Universidad de La Laguna

SCIENCES FACULTY

Physics Section

FINAL DEGREE PROJECT

Theoretical study of protein folding

Milva Beatriz Marrero Castro

Project supervisor:

Javier Hernández Rojas

7th July 2021

Me gustaría expresar mi gratitud a mi tutor, Javier Hernández Rojas, por aceptarme y guiarme en este proyecto.
A mis padres, mi tía y mi abuela, por creer en mí, por su apoyo incondicional y por estar siempre a mi lado.
A mis amigos, los de la infancia y los que han llegado a mi vida durante la carrera, por hacer de estos años una experiencia increíble y por ayudarme siempre que lo he necesitado.
A Jose, por convertir estos últimos años de carrera en una etapa más bonita, así como por su gran apoyo y ayuda durante todo este tiempo.

Dedicado a mis abuelos, que, aunque no pudieron acompañarme hasta el final de esta etapa, espero que estén orgullosos de mí y de todo lo que he conseguido.

Table of contents

List of figures

1. Summary

El presente trabajo de fin de grado consiste en el estudio teórico del proceso de plegamiento de las proteínas. Para ello, partimos de una proteína simple de 30 aminoácidos en su estado nativo (plegado), para la cual están perfectamente establecidas las posiciones de cada monómero y las posibles interacciones entre los mismos. Tras haber entendido cómo funciona el proceso de plegamiento se ha realizado un pequeño estudio de dicha proteína, el cual ha consistido en implementar un modelo de grano grueso de interacción entre los monómeros, propuesto por Hernández-Rojas y Gómez Llorente [1]. También hemos obtenido los mapas de contactos correspondientes a diferentes distancias críticas.

A continuación, se ha analizado la transición de fase del estado nativo (plegado) al desplegado a través del estudio de la energía libre. Para ello, se han utilizado 10000 configuraciones de la proteína para cada una de las tres temperaturas de estudio ($k_BT = 13$, $k_BT = 13.5$ y $k_BT = 14$). Cada uno de estos tres conjuntos se obtuvo mediante el método Parallel Tempering Monte Carlo (PTMC) por Hernández-Rojas y Gómez Llorente [1][2]. El estudio de la energía libre se ha llevado a cabo de dos formas distintas:

1. Tomando el número de contactos (Q) como coordenada de reacción.
2. Tomando la raíz de la desviación cuadrática media (RMSD) como coordenada de reacción.

Sin embargo, el procedimiento a seguir en ambos casos ha sido similar. En primer lugar, para Q se ha obtenido el número de contactos en cada una de las configuraciones, comparando las distancias entre sus monómeros con los de la proteína modelo. Una vez obtenido el número de contactos, se han calculado tanto P(Q) como la energía libre (F) para cada una de las temperaturas estudiadas, pudiendo observar claramente los distintos estados en los que podemos encontrar a la proteína. Este procedimiento se ha repetido para distintas distancias críticas, pudiendo observar cómo varían los parámetros estudiados con respecto a esta.

Finalmente, hemos tomado la RMSD como coordenada de reacción. Dicho valor se ha obtenido para cada una de las configuraciones de cada conjunto, a partir de lo cual se ha calculado también la energía libre en función de esta nueva coordenada de reacción. Así, hemos podido observar de nuevo dos estados bien diferenciados.

Con todo esto, se ha podido comprobar a qué temperatura se produce la transición de fase del estado plegado al desplegado.

2. Introduction

Abstract

Las proteínas son moléculas, generalmente de gran tamaño (macromoléculas), que químicamente pueden caracterizarse como polímeros lineales compuestos por una secuencia de aminoácidos (monómeros). La secuencia de aminoácidos determina el estado funcional o nativo de cada proteína, observándose cuatro niveles en la organización de su estructura hasta adquirir el estado nativo tridimensional. El proceso mediante el cual se adquiere dicho estado nativo se denomina plegamiento. El hecho de que este proceso sea espontáneo, entre otras cosas, ha dado lugar a numerosas investigaciones, las cuales siguen enfrentándose hoy en día al "problema del plegamiento". Este se basa principalmente en tres cuestiones: ¿Cuál es el código por el cual una secuencia de aminoácidos dicta la estructura nativa de proteína?, ¿cómo se pliegan tan rápido?, ¿es posible predecir la estructura de las proteínas a partir de sus secuencias?

It is well known that proteins underlie all kinds of metabolic processes and constitute a basic structural component of living organisms, making them essential for life. The protein's structure depends on the amino acids sequence and conditions its functionality. Thus, it is foreseeable that an error in its structure could lead to serious biological or physiological problems. This is the case for diseases like Alzheimer or Parkinson. Taking this into account, it is understandable that such different parts of the scientific community focus on studying these molecules in depth. In fact, the relation between the sequence of amino acids and the structure of the molecule as well as the properties of the process by which the sequence acquires its functional structure is the object of study in different fields such as medicine, chemistry, biology, and physics.

2.1. What is a protein?

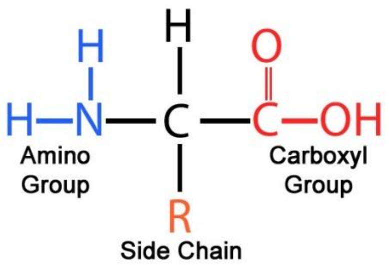


Figure 1: Structure of an amino acid [4].

Proteins are macromolecules formed by amino acids. In a coarse-grained description, one can consider proteins as heteropolymers for which the monomers are amino acids [3]. An amino acid is an organic molecule that contains an amino ($-NH_2$) and a carboxyl ($-COOH$) functional group. There are 20 kinds of amino acids in nature that can form proteins. Their main characteristic is that the amino and the carboxyl groups are connected to the same carbon atom, called α carbon. Also, there are a hydrogen atom and an additional group, called side chain (R), connected to this carbon (Figure 1). This side chain determines the properties of each amino acid. The amino acids that make up a protein are linked by a peptide bond (between the carboxyl group of one amino acid and the amino group of another) which has a partial double bond character. This prevents the rotation of the groups attached to it, making them remain in the same plane. Due to the characteristics of the peptide bond and of the side chains of the amino acids that make up the protein, it presents specific structural properties.

2.2. The structure of a protein and the folding process

Proteins are linear unbranched polymers with a specific length and amino acids sequence. The functionality of a protein depends on this structure. The amino acids sequence determines the protein's ability to assume its functional structure or native state. There are four levels in the structure of a protein:

- Primary structure: the amino acids sequence as it is mentioned in section 2.1. This structure is held together by peptide bonds of which geometry gives the protein some structural restrictions (Figure 2.a).
- Secondary structure: These restrictions mentioned above lead to local sub-structures on the polypeptide backbone chain. This is, the amino acids tend to order locally, giving rise to two types of structures: the α-helix and the β-strand or β-sheets. These are known as the secondary structure elements (Figure 2.b).
- Tertiary structure: Some pairs of amino acids have small side chains which do not give rise to large structural restrictions. This allows the protein to compact and makes the elements of the secondary structure interact with each other, leading to three-dimensional structures called the tertiary structure (Figure 2.c).
- Quaternary structure: Sometimes the protein is formed by the aggregation of various individual amino acids chains that operate as single functional units. They have a specific tertiary structure and interact with each other, producing the quaternary structure (Figure 2.d).

