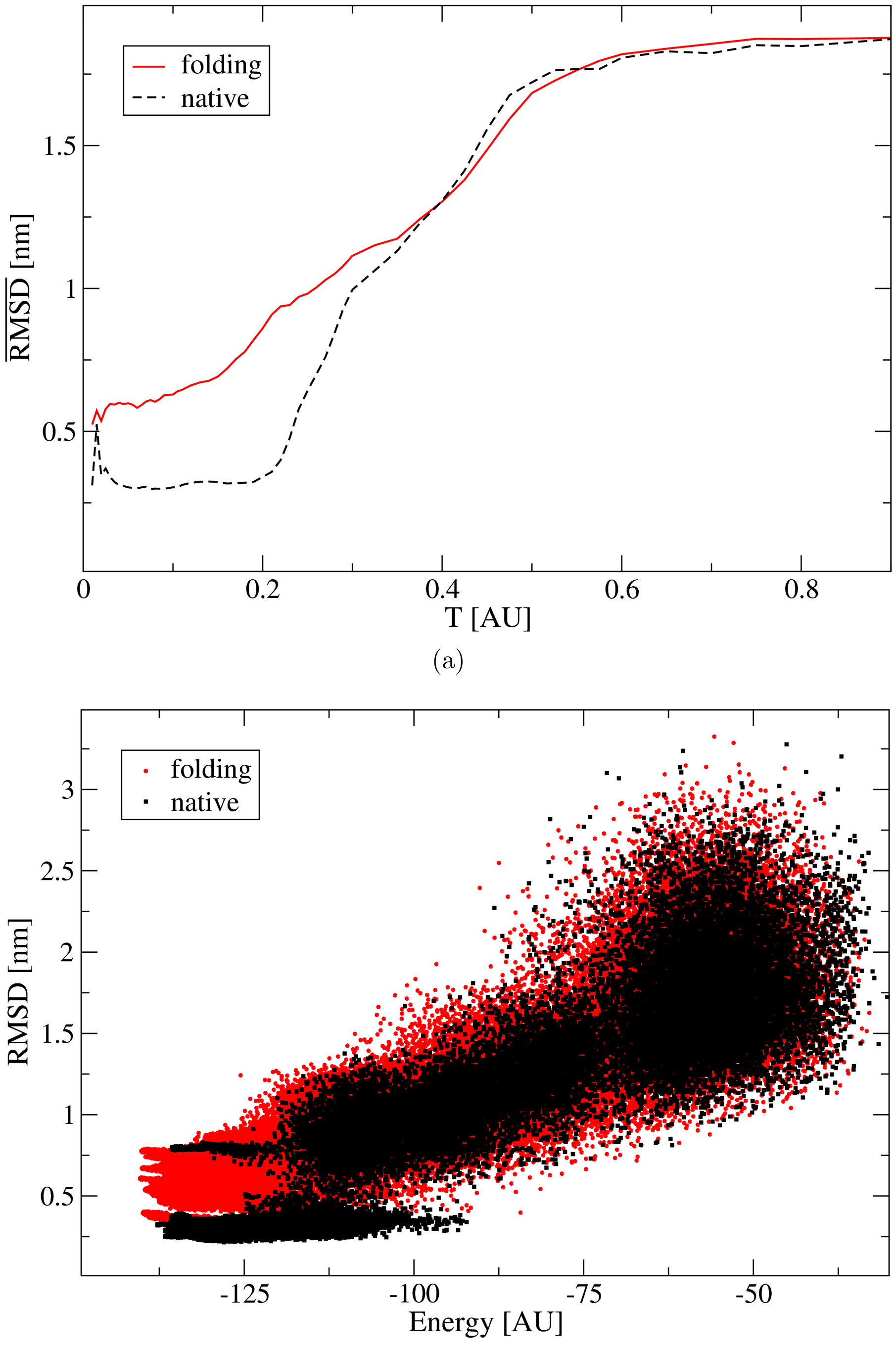
We vary DI_0 between 0.005, 0.006, 0.007, 0.0075, 0.008, 0.0085, 0.009, 0.01. The results are shown in Fig. 3.5. As for the radius, in spite of the presence of a metastable state, $DI_0 = 0.007$ best reproduces the experimental native state at low energies; therefore we keep it as the best value, which is also the best found by preliminar simulations.

Validation of conformational sampling

In the ideal case of infinite long simulations a simulation that starts from an unfolded conformation would be totally equivalent to a simulation which starts from the native configuration. Since our simulations are finite, as final test, we perform a folding of the 1BPI with the optimal parameter set ($\alpha = 0.125, e_\alpha = e_\beta = 80, r_c = 4.0\,\text{Å}$ with splice, $\mathrm{DI}_0 = 0.007$). We implement the same analysis based on the RMSD calculation and we obtain the results shown in Fig. 3.6 (in which are also represented the results of the “native” simulation). The shape of the $\overline{\text{RMSD}}$ is different at low temperatures (Fig. (a)), this probably means that the true equilibrium configuration is something in the middle of the two curves. The analysis of the second plot (Fig. (b)) provides some information: first, the fact that the energy of the minimum energy conformation is approximately the same in both cases tells us that our potential has a global minimum around the native state; furthermore the fact that there are many conformations with low energies which are not visited during the “native” simulations suggests that this minimum is narrowed. In addition to this, we have a confirmation that our potential can predict the native state of the 1BPI within a RMSD of about 5.0 Å, with the only knowledge of the amino acid sequence.

3.1.2 Parametrization of 1BVC, 1STN and 1RQM

The other systems we parametrize are the 1STN, the 1BVC and the 1RQM. As described above, we do not have enough computational power to carry out a



(a)

(b)

Figure 3.6: Red (continue line) results of the 1BPI starting from an unfolded conformation, black (dashed line) results of the same simulation started from the experimental native conformation. (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The parameters of the simulation are fixed to $\alpha = 0.125$, $e = 80$, $DI_0 = 0.007$, $r_c = 4.0\,\text{Å}$ with splice

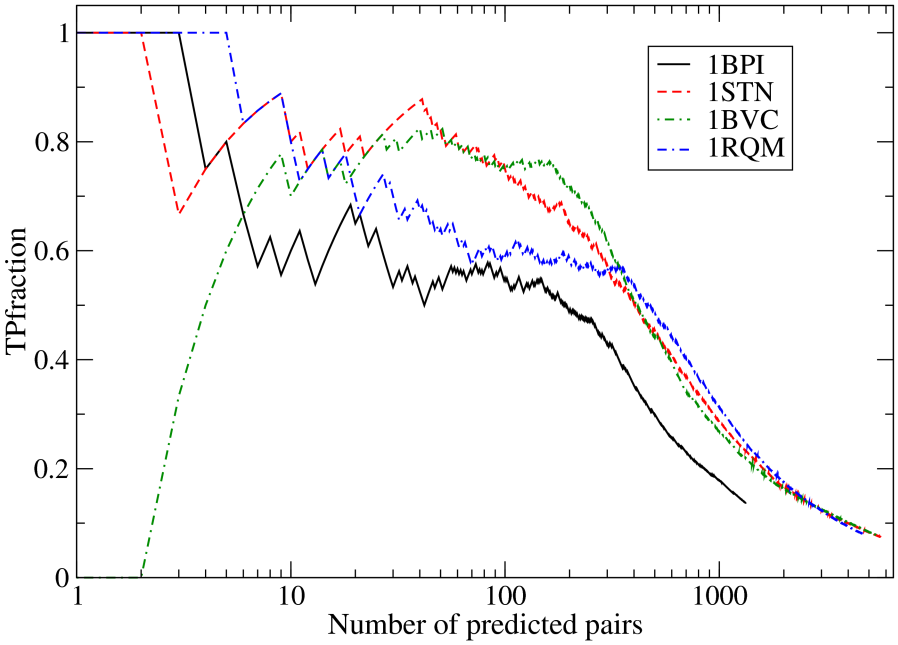


Figure 3.7: TPfraction in function of the number of predicted pairs sorted by decreasing DI for 1BPI, 1STN, 1BVC and 1RQM.

complete scan of all the parameter for all the proteins, so what we do is check if the optimal set of parameters found with the 1BPI works fine for these proteins too. Among the others the direct information threshold is a particular quantity, because while $\alpha, e_\alpha, e_\beta$ and r_c have a direct physical meaning and can thus be directly trasferred through the systems, DI_0 is not so easily interpretable. We therefore define two variables, associated to the DI, which could supply a criterion to set the DI threshold for all the proteins. The first is the mean number of residue-residue contacts per amino acid, k, which is defined as

$$k = \frac{n_{2b}}{L}, \tag{3.2}$$

where n_{2b} is the number of elements of the two-body interaction matrix different from zero (set by the value of DI_0) and L is the protein length. The second is the True Positive fraction (TPfraction)[11]. It is calculated as follows: first we identify the native contacts of the given protein (this can be easily achieved by means of CoCaInE), then we sort the amino acid pairs by decreasing order of DI, and we assign a flag to each pair, which tells us if that pair is a "native" one or not; finally we can calculate the fraction of native pairs as a function of the number of predicted pairs (sorted by decreasing DI). To each value of DI_0 corresponds thus a TPfraction value. The idea beyond this quantity is that in the ideal case the contacts which show the highest DI values are also the native contacts, so this variable can give some information about the fitness of our DI filter. We thus use it as a proxy to find good models instead of performing the full simulation. In Fig. 3.7 it is shown the TPfraction for each of the protein simulated during the parametrization.

For the 1BPI, to $DI_0 = 0.007$ corresponds $k = 3.4$ and TPfraction = 0.48, so in the simulations of the other proteins we choose a DI_0 starting value which produces comparables k and TPfraction. However we perform a DI scan for each protein, in order to check if at least one of these variables can be used to guess *a priori* the threshold value.

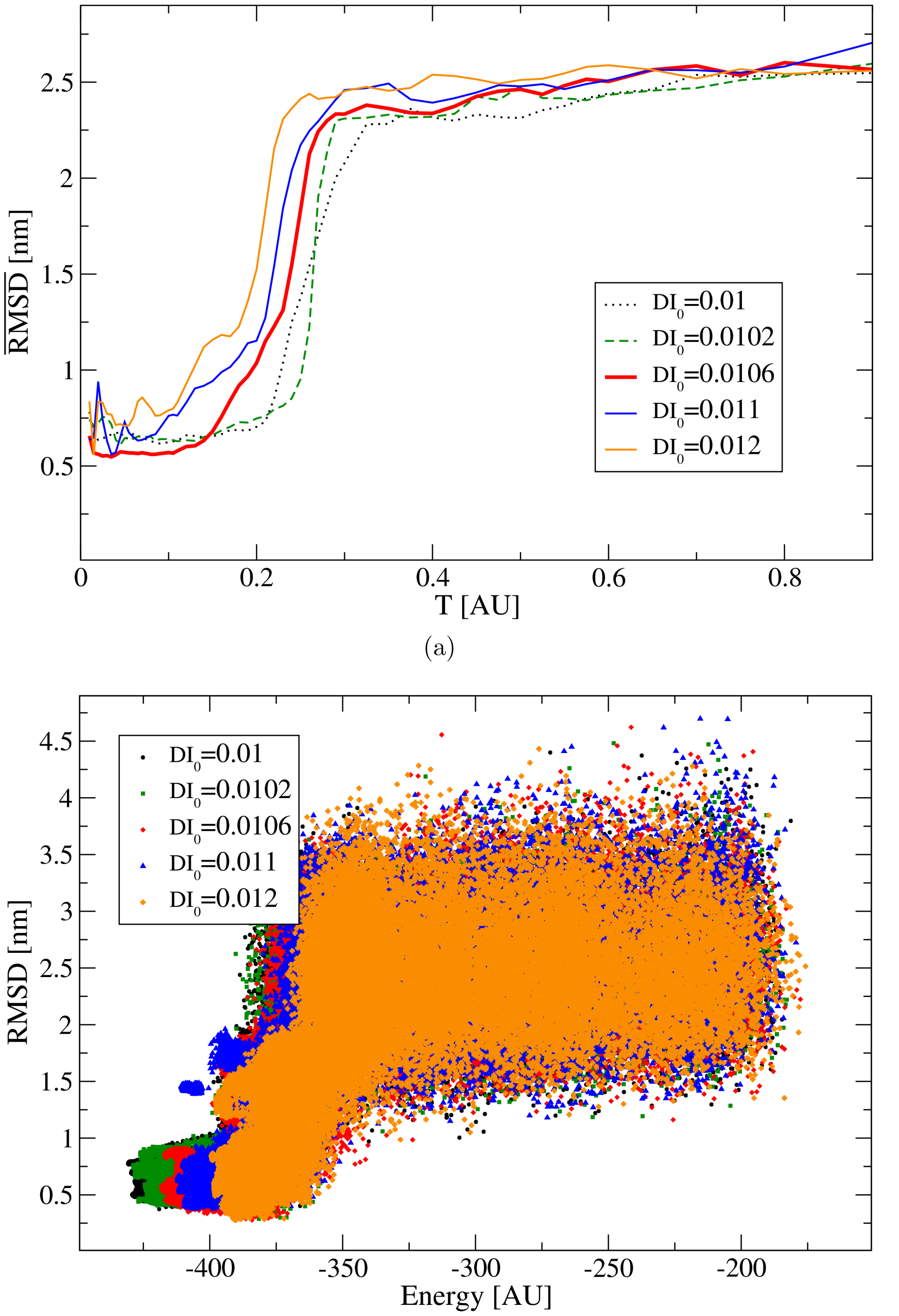
1BVC

For the 1BVC we vary DI_0 between 0.01, 0.0102, 0.0106, 0.011, 0.012, and we obtain the results shown in Fig. 3.8. We identify as the best $DI_0 = 0.0106$. As described above, being this a protein which presents only $\alpha-$helixes, we repeat the dihedral energy e_α scan, varying it between 70, 80, 90, and we obtain the results shown in Fig. 3.9. While $e_\alpha = 90$ gives worse results, $e_\alpha = 70$ and $e_\alpha = 80$ are essentially equivalent, so we retain the value $e_\alpha = 80$.