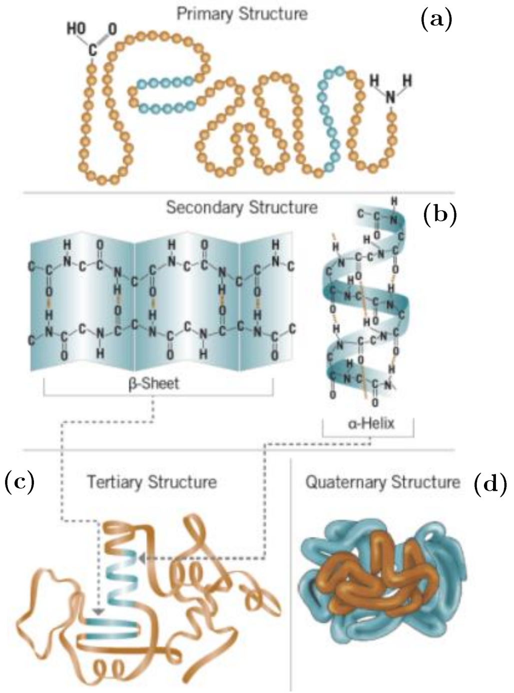


Figure 2: Levels of the protein folding: (a) primary structure, (b) secondary structure, where we can see α-helix and β-sheet, (c) tertiary structure, (d) quaternary structure. [5]

The functionality of a protein depends on whether the amino acids sequence acquires the native state or not. The process in which this happens is called folding. The proteins' ability to acquire their functional three-dimensional structure or native state not only takes place in living organisms but also *in vitro* [6]. It is generally accepted that the folding process occurs spontaneously, and the reversibility of the *in vitro* unfolding-folding process has reinforced this idea. Also, it is known that protein folding is governed by the amino acids sequence [7]. Nevertheless, there is still a large list of unanswered questions about protein folding.

2.3. State of the art

The spontaneous folding phenomenon was discovered by Anfinsen's group in 1961 [8] awakening a big interest in the scientific community. In this context, in 1968, Cyrus Levinthal noted that proteins have an astronomical number of possible conformations [9]. If a protein were to attain its correctly folded configuration by sequentially sampling all the possible conformations, it would require a time longer than the age of the universe to arrive at its correct native conformation. This is true even if conformations are sampled at rapid (nanosecond or picosecond) rates. However, most small proteins fold spontaneously on a millisecond or even microsecond time scale. Levinthal's paradox is one of the three questions that constitute the protein folding problem [10]:

- What is the physical code by which an amino acid sequence dictates a protein's native structure?
- How can proteins fold so fast?
- Can we devise a computer algorithm to predict protein structures from their sequences?

Even though they remain unanswered, many advances have been made in the protein folding problem, leaving us closer to the solution.

Regarding the first question, some factors seem to contribute: (i) Hydrogen bonds. Protein structures are composed of α-helices and β-sheets, which maintain their shape by the formation of hydrogen bonds. (ii) van der Waals interactions. The atoms within a folded protein are tightly packed, implying the importance of the same types of close-ranged interactions that govern the structures of liquids and solids [11]. (iii) Backbone angle preferences. Like other types of polymers, protein molecules have preferred angles of neighbouring backbone bond orientations. (iv) Electrostatic interactions. Some amino acids attract or repel because of negative and positive charges. (v) Hydrophobic interactions. Proteins ball up into well-packed folded states in which the hydrophobic (H) amino acids are predominantly located in the protein's core and the polar (P) amino acids are more commonly on the folded protein's surface. (vi) Chain entropy. Opposing the folding process is a large loss in chain entropy as the protein collapses into its compact native state from its many open denatured configurations [10]. These physical forces are described approximately by "force fields", which are models of potential energies that can be used in computer simulations. So far, such modelling succeeds only on a limited set of small simple protein folds [12] and it does not yet accurately predict protein stabilities or thermodynamic properties.

Concerning Levinthal's question, studies of the chain entropies in models of foldable polymers showed that more compact, low-energy conformational ensembles have fewer conformations [13], indicating that protein-folding energy landscapes are funnel-shaped (Figure 3), but we still need a “folding mechanism”. We do not yet have a general principle that explains the differences and similarities of the folding routes and rates of different proteins in advance of the data. Nevertheless, there are a few general conclusions. Proteins appear to fold in units of secondary structures [14]. A protein’s stability increases with its growing partial structure as it folds. And a protein appears to first develop local structures in the chain followed by growth into more global structures [15]. Funnelled landscapes predict that the different individual molecules of the same protein sequence may each follow microscopically different routes to the same native structure. Some paths will be more populated than others [10].

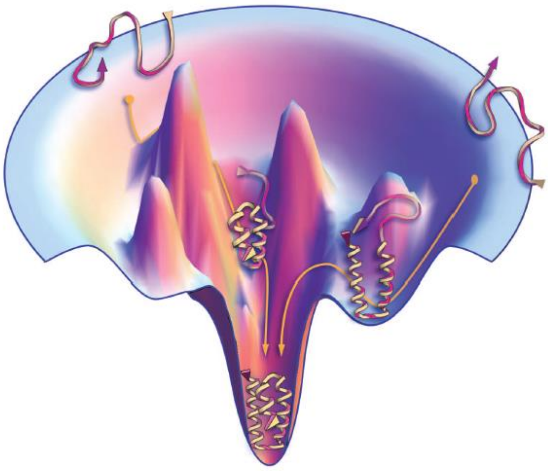


Figure 3: Funnel-shaped energy landscape of a protein [10].

Currently, all successful structure-prediction algorithms are based on assuming that similar sequences lead to similar structures. If the target protein’s sequence is related to a sequence that is already in the Protein Data Bank (PDB), predicting its structure is usually easy. In such cases, target protein structures are predicted by using “template-based modelling”. But when there is no protein in the PDB with a sequence resembling the target’s, accurately predicting the structure of the target is much more difficult. These latter predictions are called “free modelling”. Many prediction methods are hybrids, combining template-based modelling, fragment assembly, and other strategies. However, it remains a challenge to predict many protein structures accurately [10].

3. Objectives

Abstract

Los principales objetivos de este trabajo son entender el problema del plegamiento, así como los métodos de simulación existentes para abordarlo. Además, se ha estudiado la energía libre en la transición de fase de estado plegado a desplegado.

There is still a long way to go in the study of protein folding, which is the reason that has led us to carry out this work. Thus, the main objectives are to understand the folding problem as well as the existing simulation methods to address it. In addition, we studied the behaviour of the free energy of a protein in the phase transition from folded to unfolded state.

To do this, we used the data obtained by Javier Hernández-Rojas and Gómez Llorente [1]. These data consisted of the configuration of a simple protein (30 amino acids) in its native state and numerous configurations of that protein at different temperatures, obtained using a method called Parallel Tempering Monte Carlo [2].

4. Theoretical background

Abstract

Hemos visto en apartados anteriores que uno de los modelos más aceptados para entender el proceso de plegamiento es el del paisaje de energía, en el cual el estado nativo se ubica en el fondo de un embudo. Simular cómo una proteína consigue alcanzar dicho estado de mínima energía libre puede abordarse de múltiples formas. Una de las aproximaciones más tradicionales consiste en analizar la cadena proteica con todos los átomos que la componen (*all-atom models*). Este tipo de análisis implica un enorme esfuerzo computacional y solo puede ser asumido para proteínas de pequeño tamaño. Como solución se plantean los modelos de grano grueso (*coarse grained models*), en los que cada aminoácido se reduce a una estructura elemental que ocupa la posición de su C_α. En general, la simulación del proceso de plegamiento con este modelo requiere de métodos en los cuales los distintos aminoácidos de la cadena van ocupando unas determinadas posiciones hasta que se adquiere el estado nativo. En este sentido los modelos más sencillos son los de *lattice*, en los cuales los distintos aminoácidos solo pueden ocupar los nodos de una red cúbica (en general) y para ello solo pueden ejecutar determinados movimientos. Aunque estos modelos han permitido grandes avances, su simplicidad implica que las estructuras de menor energía obtenidas presentan escaso parecido con el modelo nativo. Para solventar este problema surgen los modelos *off-lattice,* en los que cada monómero de la cadena se mueve libremente dentro de un determinado rango de acción.

The simulation of the protein folding process requires a certain level of representation of the protein structure. The most traditional approximation would be to extend this structural representation to all the atoms of each amino acid that form the chain. These all-atom models (AA) present large detail and can be approached with some solvency for small proteins. However, in practice protein molecules are generally large and the simulation based on AA requires enormous computational effort [15]. To avoid this problem, it is necessary a simplification of the protein geometry and coarse-grained models (CG) represent an important advance in this way. In a CG, the different atoms that form each amino acid are grouped into a single element, generally located in the position that would be occupied by the C_α of the amino acid [17].

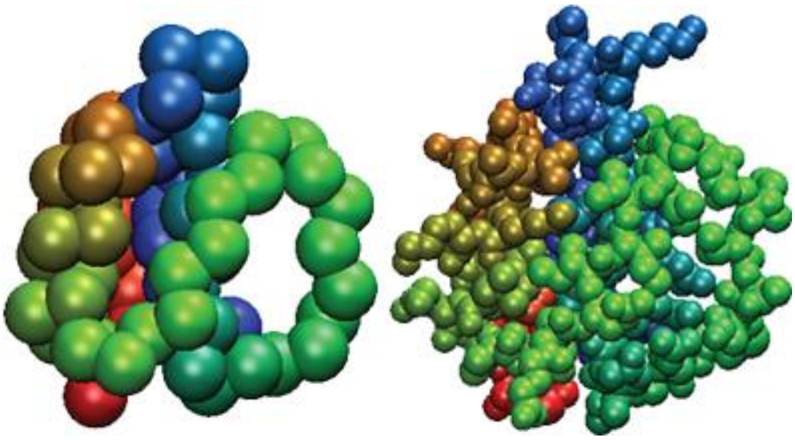


Figure 4: Coarse grained (right) and all-atom (left) representation of a protein. Obtained from [15].

Even considering the loss of detail that supposes opting for a CG, they maintain a series of advantages over AAs. Despite the rapid development of simulation technology, AAs still are limited to small systems and simple processes. However, CGs are more computationally effective and allow much longer simulations for much greater systems.

Protein folding simulation generally assumes a progressive advance towards the native structure. This is, the chain folds in successive steps, in which the different monomers occupy certain positions in space. Within the CGs, simulating this process can be approached in different ways. One of them are the lattice models, where the three-dimensional space is discretized, and the monomers are constrained to lie on the lattice nodes. Only one monomer is allowed per node, so the chain is a self-avoiding path on the lattice [18]. For simulating the folding process, moves along the lattice are restricted and chosen from a set of three possible types: a one-monomer corner flip, a one-monomer end pivot, and a two-monomer crankshaft move (figure 5).

However, lattice models present a significant drawback: the representation of the protein structure is oversimplified in a way that can hardly be considered as a representation of real molecules. In consequence, the native structures cannot be easily represented. Off-lattice models are a possible solution, since the three-dimensional space is continuous, and the monomer location is not restricted to specific positions.

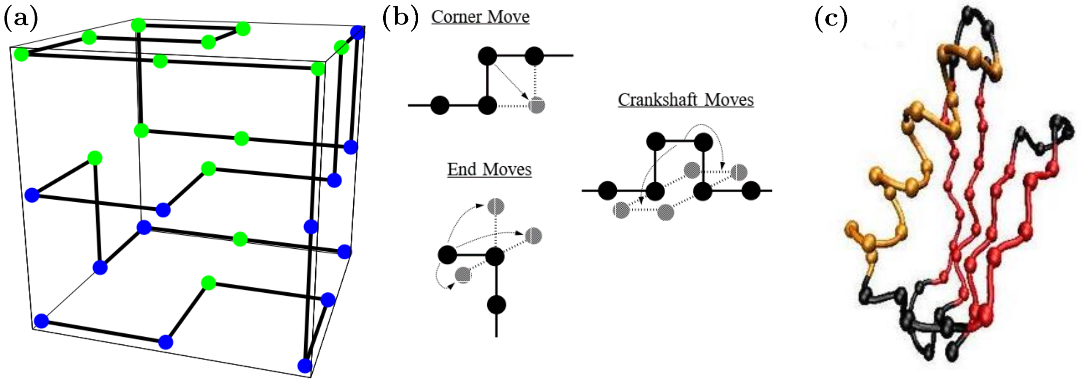


Figure 5. (a) Representation of a 27-length polymer chain on a cubic lattice. The conformation is a maximally compact cube [18]. (b) Moves used in the lattice models where grey circles represent the possible lattice points a given monomer can move to provided that that point is not occupied [18]. (c) Representation of a protein conformation using an off-lattice model [19].

In these models, we usually start from the fully unfolded protein chain, for which the monomers will be located in any position within an assignable region. The movement of each monomer within this space is determined by variables inherent to the peptide bond itself, for example the dihedral angles or non-local contact potentials [1][3].

In 1998 Clementi, Maritan and Banavar [3] proposed an off-lattice model based on contact potential. Considering an heteropolymer formed by 4 different types of amino acids, the potential between two monomers is defined as follows:

$$U_{ij} = \delta_{i,j+1} f(r_{i,j}) + \eta_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \qquad (1)$$

Where the first term corresponds to the interaction between bonding monomers (anharmonic repulsive potential) and the second one corresponds to the interaction between the non-bonding monomers (Lennard-Jones potential). For the anharmonic potential, the function $f(r_{i,j})$ is:

$$f(r_{i,j}) = a(r_{ij} - r_0)^2 + b(r_{ij} - r_0)^4 \qquad (2)$$

Where a and b are 1 and 100 respectively, r_0 is the equilibrium distance (3.8 Å) and r_{ij} is the inter-residue distance. In this model, Clementi et al. [3] used five predefined σ_{ij} values, which are: 6.25, 6.5, 7, 7.5 and 8 Å. The values of η_{ij} are: $\eta_{C,C} = 40, \eta_{\mathrm{C,O}} = \eta_{\mathrm{O,C}} = 30, \eta_{\mathrm{C,N}} = \eta_{\mathrm{N,C}} = 20, \eta_{\mathrm{C,B}} = \eta_{\mathrm{B,C}} = 17, \eta_{\mathrm{O,O}} = 25, \eta_{\mathrm{O,N}} = \eta_{\mathrm{N,O}} = 13, \eta_{\mathrm{O,B}} = \eta_{\mathrm{B,O}} = 10, \eta_{\mathrm{N,N}} = 5, \eta_{\mathrm{N,B}} = \eta_{\mathrm{B,N}} = 2$ and $\eta_{\mathrm{B,B}} = 1$.

In 2008, Hernández-Rojas & Gómez Llorente [1] published a study that uses a similar off-lattice model, although with some slight modifications so that the energy of the global minimum was as deep as possible. Thus, the potential energy is:

$$U_{ij} = \delta_{i,j+1} a(r_{ij} - r_0)^2 + (1 - \delta_{i,j+1}) 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \qquad (3)$$

Here the Lennard-Jones potential is applied only to non-covalent contacts. Moreover, in the $f(r_{ij})$ expression, the term raised to fourth power is eliminated and $a = 50$ Å^{-2}. As in the Clementi et al. model, $r_0 = 3.8$ Å. Finally, σ_{ij} values were chosen so that the native structure corresponding to the chosen sequence is significantly stabilized: $5<\sigma_{ij}<17$ Å. These values are provided in table 2 of appendix A [2]. The values of ϵ_{ij} used by Hernández-Rojas & Gómez Llorente [1] are equal to the η_{ij} values used by Clementi et al. [3] divided by four. Thus, 10 different ϵ_{ij} parameters were used: $\epsilon_{\mathrm{C,C}} = 10, \epsilon_{\mathrm{C,O}} = \epsilon_{\mathrm{O,C}} = 7.5, \epsilon_{\mathrm{C,N}} = \epsilon_{\mathrm{N,C}} = 5, \epsilon_{\mathrm{C,B}} = \epsilon_{\mathrm{B,C}} = 4.25, \epsilon_{\mathrm{O,O}} = 6.25, \epsilon_{\mathrm{O,N}} = \epsilon_{\mathrm{N,O}} = 3.25, \epsilon_{\mathrm{O,B}} = \epsilon_{\mathrm{B,O}} = 2.5, \epsilon_{\mathrm{N,N}} = 1.25, \epsilon_{\mathrm{N,B}} = \epsilon_{\mathrm{B,N}} = 0.5$ and $\epsilon_{\mathrm{B,B}} = 0.25$