1STN

For the 1STN we vary DI_0 between 0.005, 0.006, 0.007, 0.0075, 0.008, 0.0085, 0.009, 0.01, and we obtain the results shown in Fig. 3.10. We identify as the best $DI_0 = 0.0075$. In addition we carry out a simulation in which we set $r_c = 4.0\,\text{Å}$ without a splice, to decide whether or not the double well is better than the single one; as shown in Fig. 3.11, while the $\overline{\text{RMSD}}$ shape is basically the same, by looking at the plot (b) we can see how, without a splice, there are conformations which present RMSD $\simeq 0.5$ with an energy lower than the minimum RMSD ones. This shows an incorrect behaviour of the model, so we keep a double square well. Finally we perform a folding simulation, to have an additional test of our potential. The results are shown in Fig. 3.12. The results of the “native” simulation are not directly comparable to the previous ones, because in this simulation we improved MonteGrappa by adding some features concerning the rotamers, which modify the results[2]. By looking at Fig. 3.12 we can see how, as it happens for the 1BPI (Fig. 3.6), the $\overline{\text{RMSD}}$ is higher for the folding simulation than the native one, as it is the RMSD of the folding conformations for the same energy. Moreover, by inspecting the Monte Carlo trajectories, we see that while the $\alpha-$helixes are well shaped, the $\beta-$sheets are not; this is true also for the 1BPI, so we definitively

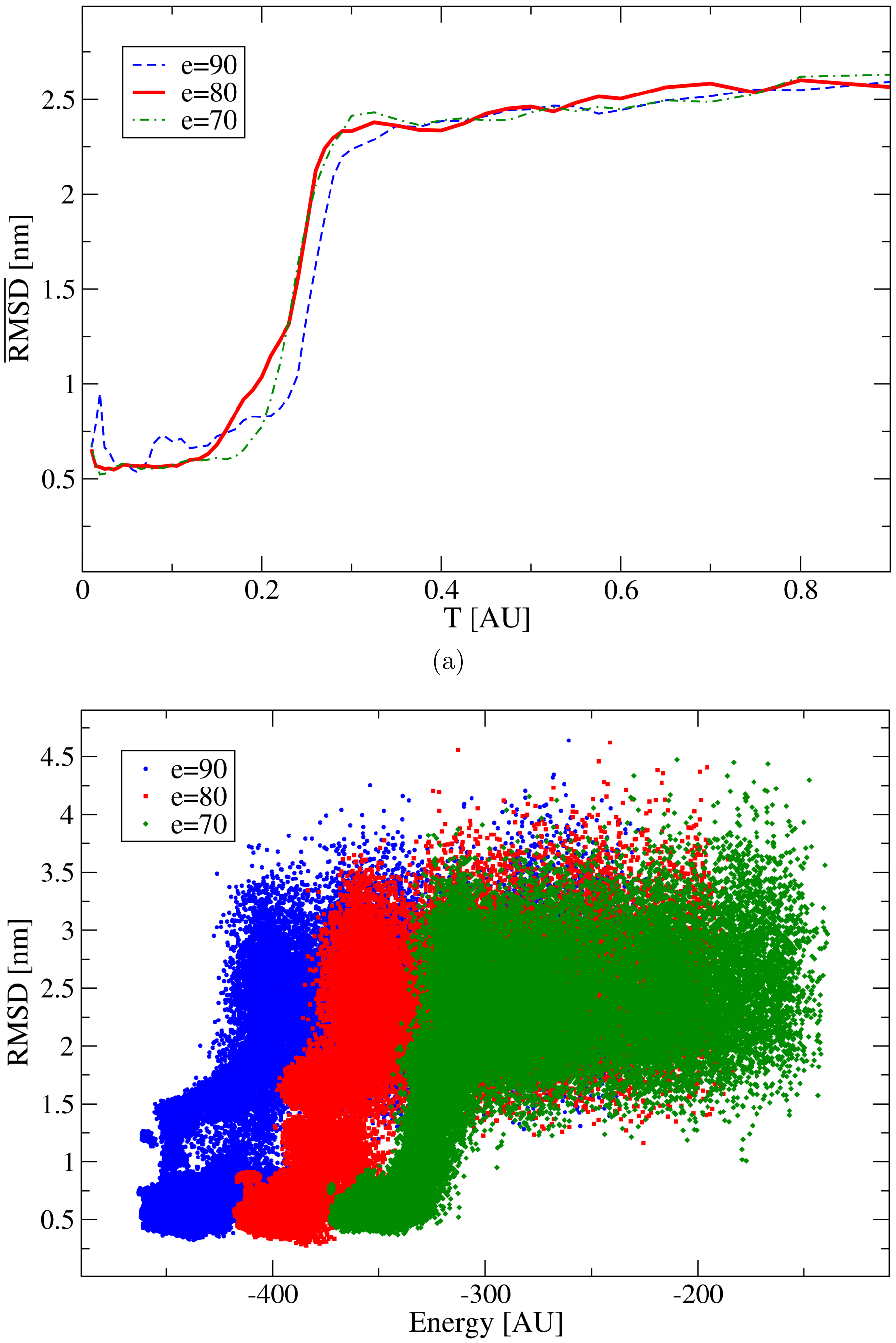
[2] We improved the way MonteGrappa handles the rotamers for the amino acids which contain aromatic rings.



(a)

(b)

Figure 3.8: 1BVC DI threshold parametrization results: (a) $\overline{\mathrm{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for $DI_0 = 0.01$, 0.0102, 0.0106, 0.011, 0.012, while the other parameters are fixed to $\alpha = 0.125$, $e = 80$, $r_c = 4.0\,\text{Å}$ with splice. The red continue line indicates the best DI_0 value, which is $DI_0 = 0.0106$.



(a)

(b)

Figure 3.9: 1BVC dihedral energy parametrization results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for e = 70, 80, 90, while the other parameters are fixed to $\alpha = 0.125$, $r_c = 4.0\,\text{Å}$ with splice, $DI_0 = 0.0106$. The red continue line indicates the chosen e value, which is $e = 80$.

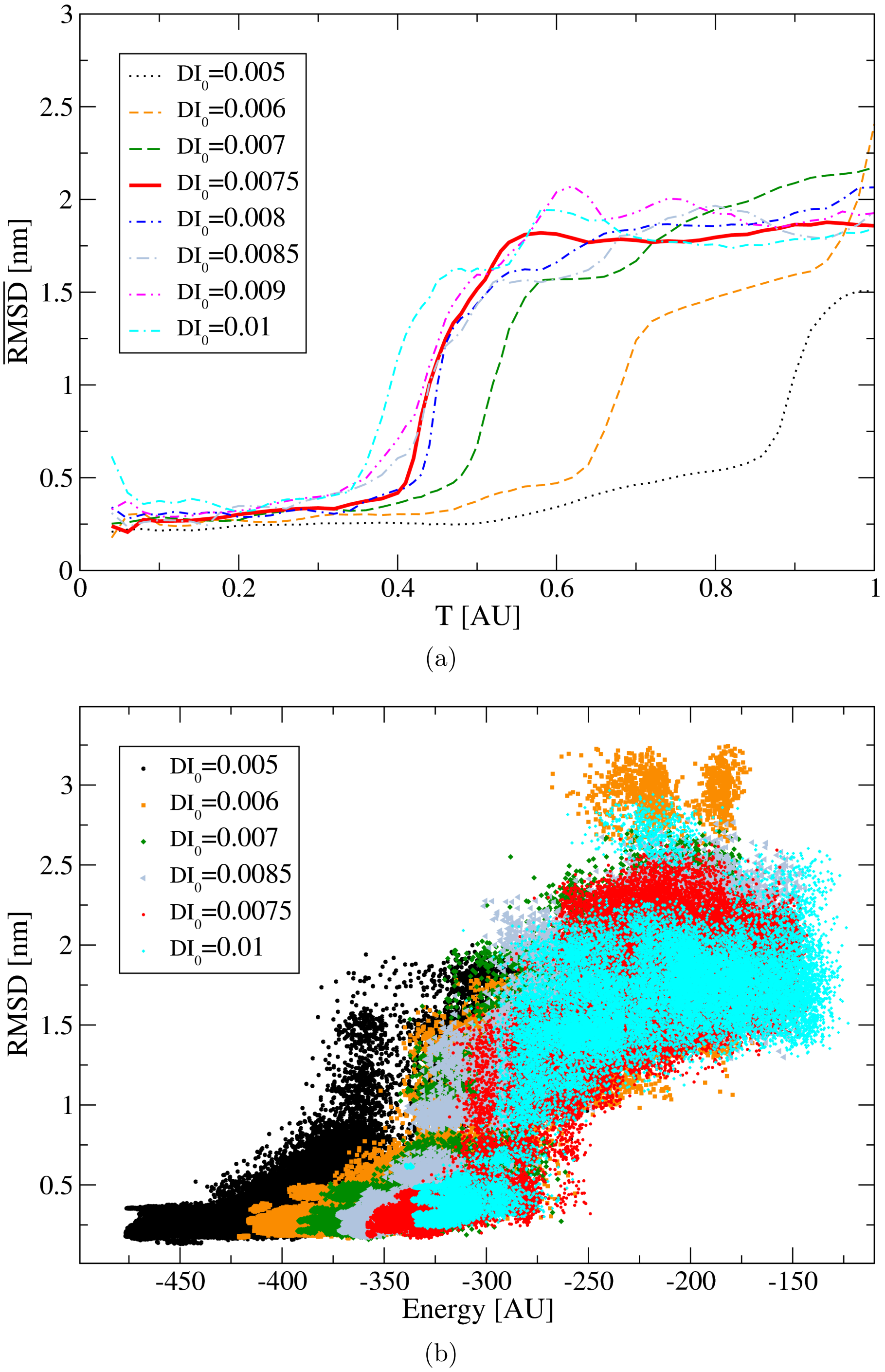
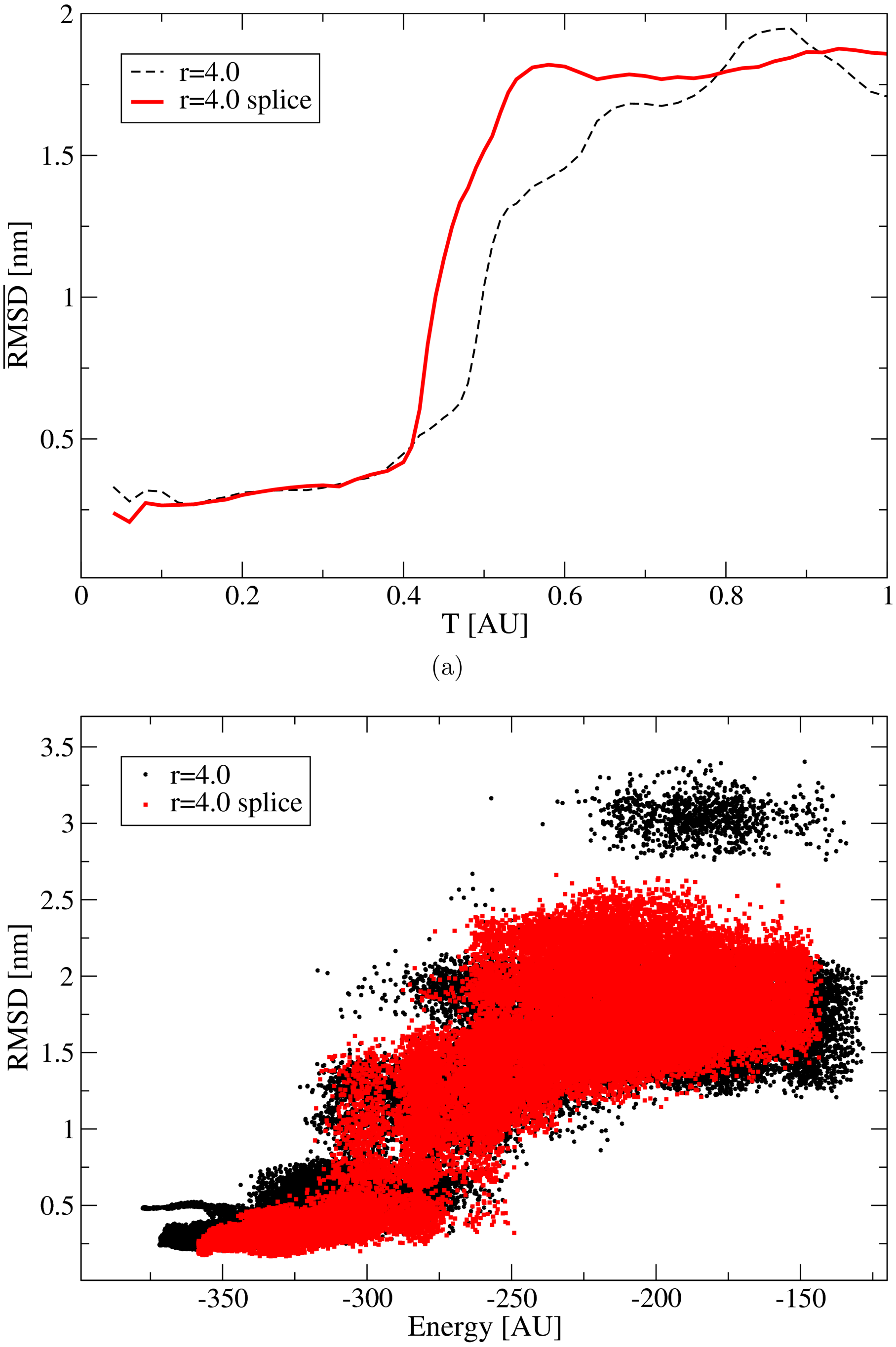


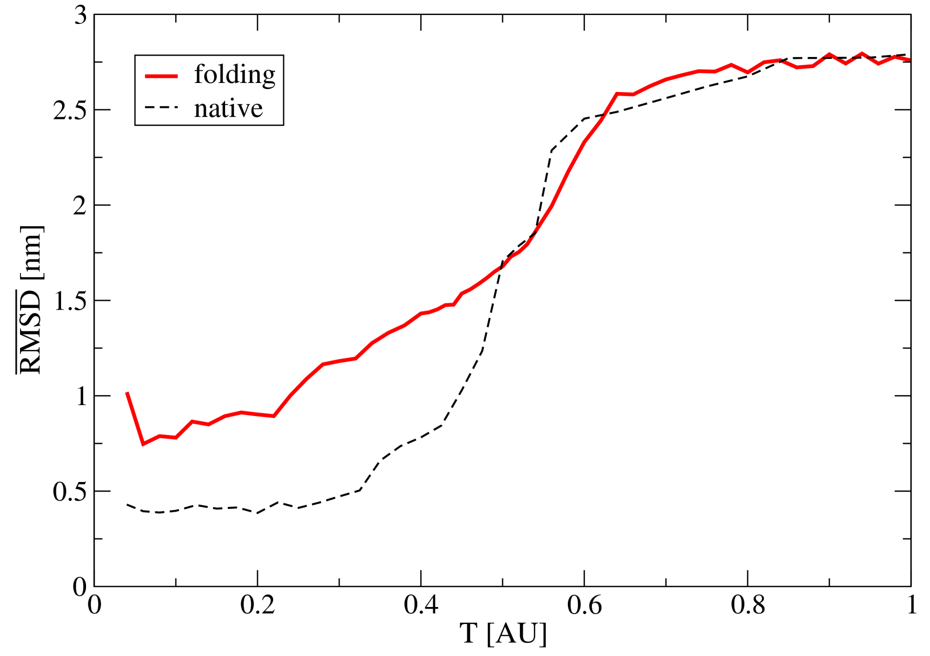
Figure 3.10: 1STN DI threshold parametrization results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for $DI_0 = 0.005$, 0.006, 0.007, 0.0075, 0.008, 0.0085, 0.009, 0.01, while the other parameters are fixed to $\alpha = 0.125$, $e = 80$, $r_c = 4.0\,\text{Å}$ with splice. The red continue line indicates the best DI_0 value, which is $DI_0 = 0.0075$.