Various authors [20] consider that to unravel the mysteries about the folding process it is necessary to know in depth the process of phase transition. This is, the tiny segment of a single molecule trajectory when the free energy barrier between states is crossed and for protein folding contains all of the information about the self-assembly mechanism. There are many questions about it, for example: is it just a random search process for the minimum energy conformation or are there forces that determine paths for the folding? Considering Levinthal's paradox, the first case is rejected. Therefore, it seems coherent to propose the existence of paths according to which the protein passes from a denatured state to the native one through a series of defined steps [9]. On the other hand, although it is very brief, the folding process requires a certain period of time, so it does not seem possible to think of a barrier-free mechanism for the acquisition of the native state structure. Thus, in late 1980 a new vision of the issue arose: the energy landscape theory [21] (already mentioned in section 2.3.). It suggests that the most realistic model of a protein is a minimally frustrated heteropolymer with a rugged funnel-like landscape biased towards the native structure. The funnel's bottom constitutes the minimum energy that corresponds to the native state and the width is related to the entropy of the system. Taking this into account, one way to tackle the folding problem might be to find the lowest energy structure for a given amino acid sequence.

The energy function that governs the folding process is unknown. However, many authors have approached this issue through reductions [3], including assimilating the energy of the native state to the Lennard-Jones potential for the pairwise interaction of the different monomers that constitute the protein. Of course, like so many others, the Lennard-Jones potential is simply an approximation, since the actual process is much more complex and includes different potentials (Coulomb interactions, hydrogen bonds, etc.). In our case, we start from a protein in its native state constituted by 30 monomers, considering 4 types of amino acids (called N, C, B, O)[2]. Table 1 of appendix A contains the sequence of amino acids as well as the position of each one in the three-dimensional space [2].

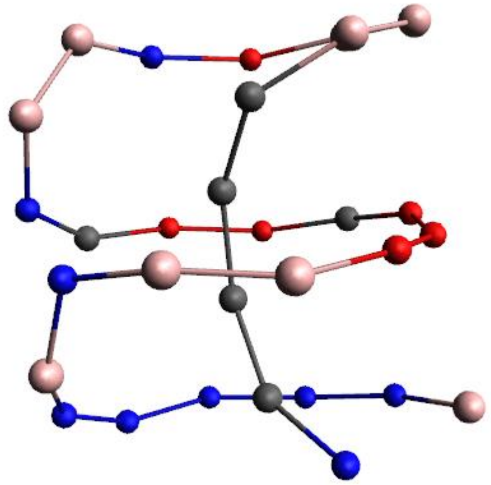


Figure 6: Graphic representation of the protein studied in its native state. Image represented by the program Avogadro.

Using equation (3) we obtain the interaction between each pair of monomers. Thus, the total potential energy is given by:

$$U_{TOTAL} = \sum_{i>j} U_{ij} \qquad (4)$$

Taking this into account, the value obtained for the potential energy of the protein in its native structure is:

$$U_{min} = -1314.69299095$$

This result corresponds to the minimum energy and, hence, to the bottom of the funneled landscape. Also, this result is the same as the one obtained by Hernández-Rojas & Gómez Llorente by using global optimization by basin-hopping [1][2]. By using global optimization, one can obtain the lowest energy configuration of an atomic or molecular system [22].

On the other hand, a minimalistic representation of a protein's structure is given by its contact map. This kind of representations are quite useful and can serve different purposes, like predicting the three-dimensional structure of the protein. Another interesting use is to perform a search in the space of possible contact maps, for a fixed sequence of amino acids, to find maps of low energy. This is, to find contact maps that correspond to highly probable structures [23]. Nevertheless, we must not mislead a contact map with a distance map. In a contact map, a minimal amount of information is available: for a pair of amino acids, we only know whether they are in contact or not. A distance matrix presents real-valued distances between pairs of amino acids. Therefore, it is considerably harder to reconstruct a structure from a contact map than from a distance matrix [23].

For a protein of N residues, the contact map is an N × N matrix **S**, whose elements are given (for all i,j=1, 2,..., N) by [24]:

$$S_{ij} = \begin{cases} 1 & if\ amino\ acids\ i\ and\ j\ are\ in\ contact \\ 0 & otherwhise \end{cases}$$

It is considered that two amino acids are in contact if the distance between them is smaller than a fixed critical distance. Thus, the contact map presents more of these contacts as the critical distance increases. As the result of representing the elements of that matrix **S**, we obtain a 2D representation of the protein's structure. With this kind of representation, secondary structures are easily detected. α-Helices appear as thick bands along the main diagonal, since they involve contacts between one amino acid and its successors. The characteristics of parallel or anti-parallel β-sheets are thin bands, parallel or anti-parallel to the main diagonal. On the other hand, the overall tertiary structure is not easily discerned [23]. In our case, we have analysed the variation of the contact map of the protein in its native state for four different critical distances:

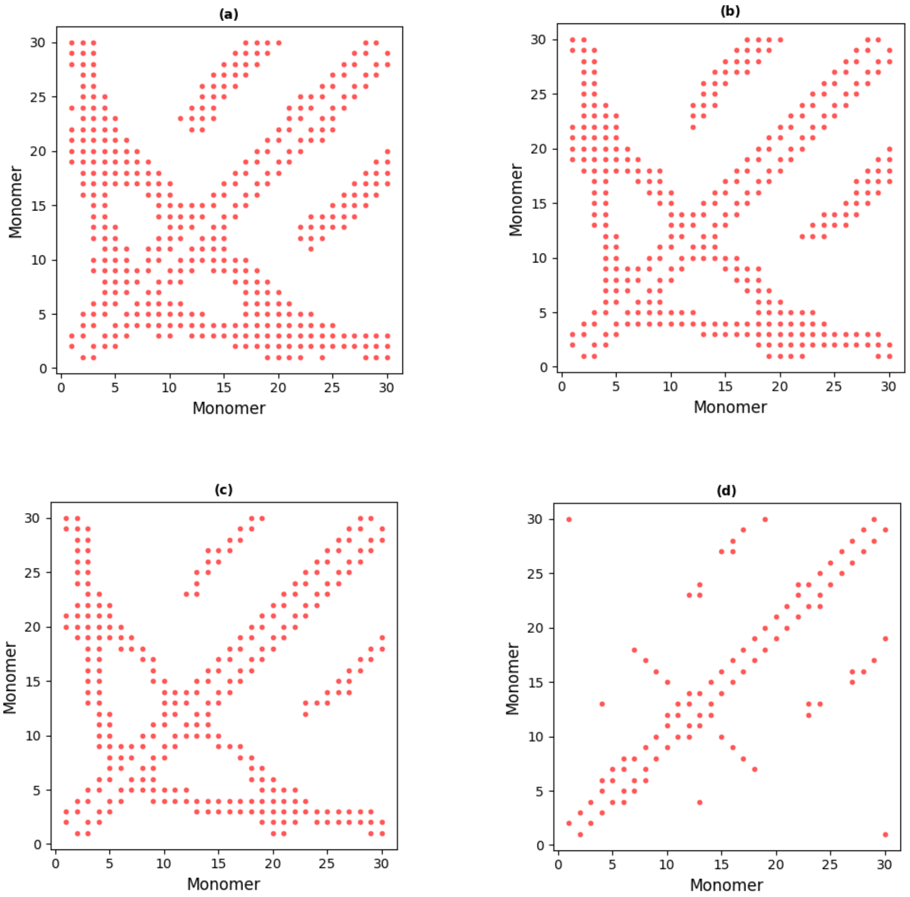


Figure 7: Changes on the contact map for the protein analysed by varying the critical distance. Each spot represents a contact. In (a) the critical distance is 10 Å; in (b), it is 9 Å; in (c), it is 8 Å and in (d) it is 7 Å.

Thus, based on the results presented in figure 7, we can conclude that the protein analysed in the present study has a helix, which coincides with the reality (figure 6). However, this helix has a completed turn in 10-12 beads, whereas natural proteins helices have 3.6 residues per turn [3]. We also observe the number of contacts' dependence on the critical distances: the lower the critical distance, the lower the number of contacts.

In this case, we have obtained the contact map corresponding to the protein studied from its three-dimensional structure. In this way, we have not only been able to verify that, indeed, both representations are equivalent, but we have also obtained the number of contacts in each case. This value will be very useful later.

5. Parallel Tempering Monte Carlo (PTMC)

Abstract

Para el estudio de la transición del estado plegado al desplegado en la proteína se han empleado numerosas simulaciones de esta a distintas temperaturas. Dichas simulaciones fueron obtenidas por Hernández-Rojas and Gómez Llorente [1] mediante Parallel Tempering Monte Carlo (PTMC). Este método consiste en trabajar simultáneamente con M réplicas de la proteína original a diferentes temperaturas T_i bien definidas. Por una parte, se perturba un monómero cualquiera de la réplica de forma que cambie su posición de forma aleatoria. Si la energía de la nueva configuración es menor que la de la anterior, se acepta el cambio. De lo contrario, se elige un número aleatorio entre 0 y 1, comparándolo con el factor de Boltzmann y si el número elegido es menor, se acepta el cambio. Por otra parte, también se propone cambiar una réplica por otra que esté a una temperatura vecina cada cierto número de pasos. El procedimiento a seguir para saber si este cambio es aceptado o no, es análogo al descrito para la perturbación de monómeros.

For the study of the transition from folded to unfolded state of a protein numerous configurations of the protein described in the previous section have been used. These configurations were obtained by Javier Hernández Rojas and Gómez Llorente [1] using Parallel Tempering Monte Carlo simulations. This method consists of working simultaneously with M replicas, each in the canonical ensemble, and at different temperatures, T_i [25]. In this case, 50 replicas were used, each one of them at a different and well-defined temperature. As we will comment below, Parallel Tempering performs two types of movements in the extended ensemble [26].

Movements within the replicas: The model used by Hernández-Rojas and Gómez Llorente in 2008 [1] consists of an off-lattice model, very similar to the one used by Clementi et al. in 1999 [3]. Thus, the positions of each amino acid are not affixed to the nodes of a cubic lattice (lattice models), so the results are more realistic. The simplest way of tackling an off-lattice model does not present significant variations to an on-lattice model and can be summarized in the following steps:

1. Choosing the initial state. In this case, it is the protein mentioned in section 4 in its native state.
2. Perturbing a random monomer from the chain, for which the position will vary (at random too) to an alternative one. This perturbation cannot be very great, because the temperature would be infinite; nor very small, since the change would be almost imperceptible. The perturbation parameter can be adjusted in the simulation so that 50% of the configurations generated are valid.
3. Calculating the energy of the new configuration and its difference with the energy of the previous configuration, this is:

$$\Delta E = E_{new} - E_{initial} \qquad (5)$$

If the energy of the new configuration is lower than the previous energy ($\Delta E < 0$), the change is accepted and continued with the perturbation of another amino acid. On the other hand, if $\Delta E > 0$, a random number between 0 and 1 is chosen and compared with the Boltzmann factor:

$$e^{-\left(\frac{\Delta E}{k_B T}\right)}$$

Where k_B is the Boltzmann constant and T, the temperature. If the chosen number is lower than the Boltzmann factor, the movement is accepted. Otherwise, it is rejected.

Movements between replicas: They consist of proposing the exchange of the complete chain of a certain replica for another one of the replicas studied. This exchange is done every certain number of steps and between neighbouring temperatures. If we denote the energy difference between the two replicas for which we propose the exchange as:

$$\Delta E_{ij} = E_i - E_j \qquad (6)$$

Where E_i is the energy of the replica at temperature T_i and E_j that of the replica that is at temperature T_j. The probability of accepting the exchange is [27]:

$$P = \begin{cases} e^{-(\beta_j - \beta_i)\cdot \Delta E_{ij}} & if\ \Delta E > 0 \\ 1 & otherwhise \end{cases} \qquad (7)$$

Where β_i is $\frac{1}{k_B T_i}$, being k_B the Bolztmann constant.

If $\Delta E > 0$, one must decide if the change is made with a certain probability. In this case, a random value between 0 and 1 is chosen and if it is lower than the calculated probability, the exchange is accepted.

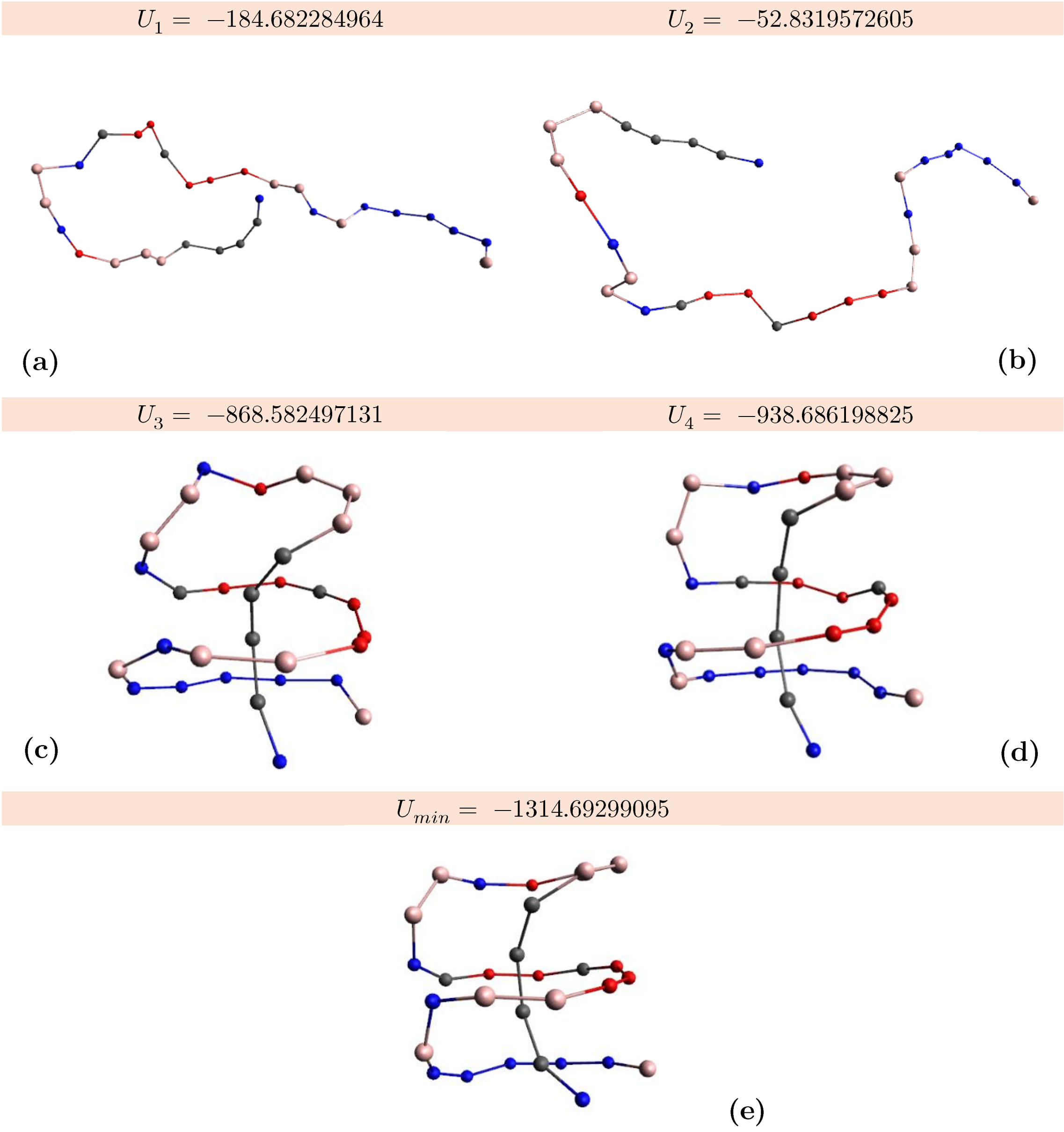
Bearing all the above in mind, there is one main question about the application of this process: How often should we do the exchange of replicas?