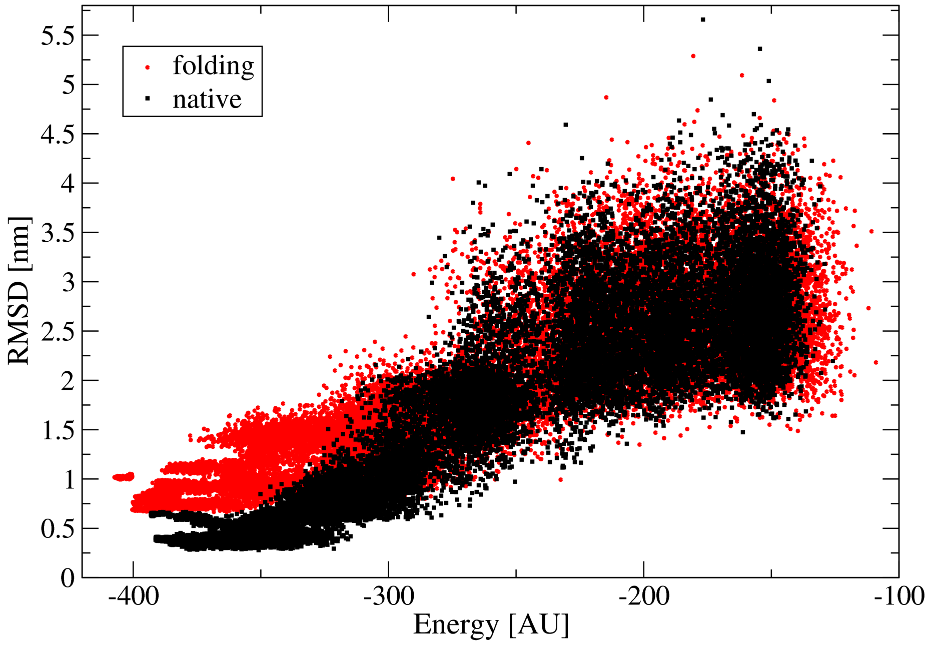
(a)

(b)

Figure 3.11: 1STN splice results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The parameters are fixed to $\alpha = 0.125$, $e = 80$, $DI_0 = 0.0075$ and $r_c = 4.0\,\text{Å}$ with and without splice. The red continue line indicates the best result, which is the rpesence of a splice.



(a)



(b)

Figure 3.12: Red (continue line) results of the 1STN starting from an unfolded conformation, black (dashed line) results of the same simulation started from the experimental native conformation. (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The parameters of the simulation are fixed to $\alpha = 0.125$, $e = 80$, $DI_0 = 0.0075$, $r_c = 4.0\,\text{Å}$ with splice

Protein	DI_0	TPfraction	k
1BPI	0.007	0.48	3.4
1STN	0.0075	0.45	4.0
1BVC	0.0106	0.49	3.6
1RQM	0.007	0.57	2.3

Table 3.1: Direct information best value for each protein, the corresponding TPfraction and mean number of contacts per residue k.

realize that we must improve our potential in order to enhance the formation of the $\beta-$sheets.

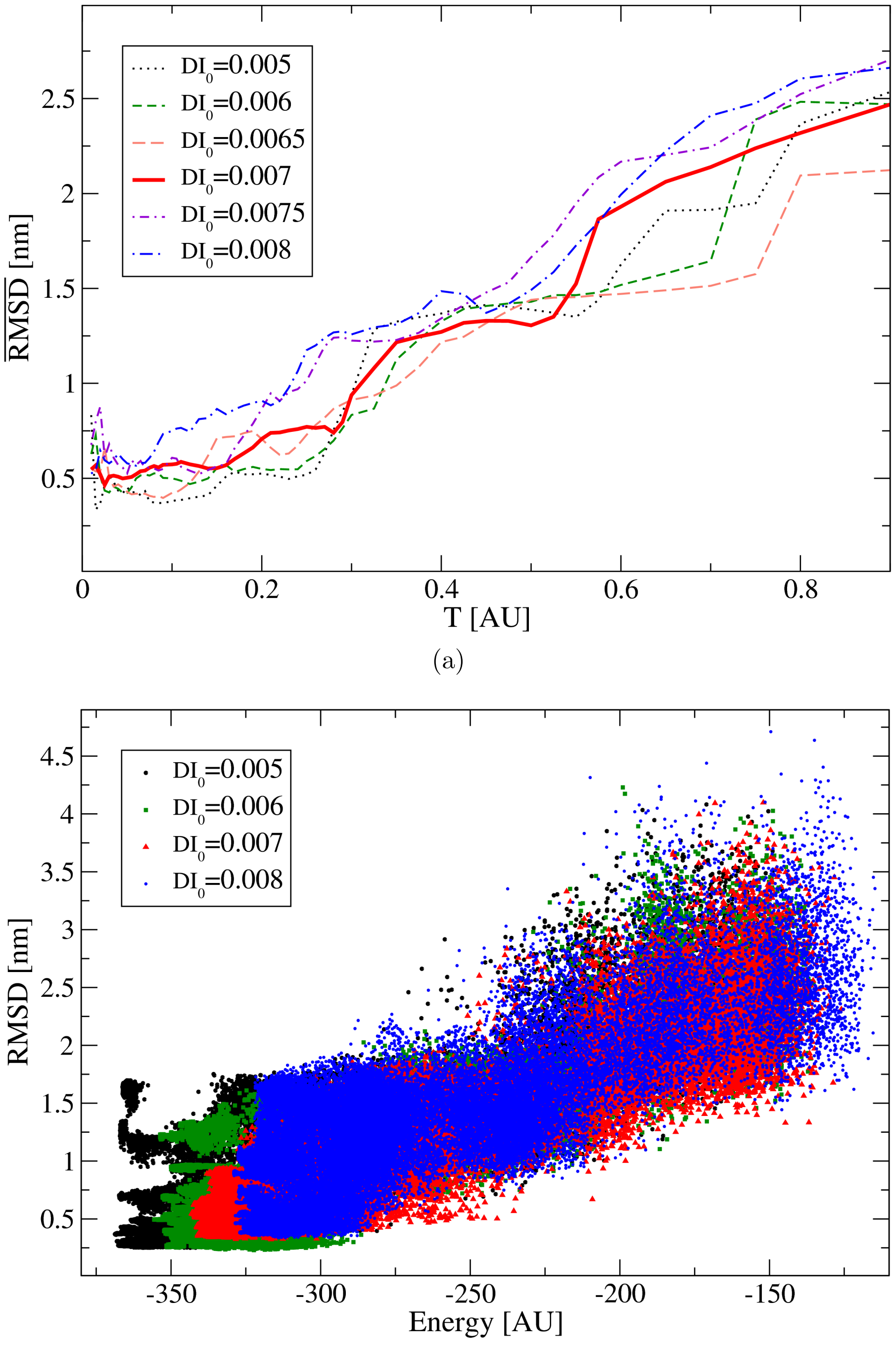
1RQM

The last system we consider is the 1RQM, another protein which presents both α and β structures. For this protein we vary DI_0 between 0.005, 0.006, 0.0065, 0.007, 0.0075, 0.008, and we choose as the best threshold $DI_0 = 0.007$ (see Fig. 3.13).

As exposed above, by means of these results we have to check if we are able to find a portable criterion for the DI threshold choice. In Table 3.1 are reported the k and TPfraction values corresponding to the best DI_0 values; by inspecting the quantities we can see how, with the exception of the 1RQM, the TPrate should be set around 0.5, while k around 3.7. However, considering the variability of these quantities among the systems, we do not have a solid criterion to set a priori the DI_0 value. Actually this is a limit of our model; we try therefore to go beyond the DI filter, implementing the strategies described in the next section.

3.2 Improvements of the model

The results obtained in the previous section indicate that the use of a coevolutionary potential is able to catch the main features of a protein structure, but also indicate that the form of the potential can be improved. We therefore try to improve our model first adding a term to the potential (1.4) which takes into account the effect of the hydrogen bonds; secondly we implement various filters on the two-body energies which could sobstitute the one based on the DI. These filters are tested on the 1BPI.



(a)

(b)

Figure 3.13: 1RQM DI threshold parametrization results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for $DI_0 = 0.005$, 0.006, 0.0065, 0.007, 0.0075, 0.008, while the other parameters are fixed to $\alpha = 0.125$, $e = 80$, $r_c = 4.0\,\text{Å}$ with splice. The red continue line indicates the best DI_0 value, which is $DI_0 = 0.007$.

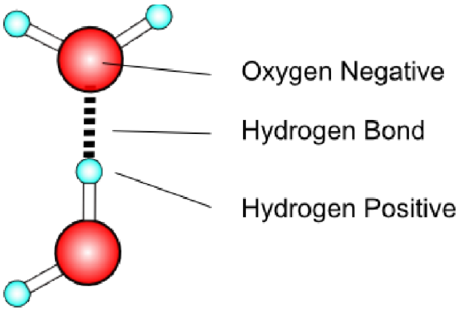
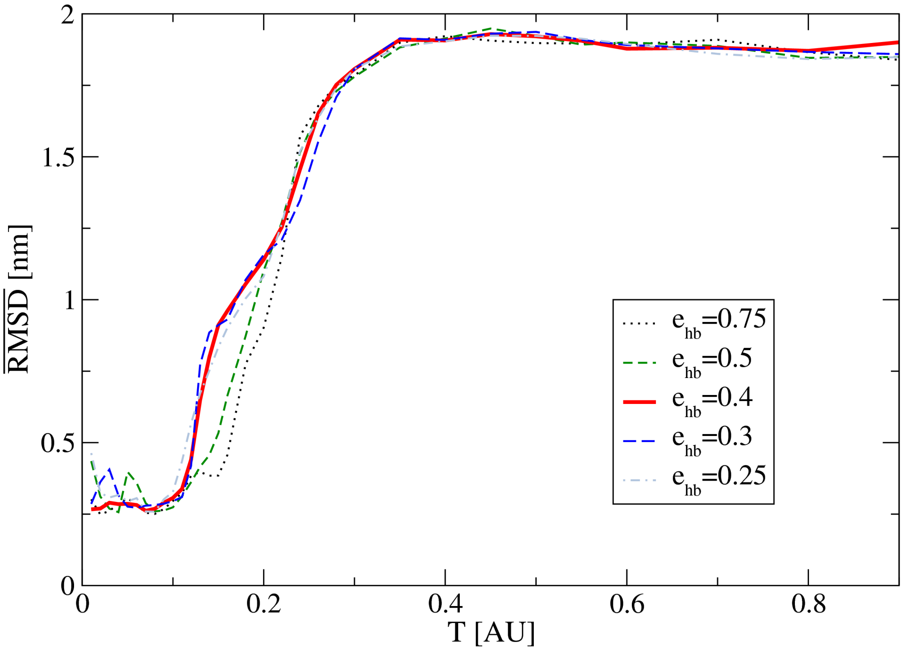


Figure 3.14: Representation of a hydrogen bond in water.

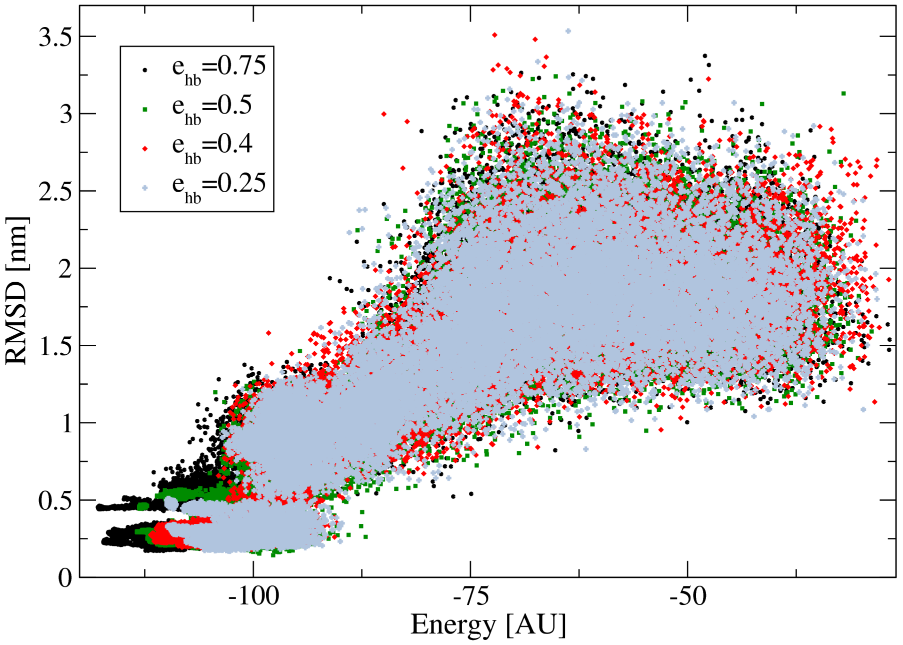
3.2.1 Hydrogen bonds

The secondary structures of a protein are stabilized also by the hydrogen bonds (HBs). These are bonds that occur between an hydrogen (H) bounded to a highly electronegative atom (such as nitrogen (N) or oxygen (O)) and another highly electronegative atom. The H atom has a partial positive charge (because it is bounded), while the other involved atom has a partial negative charge (because it is highly electronegative); if they come close to each other they experience an electrostatic attraction. Being an electrostatic interaction, the energy of this bond is higher than the one of a Van der Waals interaction, but lower than the one of a covalent bond. Such bonds occurs for example in water (Fig. 3.14). In a protein, the hydrogen bonds can arise between the H bonded to the backbone N atom and the O of the backbone carbonilic group of another amino acid. Thanks to their high associated energy, they play an important role in the stabilization of a protein (and in particular in the stabilization of the $\beta-$structures).