If the replicas are exchanged too frequently, they will not have the time necessary to evolve and reaching a state that is sufficiently different from the one they had when attempting the previous exchange. On the other hand, waiting too much to propose the exchange means wasting the benefits of the method and unnecessarily wasting processing time.

The main benefit of this method is that, as it is constantly switching between systems at different temperatures, it helps to extract simulations carried out at low temperatures from possible metastable states (local minimum in the free energy funnel). By bringing these configurations to higher temperatures, they can escape these states more easily. In addition, with this method, one can obtain well-defined temperatures allowing us to calculate the heat capacity and to estimate the transition temperature.

As we mentioned before, as a result of applying the aforementioned, we have been provided by Hernández-Rojas & Gómez Llorente with the configurations obtained in their work [1][2] for the temperatures at which the phase transition occurs. These temperatures are $k_B T = 13$, $k_B T = 13.5$ and $k_B T = 14$ (in reduced units). The PTMC was performed on 50 replicas of the protein in its native state, each of them at a different temperature. Finally, 10000 different configurations for each temperature were obtained. Each configuration is characterised by the position x, y, z of the monomers that constitute it.

For greater clarity, the graphical representation of some of the configurations obtained by PTMC is provided below, as well as the representation of the model protein (figure 6):


$U_1 = -184.682284964$ $U_2 = -52.8319572605$

$U_3 = -868.582497131$ $U_4 = -938.686198825$

$U_{min} = -1314.69299095$

Figure 8: Graphic representation of some of the configurations obtained by PTMC and of the model protein. In figures (a) and (b) we observe the protein unfolded at $k_B\mathrm{T} = 14$ and $k_B\mathrm{T} = 13.5$ respectively. In figures (c) and (d) we observe the folded structures of the protein (at $k_B\mathrm{T} = 13.5$ and $k_B\mathrm{T} = 13$ respectively). Finally, the protein its native state is represented in (e). The value of U for each configuration denotes its potential energy, obtained with equation (4). Images represented by the program Avogadro.

In images 8a and 8b we observe the protein completely unfolded. Also, in figures 8c and 8d, the configurations obtained for the folded protein are quite similar to the protein in its native state. Another interesting fact is that, as we can see, the potential energy for the unfolded configurations of the protein (U_1 and U_2) is much higher than for the folded structures (U_3, U_4 and U_{min}), which coincides with the theory. This proves the proper functioning of the method used to obtain the configurations.

6. Results and discussion

Abstract

El análisis de la energía libre en la transición de fase se ha realizado por dos métodos distintos: (1) Usando el número de contactos (Q) como coordenada de reacción y (2) usando la raíz de la desviación cuadrática media (RMSD) como coordenada de reacción. Los resultados obtenidos con ambos métodos son bastante similares: se observan dos mínimos, correspondientes al estado plegado y al desplegado. Hemos observado que para las temperaturas k_BT = 13 y k_BT = 13.5, la proteína está pasando del estado plegado al desplegado, mientras que para k_BT = 14 la proteína ya está prácticamente desplegada. Además, al estudiar la energía libre tomando Q como coordenada de reacción se ha comprobado que el resultado varía con respecto a la distancia crítica.

Since we already have the theoretical background, we can address the analysis carried out in this work. To study the phase transition, the protein described in section 4 (in its native state) have been compared with numerous different configurations of the same amino acid chain at different temperatures. These configurations have been obtained using Parallel Tempering Monte Carlo (PTMC), as we have mentioned above. To compare them with the protein under study, the analysis of the free energy has been carried out in two different ways:

1. Using the number of contacts as the reaction coordinate.
2. Using the root-mean-square deviation (RMSD) as the reaction coordinate.

Nevertheless, before commenting the results, we need to know what a reaction coordinate is.

6.1. Reaction coordinates

Traditionally, there is a clear tendency to assimilate the folding process with a chemical reaction in which the reactant (unfolded protein) becomes a product (folded protein) and, as in any other reactant-product process, the reaction is governed by a limiting step called transition state. The transition state in a simple chemical reaction is a single conformation with unfavourable energy, which represents the main barrier between reactants and products. By similarity, in the case of the folding of a protein there should be some conformation that acts in a similar way to the transition state. In general, we can admit that if we consider that the native state (completely folded) of a protein admits only one possible structure, the number of possible configurations increases as we move away from this native state. Therefore, we can establish a parameter R based on the number of structures with a certain measure of similarity with the native structure.

In this way, like any other simple chemical reaction, the progress from the unfolded state to the native state can be represented by this value of R, that could work as a one-dimensional reaction coordinate. Thus, the process can be understood as the difussion throughout it, being able to represent the free energy through the folding process as a function of said reaction coordinate (figure 7). The free energy in function of this reaction coordinate (R) defined by [29] is:

$$F(R) = -k_B T \ln (P(R)) \quad (8)$$

Where P(R) is the probability of obtaining the value studied of the reaction coordinate R.

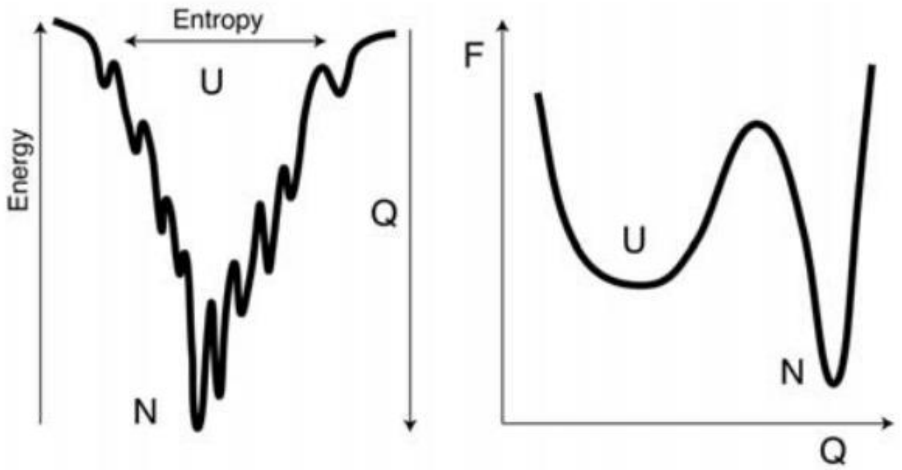


Figure 9. Left: the line shows an energy landscape funnel of the folding process at high energies many configurations are possible, and the entropy is large. The funnel guides the protein to the minimum, where it has less configuration freedom. During this process, the reaction coordinate (Q) increases. Right: Schematic free energy landscape of the protein folding process as a function of the reaction coordinate Q, two stable states separated by a barrier (the transition state) are shown [28].

Taking all the mentioned below into account, we also can theoretically construct a Cartesian space in which we represent free energy against R. With high temperatures, folding is a thermodynamically uphill process, so it is exponentially suppressed and states with low similarity to the native structure are favored, corresponding to an unfolded state. At low temperatures, states very similar to the native structure predominate (folded state). Logically, in between, we can define a situation where the folded and unfolded states predominate equally, appreciating two minima for free energy with an energy barrier between them, which arises from the incomplete cancellation of the entropy by the energy (figure 9). In fact, the free energy F incorporates the balance between two terms: the energy, that decreases as the native state is approached, and the entropy, which increases with unfolding.