Up to this point we did not consider explicitely the HBs because they play basically the same role of the potential term which acts on the Ramachandran dihedrals but, according to the results of the folding simulations, we need an additional term to stabilize mainly the $\beta-$sheets. The HBs are a two-body interaction, but it is important to note that they cannot be automatically caught by a coevolutionary potential, as it is ours, where the two-body interaction energies are calculated basing on the analysis of the correlation patterns in a MSA. This because each amino acid is identical to the others from the HBs point of view: they all have the H and the O in the same position, so when in a site an amino acid is subsituted by another one, the H and the O do not change, therefore the associated change in energy is not due to the formation or to the break of a HB. This means that the correlation pattern would be the same even if all the HBs were neglected, so would be the two-body energies.

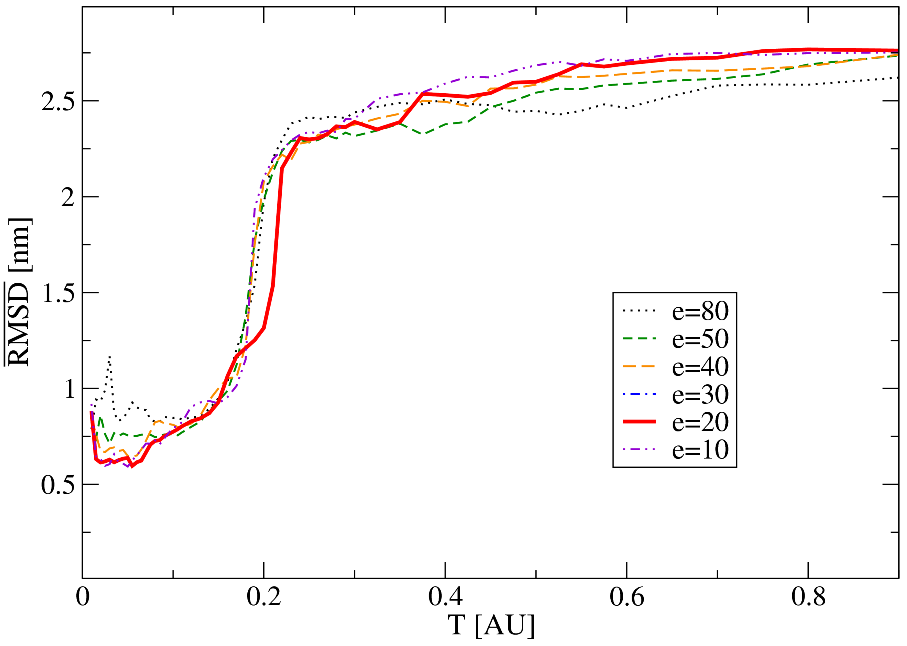


(a)

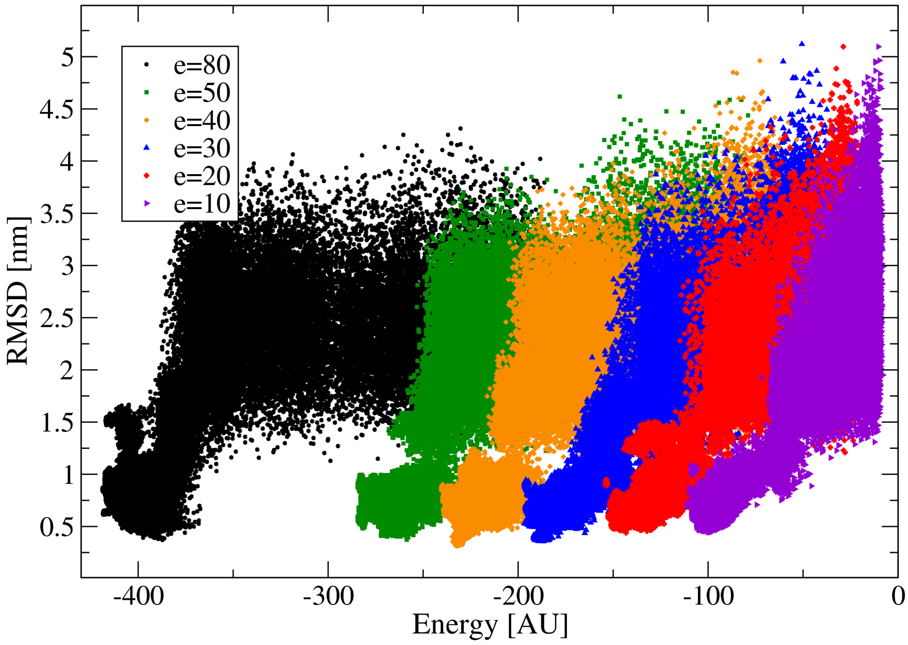


(b)

Figure 3.15: 1BPI hydrogen bonds energy parametrization results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for $e_{hb} = 0.25$, 0.3, 0.4, 0.5, 0.75, while the other parameters are fixed to $\alpha = 0.125$, $e_\alpha = e_\beta = 80$, $DI_0 = 0.007$, $r_c = 4.0$ with splice.

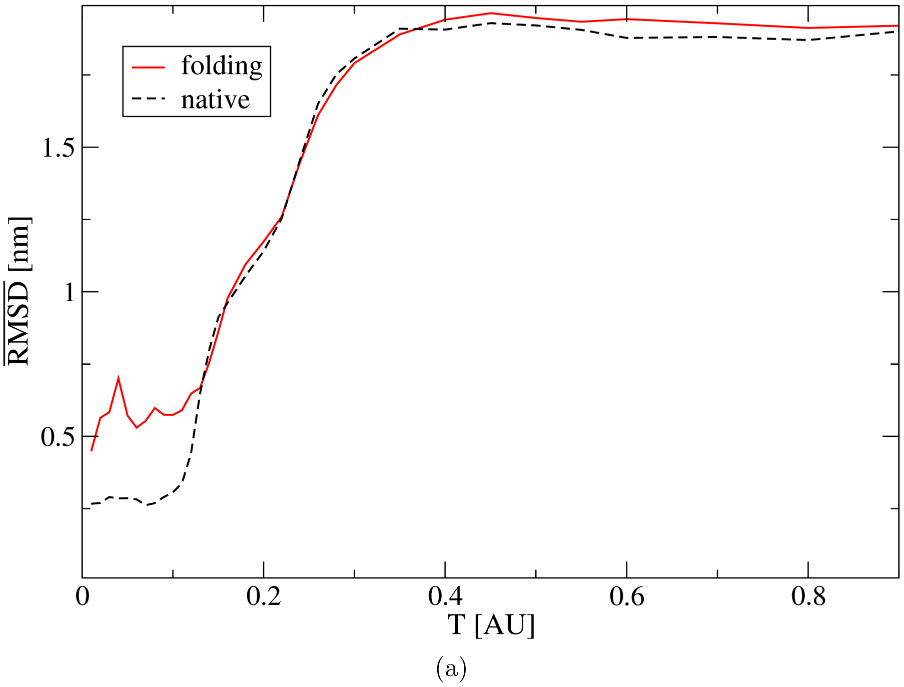


(a)



(b)

Figure 3.16: 1BVC dihedral energy energy parametrization results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for $e = 80$, 50, 40, 30, 20 and 10, while the other parameters are fixed to $\alpha = 0.125$, $DI_0 = 0.0106$, $r_c = 4.0$ with splice and $e_{hb} = 0.4$.



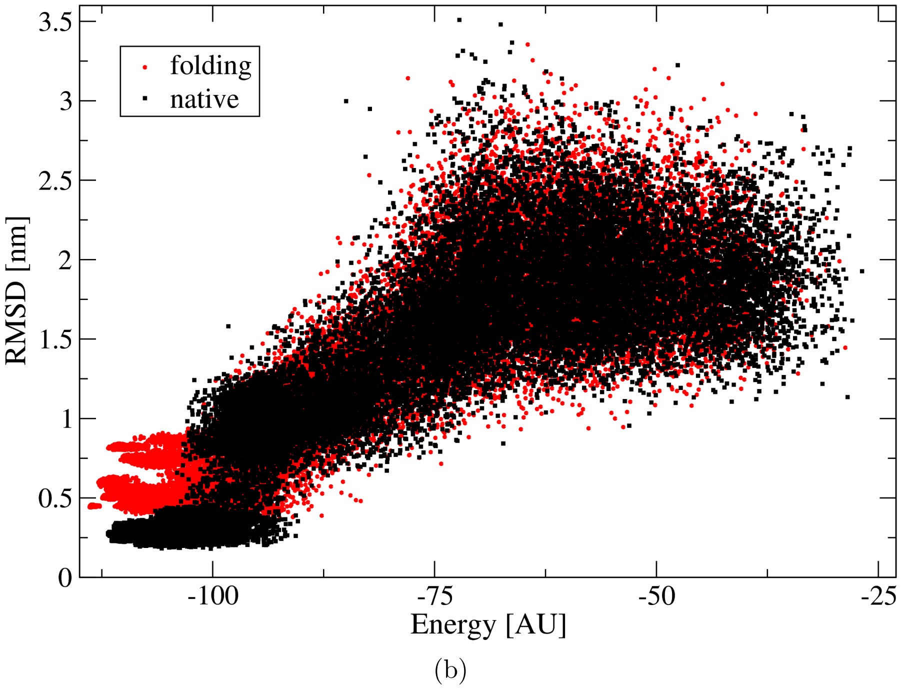


Figure 3.17: Red (continue line) results of the 1BPI starting from an unfolded conformation with the HBs term, black (dashed line) results of the same simulation started from the experimental native conformation. (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The parameters of the simulation are fixed to $\alpha = 0.125$, $e = 80$, $s = 0.007$, $DI_0 = 0.007$, $r_c = 4.0\,\text{Å}$ with splice, $e_{hb} = 0.4$.

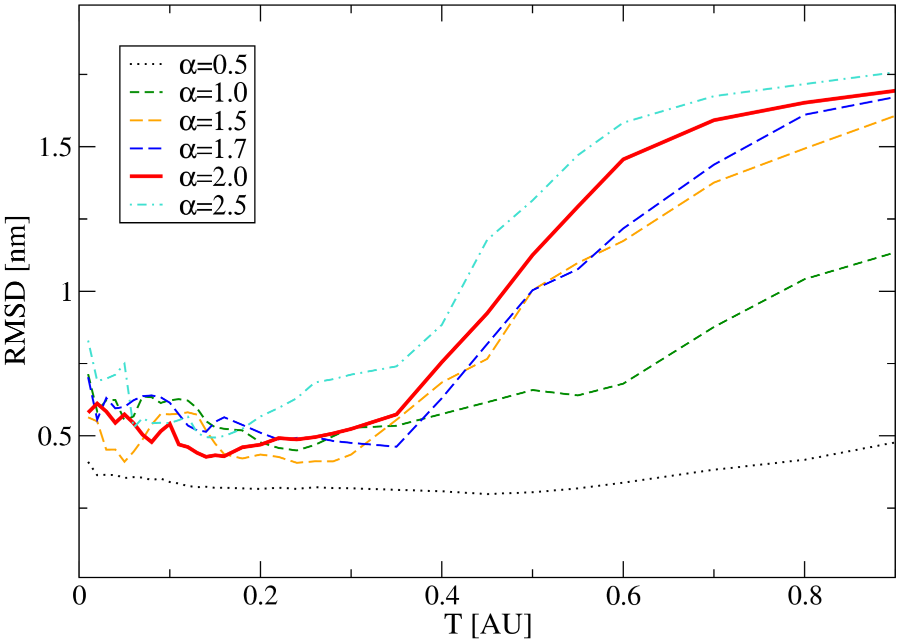
We therefore add an explicit term at the potential 1.4, which is

$$U_{hb} = -\sum_{k=1}^{N_H} \sum_{l=1}^{N_O} e_{hb} \Theta(r_{hb} - r_{kl}), \tag{3.3}$$

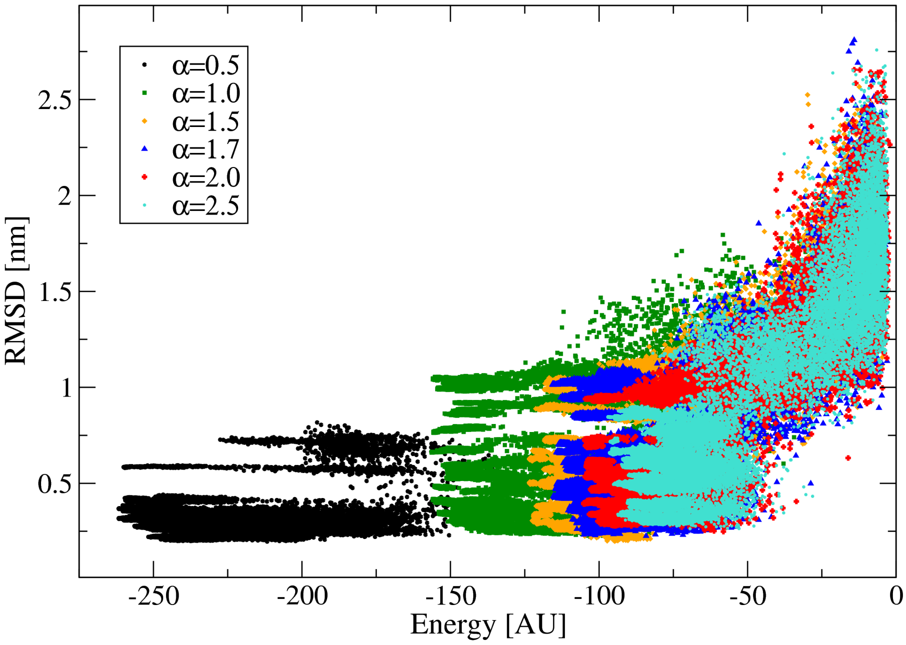
where N_H and N_O are respectively the number of hydrogens and oxigens involved in this kind of interaction, e_{hb} is the energy associated to the formation of a single HB, Θ indicates a step function which is 1 if $r_{hb} < r_{kl}$ and 0 otherwise, r_{kl} is the distance of the atoms and r_{hb} is the maximum distance at which an H and a O form an HB (we set $r_{hb} = 1\,\text{Å}$). This term is so a square well potential, whose depth e_{hb} has to be chosen by carrying out some simulations. By introducing this term we add a new type of atom, the hydrogen, which was not simulated up to this point. This because there are many hydrogens in a protein, which does not play a relevant role in the structure stabilization; we can therefore neglect them to have a significant gain from the computational point of view. The interaction energy between the hydrogens and all the other atoms is so set to zero, nor there is an hard-core repulsion, because these hydrogens are dummy atoms whose only aim is to form HBs.