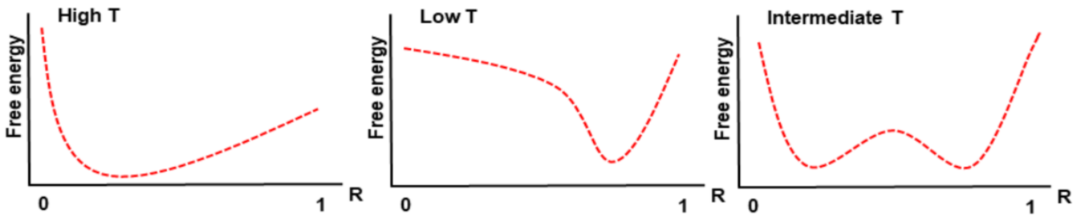


Figure 10: Scheme for the behavior of free energy against a folding reaction coordinate assumable to the similarity with the native state in three different temperature scenarios. The value R=1 corresponds to the native state. On the left the situation for high temperatures, with a single minimum for the free energy function near n = 0 (unfolded states predominate). At the middle, the scenario at low temperatures, where the minimum near 1 corresponding to a proximity with the native structure. At right, a scenario with an intermediate temperature between the two previous scenarios, where the free energy develops a two minimum, one of them similar to native structure and the other similar to the unfolded states, and with a free energy barrier between these minima.

In the study of protein folding, many variables have been used as reaction coordinates. Some of them are the fraction of native contacts (Q) and the root-mean-square deviation (RMSD), on which we will focus below.

6.2. The number of contacts as reaction coordinate

The number of non-covalent contacts shared between the native structure and any other configuration of the amino acid chain is often considered a good reaction coordinate, since they are assumed to be the only ones that play a role in the folding mechanism [30] in such a way that only those contacts are energetically favourable.

This assumption leads to a good correlation between energy and "nativity" [31] and, although it remains a mere simplification, it has allowed the development of models that successfully predict a wide range of properties of the folding process [32]. What we have called "nativity" is just the fraction of native contacts in a certain structure with respect to those existing in the native structure. Said fraction of native contacts (Q) varies between 0 (there is no native contact in the structure studied) and 1 (the structure studied has the same number of native contacts as the native structure and therefore, they are very similar). Thus:

$$Q = \frac{Q_S}{Q_T} \tag{9}$$

Where Q_S is the number of non-covalent contacts shared between the simulated and the native structures. Q_T is the total number of non-covalent contacts in the native structure.

We need to realise that with the assumption made, non-native contacts (those non-covalent contacts that are not present in the native structure) do not play any specific role in the folding mechanism, a simplification that has been questioned with some frequency. In fact, simply by pure logic and considering that the roughness that characterises the energy landscape, making and breaking the non-native contacts may lead to a slowdown of the folding process, leading to an increase of the energy barrier that surrounds a local minimum. The opposite can also happen since these interactions could accelerate folding by reducing the free energy barrier [33]. In this way, although a funnel energy landscape eases the folding towards a global minimum, alternative scenarios in which certain non-native contacts help guide the folding cannot be rejected *a priori.*

But even with the aforementioned, in general it is accepted that the value of Q is an adequate reaction coordinate to study the folding process, especially for small proteins, and that in their folding process they have to overcome an energy barrier associated with a transition state [34]. Thus, when representing Q against free energy it is usual to observe two minimums (unfolded and folded state) separated by a maximum that corresponds to the transition state (figure 9).

One of the advantages of using Q as reaction coordinate is that in its definition it is possible to control the resolution of the order parameter, letting us modify the notion of similarity.

As we have mentioned, we obtained numerous configurations for the protein studied at different temperatures by Parallel Tempering Monte Carlo. The temperatures studied in this work are $k_BT = 13$, $k_BT = 13.5$, $k_BT = 14$ and 10000 configurations were obtained for each one of them.

These temperatures correspond to the phase transition from the native state to an unfolded structure.

First of all, we compared each configuration with the native protein, obtaining the number of native contacts. In this case, we considered that two amino acids in a configuration are in contact if the distance between them is contained in the interval:

$$\sigma_{ij}\delta_{-} < r_{ij} < \sigma_{ij}\delta_{+}$$

Being $\delta_{\pm} = \left(\frac{\left(1\pm\sqrt{\frac{1}{3}}\right)}{2}\right)^{-1/6}$ and σ_{ij} is the distance for which the potential between particles is 0, as we have defined before.

Once the number of native contacts for each configuration was obtained, we calculated the native-contact fraction (9) and its probability, P(Q). For doing this, we divided the interval that contains all the possible values of Q in 150 subintervals (x-axis) and normalized the probability of obtaining a value of Q in any of those subintervals (y-axis). By representing P(Q) against Q for each of the temperatures analysed, we obtain:

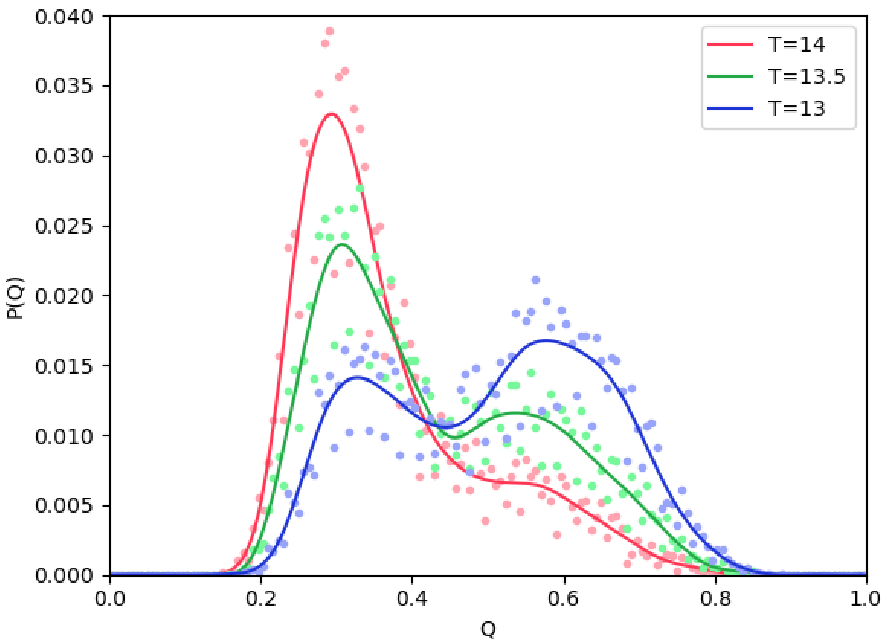


Figure 11: Probability of obtaining a certain value of the native-contact fraction in the ensemble of configurations for each one of the temperatures analysed.

The results present two maximums and one minimum, whose size varies depending on the temperature. The maximums represent the two possible states of the amino acid chain, this is the unfolded (fewer contacts) and the folded (more contacts) states, respectively. For the lowest temperature (k_BT = 13) the second maximum is higher, indicating a higher probability of finding the protein in its native state (folded structure). However, the size of both maximums is very similar for this temperature denoting that, since we can also find the amino acid chain unfolded, the phase transition has already started. Even though for k_BT = 13.5 it is more likely to find the protein unfolded, there still are possibilities of finding it folded. This is because the transition process occurs between k_BT = 13 and k_BT = 13.5. Finally, for the highest temperature (k_BT = 14) the first maximum is much bigger, being the second one almost imperceptible. This means that the protein is totally unfolded at this temperature.