In order to set the optimal e_{hb} value we perform a parametrization of this quantity for the 1BPI, by varying it between 0.25, 0.3, 0.4, 0.5, 0.75. According to the results shown in Fig. 3.15 we set $e_{hb} = 0.4$. By inspecting the RMSD relative to this value we can also see how the HBs improve the results of the simulations if compared to those obtained without them: they decrease the $\overline{\text{RMSD}}$ at low temperatures, and they remove the metastable state at low energies. Being the HBs' potential a term that stabilizes the secondary structures, we have to check if the dihedrals energy e_α has to be rescaled: we added this potential term to improve the formation of the $\beta-$sheets, but it also reinforces the $\alpha-$helixes, which were already stable enough; it is so possible that the dihedral term plus the HB term makes them too strong. We therefore perfom a new dihedral energy scan for the 1BVC, having set $e_{hb} = 0.4$, and we obtain the results shown in Fig. 3.16. The best simulation is the one in which we have set $e_\alpha = 20$, which is slightly worse than the one without HBs, but, since they add a physical feature to the potential, we keep them for the successive simulations.

The HBs should improve also the folding process, so we carry out a folding simulation on the 1BPI, and we obtain the results shown in Fig. 3.17. Even with the HBs the folding simulation cannot reproduce the native state as the only conformation at low energy; this is a limit of our model.

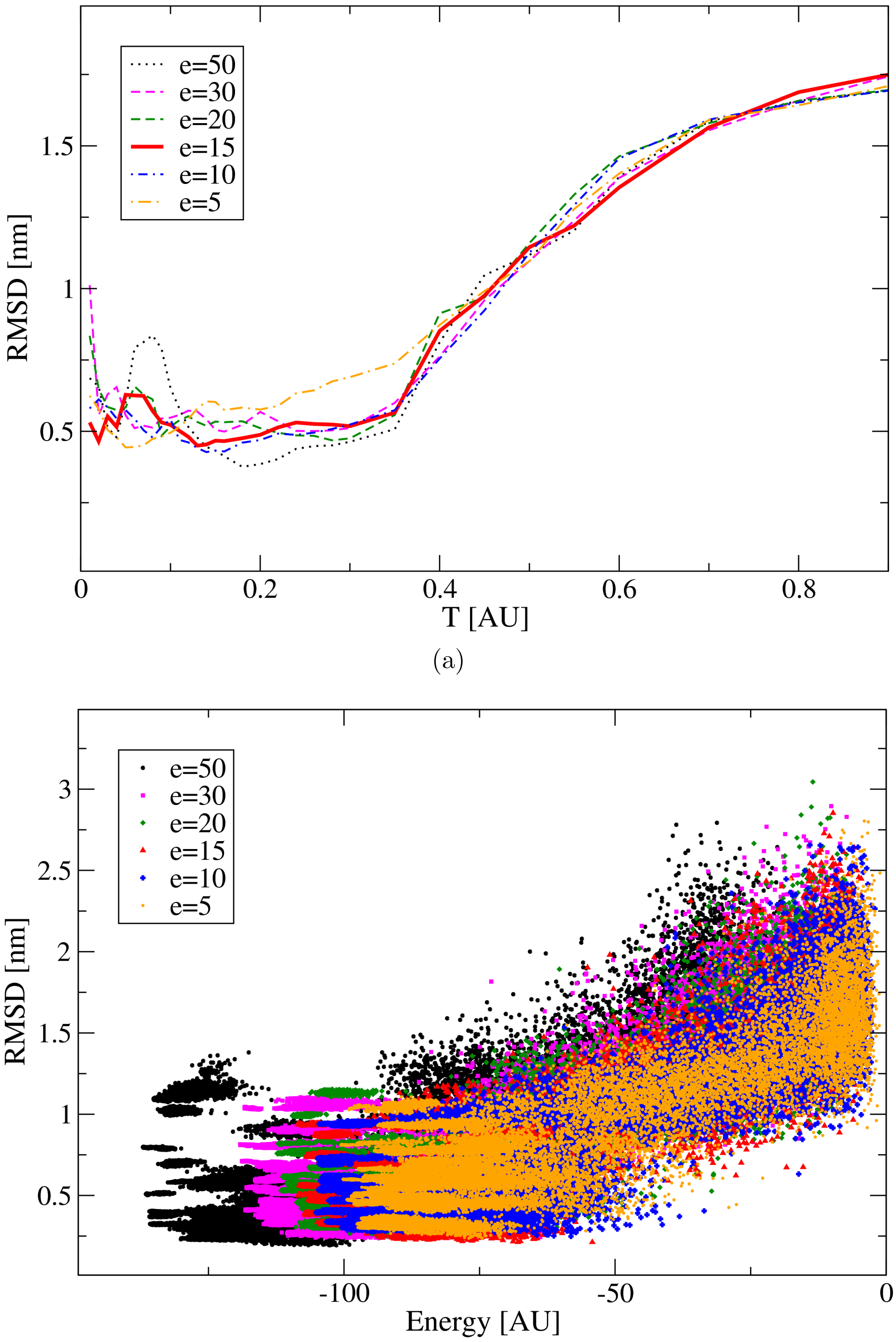


(a)



(b)

Figure 3.18: 1BPI α parametrization results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for α = 0.5, 1.0, 1.5, 1.7, 2.0, 2.5 while the other parameters are fixed to $e_\alpha = e_\beta = 10$, $r_c = 4.0$ Å with splice. The red continue line indicates the best α value, which is $\alpha = 2.0$.



(a)

(b)

Figure 3.19: 1BPI dihedral energy parametrization results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed for $e = 5$, 10, 15, 20, 30, 50 while the other parameters are fixed to $\alpha = 2.0$, $r_c = 4.0\,\text{Å}$ with splice.

3.2.2 Two-body energies reweighting

Being the optimal DI_0 value a quantity which is hardly portable among the systems, we try to find a way to not use this filter for the two-body energies. The simplest attempt is to not use any filter at all, but, from the results obtained from preliminary simulations, this has proved to be a bad solution. We have therefore to find a way to both take into account the Direct Information value and to not filter the contacts. Being the DI the fraction of Mutual Information which comes from the direct coupling alone, we reweight the two body energies by the DI/MI ratio: we compute the MI for each amino acid pair according to Eq. (2.3), and we define the new two-body matrix elements as

$$u'_{ij}(\sigma_i, \sigma_j) = u_{ij}(\sigma_i, \sigma_j)\frac{DI_{ij}}{MI_{ij}}; \tag{3.4}$$

then, by applying the two normalizations described in Sec. 2.1.1, we obtain the M_{ij} used in the simulations.

In order to test this method it is necessary to perform a new parametrization, because the two-body/dihedral/h-fields ratio is completely different from the previous one. We then perform a partial parametrization scanning the $\alpha-$values and the dihedral energies e_α and e_β with the methods described above: we vary α between 0.5, 1.0, 1.5, 1.7, 2.0, 2.5 and, after having fixed $\alpha = 2.0$, we vary the dihedral energy between 5, 10, 15, 20, 30, 50. The results are shown in Fig. 3.18 and 3.19, from which we can see how these are actually worse than the ones obtained by filtering the contacts.

3.2.3 Relative error filtering

As described in Chapter 2, being the statistics limited, the two-body energies u_{ij} obtained from CoCaInE are susceptible to error; bigger is the error, the more the matrix element is unreliable, so we implement a strategy to quantify its value, in order to filter the contacts retaining only the ones which present a small error value. First we adopt a bootstrap procedure [25]: from the MSA of the 1BPI we select randomly a number of sequences equal to the one of the original MSA, building thus a new modified MSA, in which some of the original sequences appear several times, while others do not appear. We repeat this twenty times, generating thus twenty different MSA; for each one we then run CoCaInE, and we obtain twenty two-body energy matrixes. For each matrix element u_{ij} we then calculate the mean value and its error. As for the DI, we calculate the TPfraction, both for the absolute error and for the relative one (now the pairs are ordered for increasing value of the error). From Fig. 3.20, we can see how the only meaningful

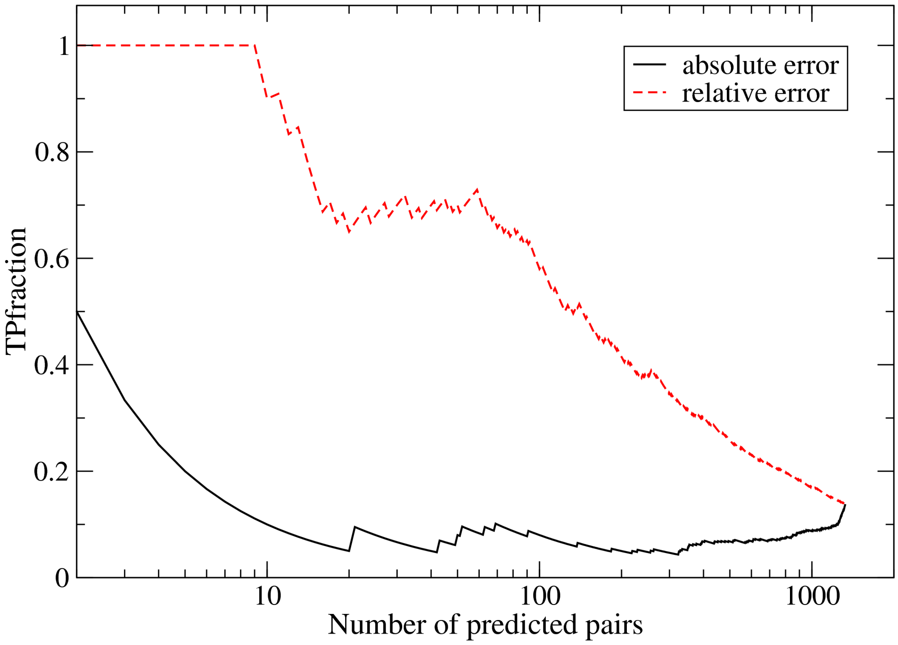


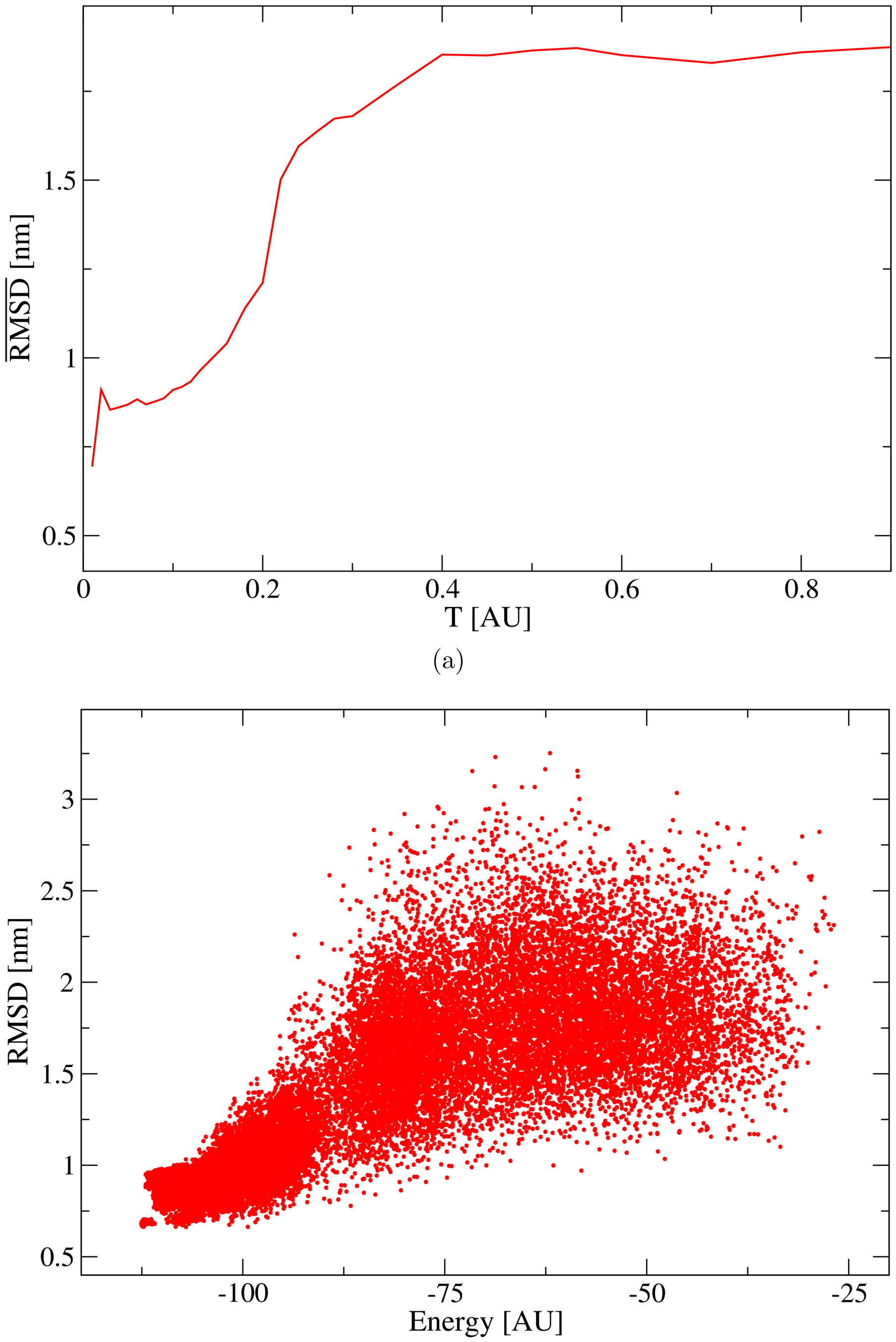
Figure 3.20: TPrate in function of the number of predicted pairs sorted by increasing absolute energy error (black solid line) and relative error (red dashed line).

quantity is the relative error, we therefore implement a filter on it: as done for the DI, we retain the u_{ij} which present a relative error below a threshold value RE_0. Finally we run a simulation in the 1BPI using the parameters $RE_0 = 0.08$, $\alpha = 0.125$, $e_\alpha = e_\beta = 80$ and $r_c = 4.0\,\text{Å}$ with splice ($RE_0 = 0.08$ was chosen because it corresponds to a TPfraction $\simeq 0.5$, which was found as the best value when using the DI filter). The results of the simulation are shown in Fig. 3.21; they are considerably worse than the one obtained using the DI filtering (Fig. 3.15), so we decide to not carry out a parametrization on the RE_0 value, because it is highly unlikely that we could obtain decent results.

3.2.4 PCA filtering

The last attempt of improvement of the model is the implementation of a Principal Component Analysis (PCA) [26] to obtain a valid filter on the two-body energies. For each amino acid pair (i, j) we consider three variables:

- the distance in residues between i and j, defined as $|i - j|$;
- the direct information, DI_{ij};
- the relative error of u_{ij}, RE_{ij}.



(a)

(b)

Figure 3.21: 1BPI relative error filtering results: (a) $\overline{\mathrm{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed using $RE_0 = 0.08$, $\alpha = 0.125$, $e_\alpha = e_\beta = 80$, $r_c = 4.0\,\text{Å}$ with splice, $e_{hb} = -0.4$.

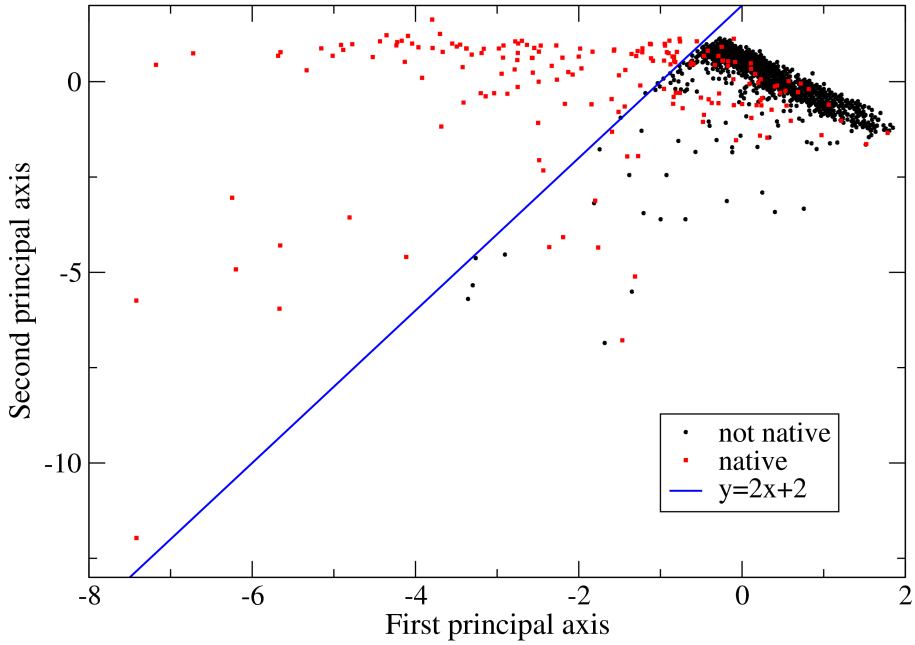
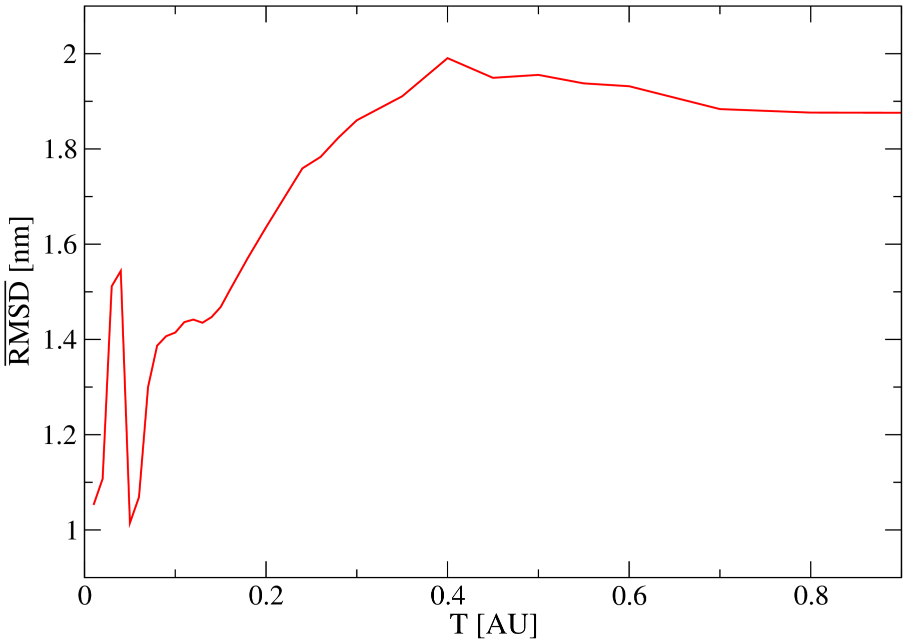


Figure 3.22: Projection on the first two principal axes of the triple (distance, DI, RE) of the points corresponding to the native contacts (red squares) and not native contacts (black circles). The blue solid straight line indicates the filter implemented to select mainly the native contacts.

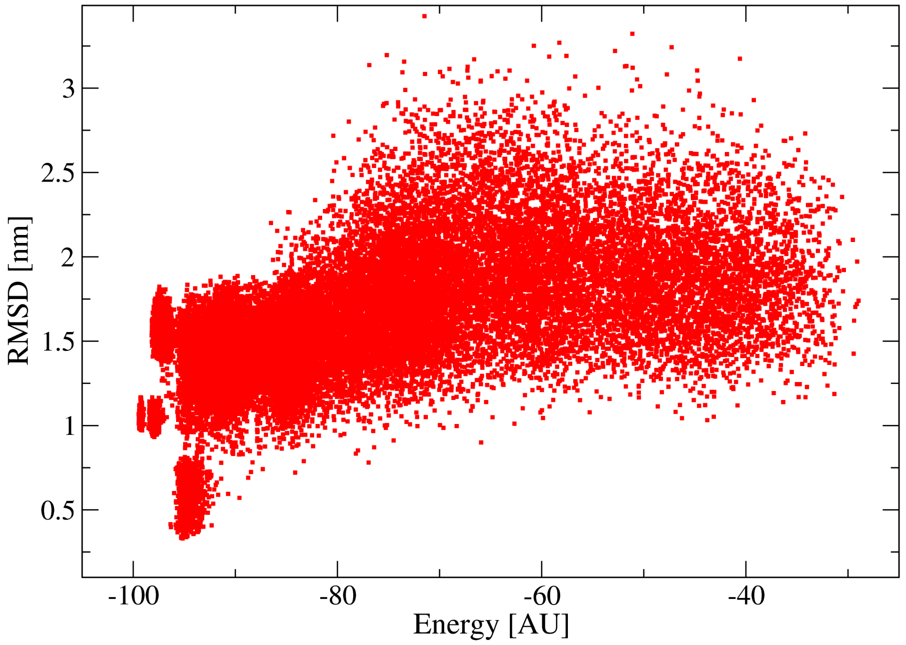
The idea is to find some linear combination of these variables that allows to identify the native contacts: that means to find a basis of the 3D-space generated by the three variables in which the points representing the amino acid pairs divide in two clusters, one corresponding to the native contacts and the other to the non-native ones. By using directly the basis generated by the three considered variables, we are not able to separate the native contacts from the others, so we implement a PCA procedure. Using it we find a basis in which the original (possible correlated) quantities are represented in the basis in which they are linearly uncorrelated. By projecting the 3D-space into the first two components (Fig. 3.22), we can see how a partial clusterization is obtained: the most native contacts live above the line

$$f(x) = 2x + 2. \tag{3.5}$$

We therefore implement a filter which retains only the amino acid pairs whose representative points live above this line, and we obtain a TPfraction $\simeq 0.97$. Finally we run a simulation by using the optimal set of parameters ($\alpha = 0.125$, $e_\alpha = e_\beta = 80$ and $r_c = 4.0\,\text{Å}$ with splice). The results are shown in Fig. 3.23; as before, they are sharply worse the the ones obtained by using the DI filter.



(a)



(b)

Figure 3.23: 1BPI dihedral PCA filtering results: (a) $\overline{\text{RMSD}}$ in function of the temperature, (b) RMSD versus the energy of the corresponding conformation. The scan is performed using $\alpha = 0.125$, $e_\alpha = e_\beta = 80$, $r_c = 4.0\,\text{Å}$ with splice, $e_{hb} = 0.4$.

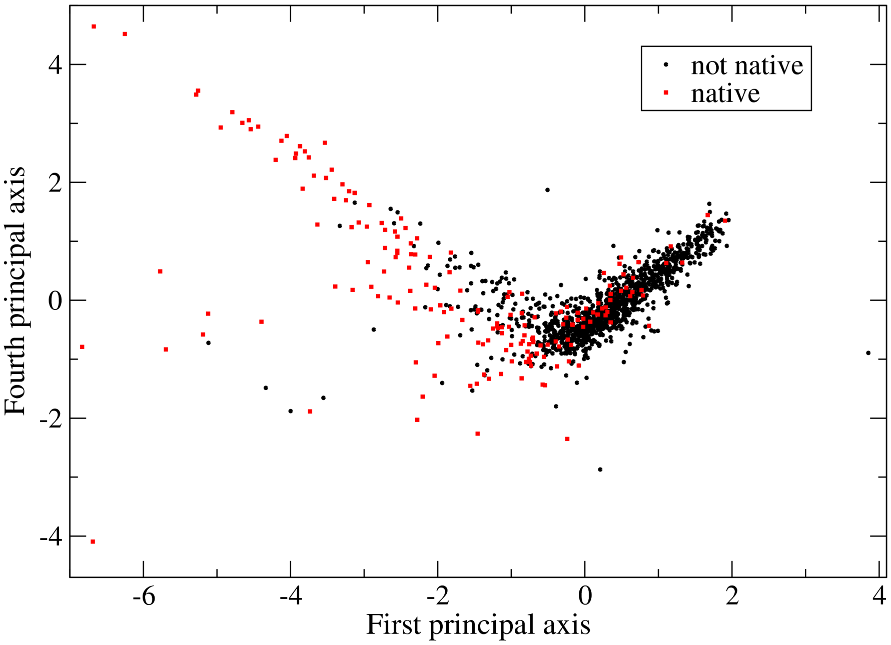


Figure 3.24: Projection on the first and on the fourth principal axes of the quadruple (distance, DI, RE, energy) of the points corresponding to the native contacts (red squares) and not native contacts (black circles).

In spite of the optimal TPfraction value the results are bad, this means that the TPfraction is not completely able to catch the ideality of the filtering. In fact we analyzed the contacts retained by the DI filter and by the PCA filter and we have found a substantial difference: by fixing $DI_0 = 0.007$ we have 104 native contacts, of which 50 relatives to pairs (i, j) such that $|i - j| > 2$ (meaning that they truly interact in MonteGrappa), while for the PCA filter we have 115 native contacts, of which only 31 truly interact. The DI filter is therefore more efficient in identifing non-local contacts.

Finally we try to improve the PCA analysis by adding the two-body energy u_{ij} as fourth variable. We perform the same PCA analysis and, by projecting the 4D-space into the varius combinations of subspaces, we find that the most representative projection is on the first and on the fourth principal axes (Fig. 3.24); however there is not a significant clusterization, so we definitively abandon this filter, and we confirm the DI filter as the best one.

Chapter 4

Aggregation

The final part of this thesis is devoted to a preliminar study of the aggregation process between different copies of the same protein. It is experimentally known that some proteins can cluster together forming different types of aggregates. The aggregation is a phenomenon that is biologically relevant when the proteins are susceptible to misfolding or unfolding, which lead hydrophobic regions to be exposed to the solvent. In such cases, the exposed hydrophobic regions of two proteins may interact to minimize their exposition, forming an aggregate. There are several conditions that can lead a protein to a misfolded/unfolded state: for example one is the enviromental stress, such as an extreme temperature or an extreme pH value [27]. The aggregation is of fundamental medical interest, because many diseases, such as Alzheimer's or Parkinson's disease, are related to this phenomenon [28].

The protein 1BPI can undergo such a process, forming a decamer [29, 30, 31], whose properties are not fully understood. Being our approach suitable to study large protein systems, we implement a strategy to reproduce this phenomenon for

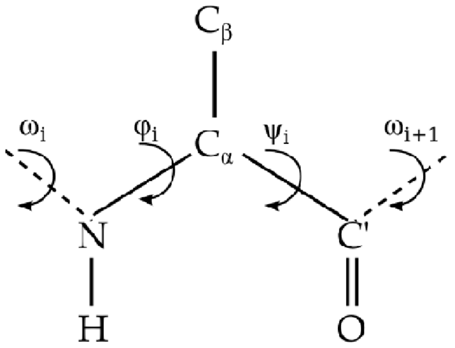
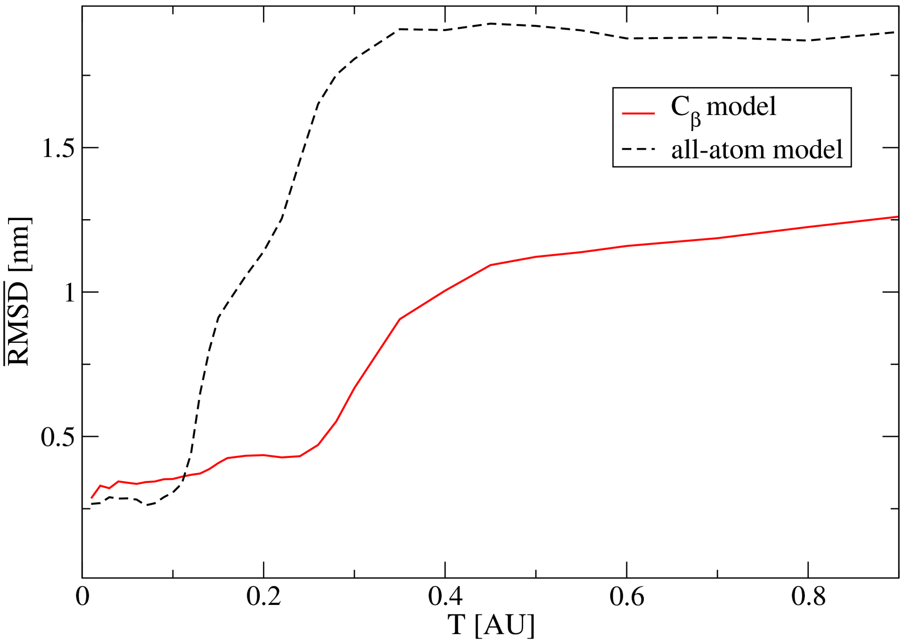
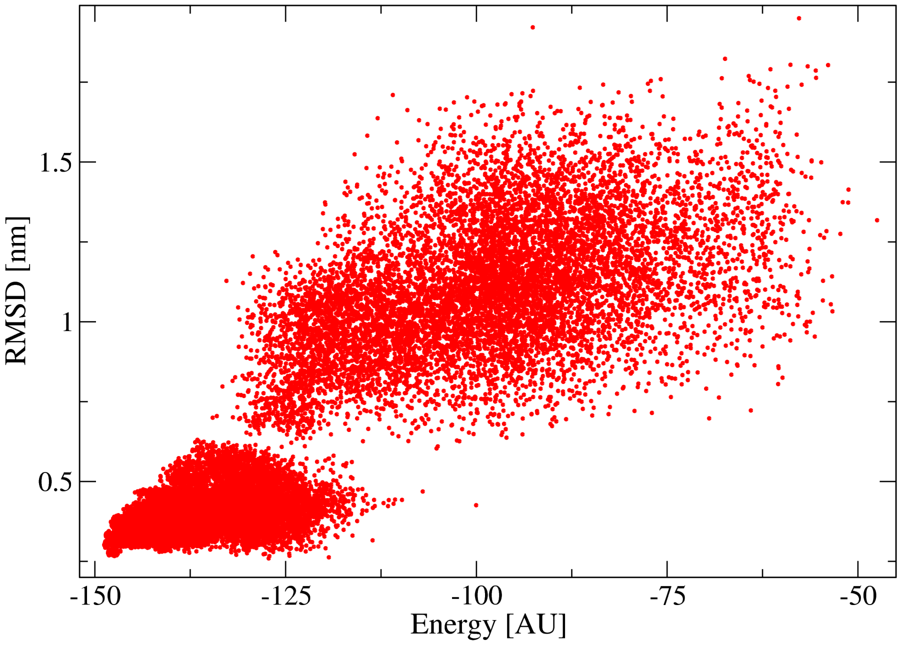


Figure 4.1: Schematic representation of an amino acid in the C_β model.



(a)



(b)

Figure 4.2: 1BPI C_β model results: (a) $\overline{\text{RMSD}}$ in function of the temperature for the all atom model (black dashed line), and for the C_β model (red solid line); (b) RMSD versus the energy of the corresponding conformation for the C_β model.

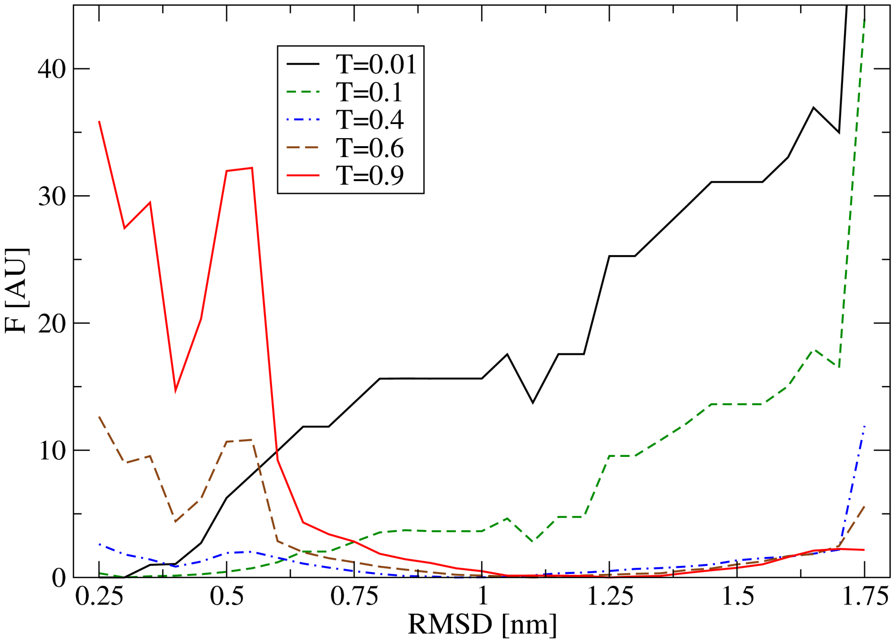


Figure 4.3: 1BPI free energy profile in function of the RMSD, at the temperatures $T = 0.01$, 0.1, 0.4, 0.6, 0.9.

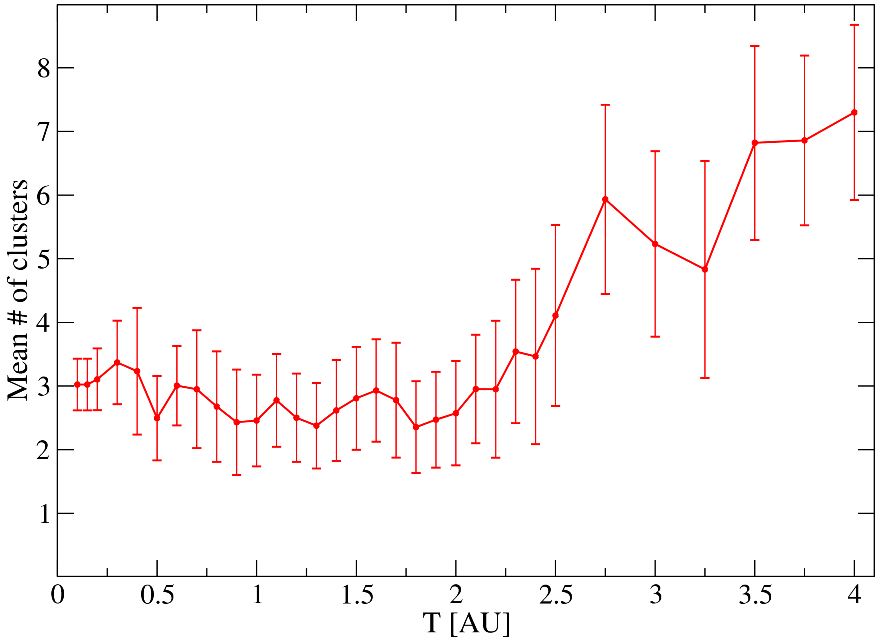


Figure 4.4: Mean number of clusters in function of the temperature for the 1BPI aggregation simulation. The error bars show the fluctuations between the different clusterization states.

the 1BPI. We simulate a system composed by ten copies of the 1BPI in a box[1] of side 15 nm; the side has been chosen in such a way to obtain a realistic molarity of 5 mM. Being now the system ten times bigger, we semplify the model in order to accelerate the simulations: we move from an all-atom description to a C_β one, doubling thus the speed of the simulations. Now the properties of the amino acid sidechiains are summarized in their C_β, artificially constructed in the geometric center of mass of the sidechain (an amino acid is now modelized as shown in Fig 4.1). The two-body term of the potential and the h-fields one have to be consequently slightly modified: the matrix elements u_{ij} do not have to undergo the normalization over the maximum number of atomic contacts; the h-fields $\tilde{h}_i$ now have to be normalized over the maximum number of contacts made between the C_β in all the systems. Moreover, it is necessary to perform a new complete parametrization on a single 1BPI, because the system is significantly different. We carry out the parametrization with the methods described in Section 3.1, and we obtain the optimal parameter set for this model:

- $\alpha = 0.4$;
- $e_\alpha = e_\beta = 80$;
- $r_c = 6.5\,\text{Å}$;
- $DI_0 = 0.007$;
- $e_{hb} = 0.5$.

The results of the simulation carried out with this optimal set is shown in Fig. 4.2. By inspecting it we can see how the $\overline{\text{RMSD}}$ value is slightly worse than the one corresponding to the best all-atom simulation, however the results are still good. In addition we compute the free energy profile in function of the RMSD for some temperature values (Fig. 4.3) by means of the WHAM method [32, 33]. For low temperatures the free energy minimum is in correspondence of low RMSD values, while increasing the temperatures it is shifted to higher RMSD. These results indicate that even a simplified C_β model is able to catch the main thermodynamic features of the system, so we can actually study the aggregation process within this semplification.

Finally we run an aggregation simulation; in doing this we add some new Monte Carlo moves besides the ones described in Appendix A, whose keywords are

[1] The presence of a box is essential, because if the system were not confined, the equilibrium state would be the one in which the chains are infinetely distant (because the entropy diverges).

- *BackRub*, which is a flip of the backbone around the axis defined by two non-consecutive C_α;
- *BackSideRub*, which is a combination of a backrub and of a sidechain move;
- *MoveCom*, in which a random chain is shifted;
- *ComCluster*, in which a cluster is shifted.

The first two are *local* moves, meaning that they change the position of some localized atoms, and they have been added to let the amino acids adjust their position within a cluster; the last two let the chains move and join together. The temperature range is $0.1 < T < 4.0$; the highest temperature has been chosen to let the equilibrium state be the one in which all the monomers are separated.

The inspection of the Monte Carlo trajectories reveals a partial clusterization of the proteins. In particular, as shown in Fig. 4.4, we compute the mean number of clusters[2] in function of the temperature, where the mean is calculated starting from the step 10^8 to the step $13 \cdot 10^7$. The error bars represent the standard deviation from the mean value, and indicates thus the fluctuations among the different clusterization states. For the lowest temperatures we observe that proteins join together forming one, two or three clusters, while increasing the temperatures the number of disjoint clusters grow (because the entropy wins over the energy). These preliminar results, although are not compared to experimental observables, suggest that the aggregation process can be studied within this framework.

[2] Two chains belong to the same cluster if there is at least an atomic contact between them, with $r_c = 6.5\,\text{Å}$.

Chapter 5

Conclusions and outlook

The potential we have developed starting from a coevolutionary model is able to catch the main structural features of the investigated proteins. The parameters on which it depends seem portable among different systems, with the limits related to the DI threshold exposed in Chapter 3. The coevolutionary quantities (two-body energies and h-fields) can be easily obtained for any protein belonging to a large enough family, so the potential can be used to study a large number of proteins. Studies on different proteins could also lead to an effective *a priori* choice of the the DI threshold, or to the development of an alternative contact filtering scheme. Once this critical point will be solved, our potential will be able to give information about the native state of some unknown protein, since all the paramenters are set without any reference to it. Furthermore, the last investigations on the 1BPI show that even a simplified C_β model can be used to obtain significant results. This is an important fact especially when our coevolutionary potential is used to study large proteins or protein systems, as we have done in the last part of the thesis. The preliminar simulations we have carried out on the aggregation show that the potential is able to partially reproduce this phenomenon: as outcome, subsequent studies will focus on the aggregation, developing strategies to reproduce the physical conditions that lead to it within our framework, and comparing some observable with experimental data.

Appendix A

Overview of MonteGrappa

The code we use to implement our model is MonteGrappa[34]; it is a C written program which implements the Monte Carlo method described in Sec. 2.2. Among the many possibilities offered by the program, we use the parallel tempering technique at a fixed potential, which is implemented through the MPI libraries. To run, MonteGrappa needs three input files: `file.par`, `file.pol` and `file.pot`; the `.par` contains all the parameters of the simulation (e.g. the number of Monte Carlo steps, the temperatures of the replicas etc.), the `.pol` contains all the geometric information of the starting configuration of the system, while the `.pot` contains the information regarding the potential which leads the dynamics. While the `.par` is prepared by hand, the `.pol` and `.pot` are generated by a supplementary tool, Grappino (distributed alongside MonteGrappa): giving it as input a structure in the PDB format and all the parameters that appear in the potential, it translates the structure and the potential in a MonteGrappa readable format. The potential parameters are chosen as described in Chapter 2, while the technical parameters of the simulation are chosen as described in Sec A.1. We use three types of Monte Carlo moves in the simulations, whose keywords are

- *Pivot*, in which a random dihedral is changed by a random angle;
- *Multiple Pivot*, which is a sequence of pivot moves applied to an arbitrary number of consecutive residues;
- *Sidechain*, in which the residues of the amino acids are moved in one of their possible conformations (the so-called *rotamers*);

the moves are summarized in Fig. A.1.

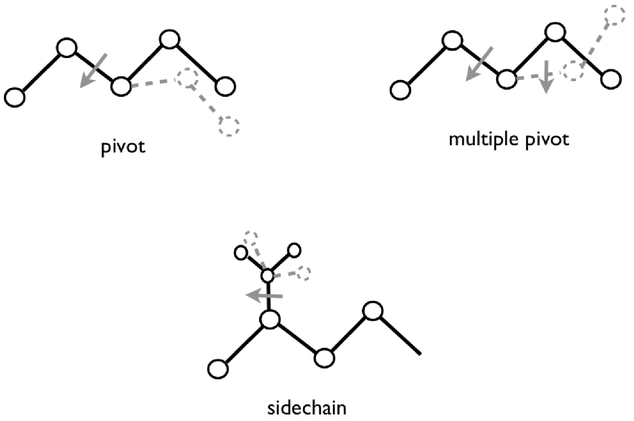


Figure A.1: Scheme of the Monte Carlo moves used in the simulations.

A.1 Technical parameters

A crucial point in all the parametrization process is that each simulation has to reach a reasonable equilibrium state, in the sense that we must let the proteins sampling completely the phase space. Actually, we cannot be sure that the equilibrium condition is reached within a finite-length simulation, what we can do is monitor an observable of the system, for example the energy, and see when it stops varying significantly. At this stage we have a necessary but not sufficient condition to check the equilibration state of the systems. Keeping this in mind, we have to fix the technical parameters of the simulations in order to reach the "equilibrium" state in the shortest possible duration. Since we use a parallel tempering, the main parameters are:

- the number of replicas N_r;
- the number of Monte Carlo steps n_s;
- the set of temperatures T_i with $i = 0..N_r - 1$.

The set of temperatures should have a range wide enough to catch both the native state and the denatured one as equilibrium conformations (in order to see the folding point and to facilitate the equilibration), and narrow enough to allow the replicas exchange efficiently. While the former condition is a system property, the latter can be achieved by having a high number of replicas; we therefore choose N_r as the biggest number of cores we can use simultaneusly, that is, depending on the machine, $N_r = 56$ or $N_r = 32$. By carrying out some

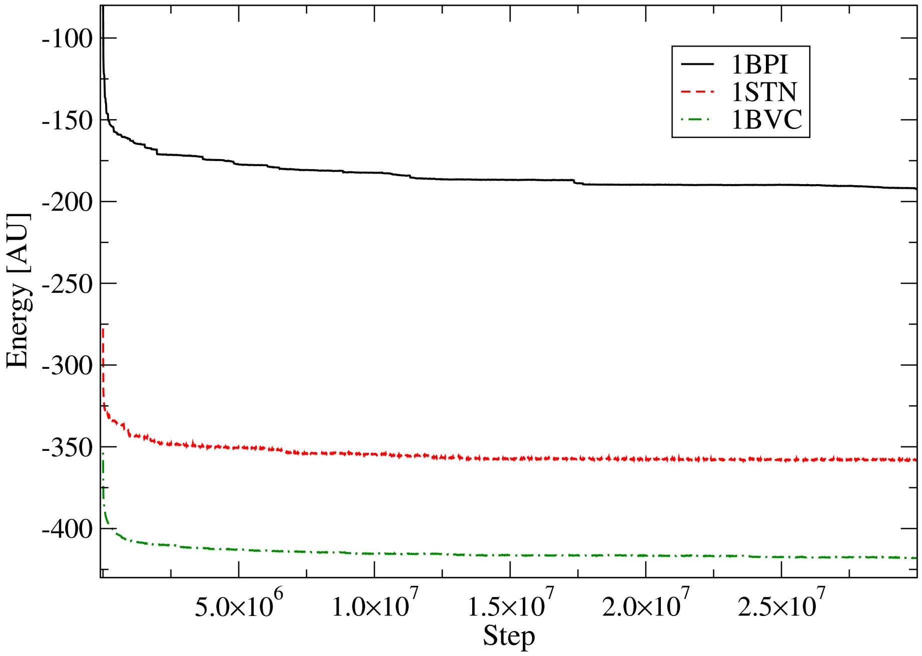


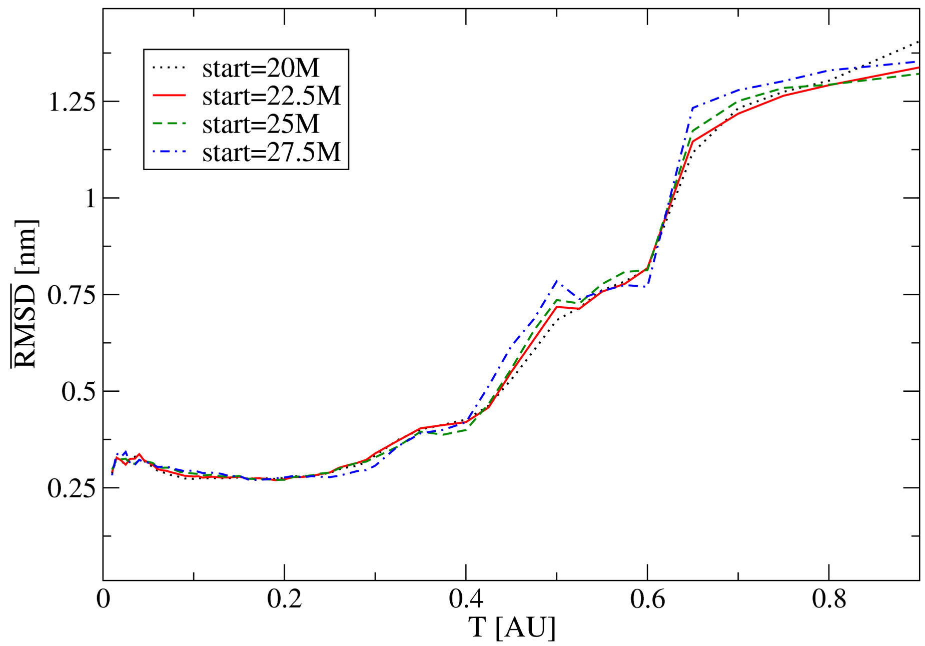
Figure A.2: Energy of the lowest replica in function of the Monte Carlo time for the 1BPI, 1STN and 1BVC.

preliminar simulation we also fix the temperatures range[1] (in principle different for each protein), and we find that $0.01 < T < 1.0$ basically fits all the systems under investigation (we vary it sligthly switching from system to system); for the reasons explained in Sec. 2.2, the low temperatures are set close to each other, while the higher ones are more separated. The last technical parameter is n_s; we have to find its minimum value that allows the systems to converge to equilibrium, and that let us make a statistical significant analysis. Obviously this parameter depends critically from the initial conditions of the simulations, so we choose (for the first part of the parametrization procedure) to start the simulations from the experimental native structure of the proteins (taken from the PDB), in order to minimize the time required by the equilibration process[2]. Once again we perform some simulations on different systems, and we find that $n_r = 3 \cdot 10^7$ represents a good trade off between the time needed by a simulation and the realization of the equilibrium state (in the sense exposed above). By inspecting the energy of the lowest replica in function of the Monte Carlo time for some simulations carried out on different proteins (see Fig. A.2) we find how the relative energy change

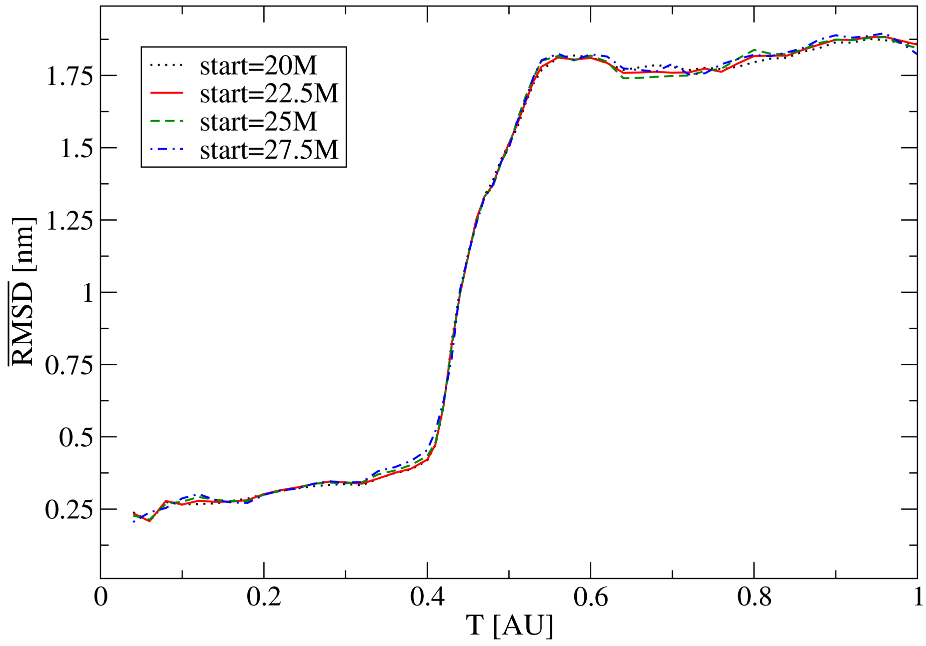
[1] Note that, since we fix $k_b = 1$, the temperature is measured in the same (arbitrary) measurements unit of the energy.

[2] This is actually true if our potential gives rise to a the free energy profile which shows a minimum in the protein native state.

between the step $2 \cdot 10^7$ and the end of the simulation is very small, so we decide to calculate all the equilibrium properties starting from this point. In addition to this we calculate the $\overline{\mathrm{RMSD}}$ starting at different steps of the simulations, for two simulations carried out on the 1BPI and on the 1STN; as it is shown in Fig. A.3, the results do not substantially change varying the starting point of the calculation, meaning that we have reached a good equilibration.



(a)



(b)

Figure A.3: $\overline{\text{RMSD}}$ versus the temperature for the 1BPI (a) and for the 1STN (b) calculated starting from different steps of the simulations.

Bibliography

[1] C. Anfinsen, Biochem. J. **128**, 737-749 (1972).

[2] C. Anfinsen, Science **181**, 223-230 (1973).

[3] L. A. Mirny and E. I. Shakhnovich, J. Mol. Biol. **264**, 1164-1179 (1996).

[4] M. Y. Shen and A. Sali, Protein Sci. **15**, 2507-2524 (2006).

[5] S. J. Sammut, R. D. Finn and A. Bateman, Brief. Bioinform. **9**, 210-219 (2008).

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Giggiland, T. N. Bath, H, Weissig, I. N. Shindyalov, P. E. Bourne, Nucleic Acids Res. **28 (1)**, 235-242 (2000).

[7] S. Parkin, B. Rupp, H. Hope, Acta Crystallogr. **52**, 18-29 (1996).

[8] U. G. Wagner, N. Muller, W. Schmitzberger, H. Falk, C. Kratky, J. Mol. Biol **247**, 326-337 (1995).

[9] T. R. Hynes, R. O. Fox, Proteins **10**, 92-105 (1991).

[10] M. Leone, P. Di Lello, O. Ohlenschlager, E. M. Pedone, S. Bartolucci, M. Rossi, B. Di Blasio, C. Pedone, M. Saviano, C. Isernia, R. Fattorusso, Biochemistry **43**, 6043-6058 (2004).

[11] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sanders, R. Zecchina, J. N. Onunic, T. Hwa and M. Weigt, Proc. Natl. Acad. Sci. U.S.A. **108**, E1293 (2011).

[12] S. Lui and G. Tiana, J. Chem. Phys. **139**, 155103 (2013).

[13] A. Contini and G. Tiana, J. Chem. Phys. **143**, 025103 (2015).

[14] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman and R. D. Finn, Nucleic Acids Res. **40**, D290-D301, (2012).

[15] S. R. Eddy, Bioinformatics Rev. **14**, 755-763 (1998).

[16] The UniProt Consortium, Nucleic Acids Res. **43**, D204-D212, (2015).

[17] E. T. Jaynes, Phys. Rev. **106**, 620-630 (1957).

[18] A. Contini, *Analysis of frustration in proteins from coevolutionary data*, MSc thesis (Università degli Studi di Milano, 2015), `http://www.mi.infn.it/~tiana/Biophysics/People.html`.

[19] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **106**, E1293 (2009).

[20] D. T. Jones, J. Mol. Biol. **292**, 195-202 (1999).

[21] D. W. A. Buchan, F. Minneci, T.C.O. Nugent,K. Bryson, D. T. Jones, Nucleic Acids Res. **41 (W1)**, W340-W348 (2013).

[22] W. Xu, S. C. Harrison, M. J. Eck, Nature **385**, 595-602 (1997).

[23] M. J. Abraham, D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team, *GROMACS User Manual version 5.1.2* (2016), `www.gromacs.org`

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

[25] B. Efron, *Bootstrap methods: another look at the jackknife* (Springer, 1992).

[26] I. Jolliffe, *Principal Component Analysis* (Wiley StatsRef: Statistics Reference Online, 2014).

[27] C. J. Roberts, Biotechnol. Bioeng. **98**, 927 (2007).

[28] M. Stefani and C. M. Dobson, J. Mol. Med. **81**, 678-699 (2003).

[29] C. Hamiaux, J. Perez, T. Prangè, S. Veesler, M. Riès-Kautt and P. Vachette, J. Mol. Biol **297**, 697-712 (2000).

[30] M. Gottschalk, K. Venu and B. Halle, Biophys. J. **84**, 3941-3958 (2003).

[31] A. Bernini, O. Spiga, A. Ciutti, V. Venditti, F. Prischi, M. Governatori, L. Bracci, B. Lelli, S. Pileri, M. Botta, A. Barge, F. Laschi, N. Niccolai, Biochim. Biophys. Acta **1764**, 856-862 (2006).

[32] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg, J. Comput. Chem. **13**, 1011-1021 (1992).

[33] J. K. Noel, P. C. Whitford, K. Y. Sanbonmatsu and J. N. Onuchic, Nucleic Acids. Res. **38**, W657-61 (2010).

[34] G. Tiana, F. Villa, Y. Zhan, R. Capelli, C. Paissoni, P. Sormanni, E. Heard, L. Giorgetti and R. Meloni, Comp. Phys. Comm. **186**, 93-104 (2015).

Ringraziamenti

Giunti alla fine è tempo di ringraziamenti.

Ringrazio innanzitutto la mia famiglia: mamma, per avermi sempre ascoltato e sostenuto in questi anni; papà, per aver sempre creduto in me; Sara, per avermi aiutato in tutte le decisioni importanti, ascoltandomi e capendomi sempre.

Ringrazio poi Guido, per avermi accolto nel suo gruppo stimolandomi sempre con ottimismo, e per rendere il secondo piano un ambiente così piacevole.
Grazie a Cape, che con la sua infinita disponibilità mi ha pazientemente guidato in questi mesi, sopportando paranoie a non finire.
Grazie a Sara, Bob e a tutto il gruppo di biofisica del secondo piano, per tutto il tempo passato insieme e per i kaffeeeee che hanno contribuito a rendere positiva questa esperienza.

Grazie ai miei compagni e amici dell'università: Silvano, Poli, Panzeri, Maci, Campa, Stefania, Luca Vis, Antonio, Miglietta, Matteo e tutti coloro che hanno condiviso una parte di questo percorso con me. Grazie a tutte le giornate e le serate passate insieme a voi questi sono stati anni bellissimi.

Grazie ai miei amici di sempre: Ruggi, Sam, Bea, Adri, Angus, Peolo, Guidano, Kikko, Eldiano, Sara, Bala, per aver sempre ravvivato la mia permanenza in valle.