Next, we studied the free energy in function of the native-contact fraction. Since Q is the reaction coordinate, and taking into account the equation (8), the free energy was obtained as follows:

$$F(Q) = -k_B T\ ln\ (P(Q)) \tag{10}$$

But for simplicity we calculated $\frac{F\ Q}{k_B}$, obtaining the following results:

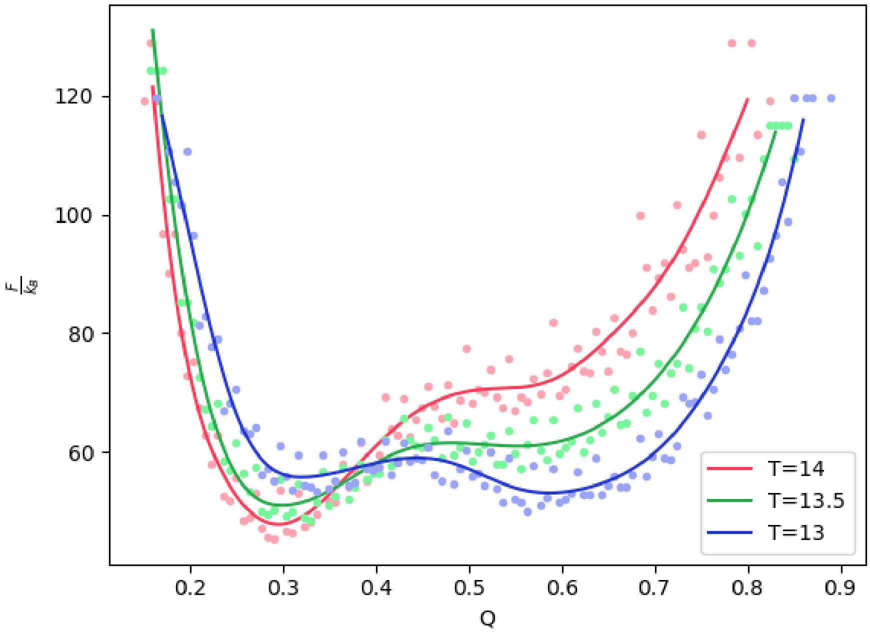


Figure 12: Profile of the free energy landscape as a function of the native-contact fraction for the ensemble of configurations for each of the temperatures analysed in this work.

The results presented in figure 12 are in some way similar to those of figure 11. At the lowest temperature we detect two minima and one maximum (as expected from section 6.1). The first minimum corresponds to the unfolded structure, and the second one to the folded structure. We observe that the folded structures present high probabilities. Nevertheless, as for k_BT = 13.5, in k_BT = 14 only one minimum is clearly visible indicating that unfolded states predominate.

Now that we have a general vision of the behaviour of the protein object of study at each one of the temperatures analysed, we decided to study the variation of P(Q) and $\frac{\mathrm{F(Q)}}{\mathrm{k_B}}$ for different contact maps:

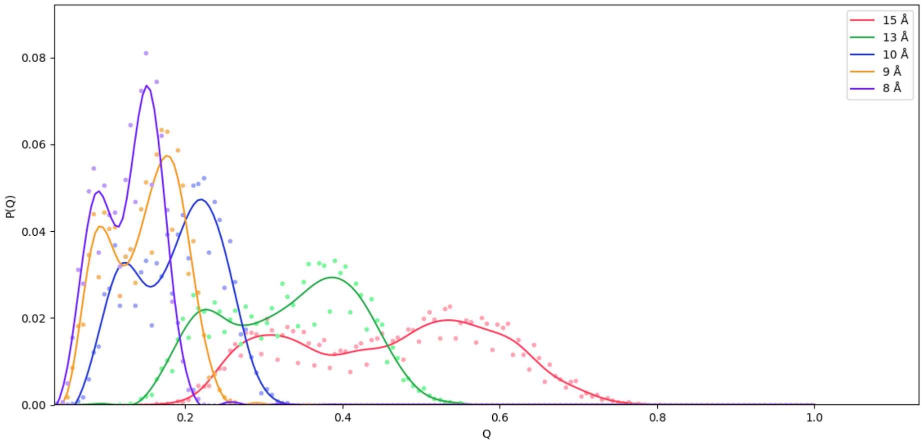


Figure 13: Probability of the native-contact fraction in the ensemble of configurations for k_BT=13. The figure shows the results obtained from different contact maps with critical distances between 8Å (purple) to 15Å (pink).

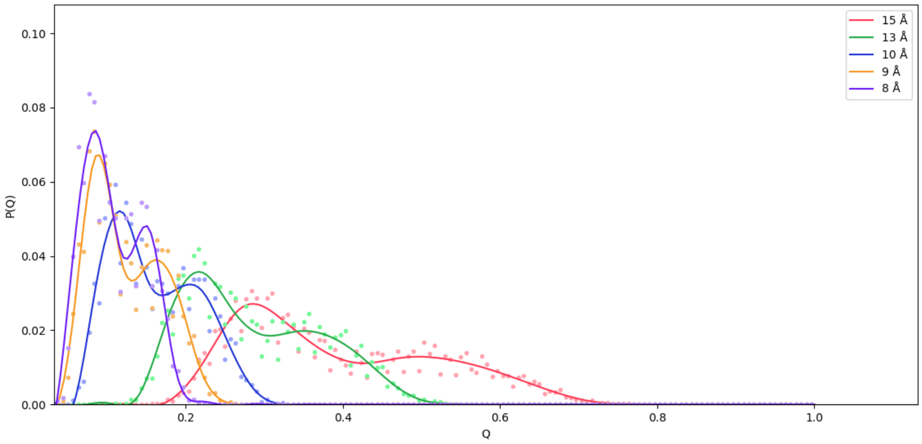


Figure 14: Probability of the native-contact fraction in the ensemble of configurations for k_BT=13.5. The figure shows the results obtained from different contact maps with critical distances between 8Å (purple) to 15Å (pink).

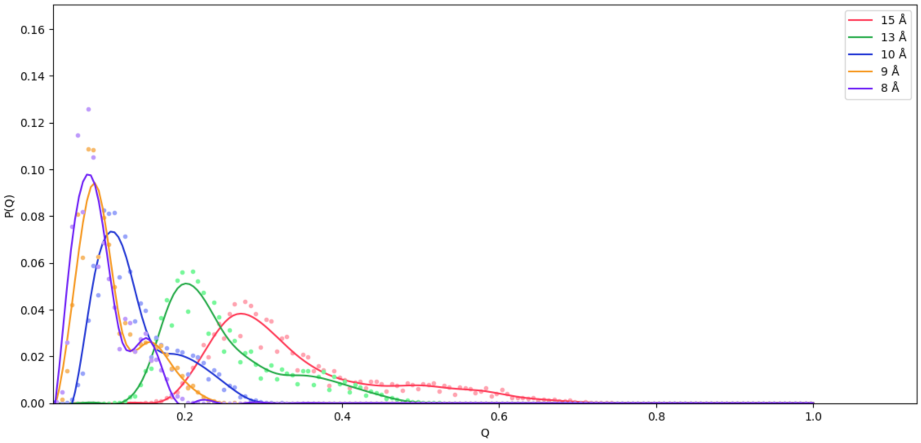


Figure 15: Probability of the native-contact fraction in the ensemble of configurations for k_BT=14. The figure shows the results obtained from different contact maps with critical distances between 8Å (purple) to 15Å (pink).

Comparing figures 13 and 14, it is totally clear the effect that we have mentioned before: we observe two maximums for both temperatures, but the one that corresponds to the folded state is bigger for $k_B\mathrm{T} = 13$ and the one that corresponds to the unfolded state is bigger for $k_B\mathrm{T} = 13.5$. This reaffirms that the transition process occurs between both temperatures. Also, the results are clearer for the lower critical distances, i.e., 10 Å, 9 Å and 8 Å. However, for very low critical distances like 7 Å, the two minima do not appear, causing the loss of information about one of the states of the protein. For this reason, we decided to discard this measure.

In the case of $k_B\mathrm{T} = 14$ it is clear that the protein is unfolded independently of the critical distance studied (figure 15).

The results obtained for the free energy for the different contact maps are as expected. We observe two minimums and one maximum. For $k_B\mathrm{T} = 13$ (figure 16) and $k_B\mathrm{T} = 13.5$ (figure 17), both minimums are clear but one of the is bigger in each case, indicating how the protein goes from the folded state to the unfolded one. The unfolded state is completely reached at $k_B\mathrm{T} = 14$ (figure 18), where we can only see one minimum.

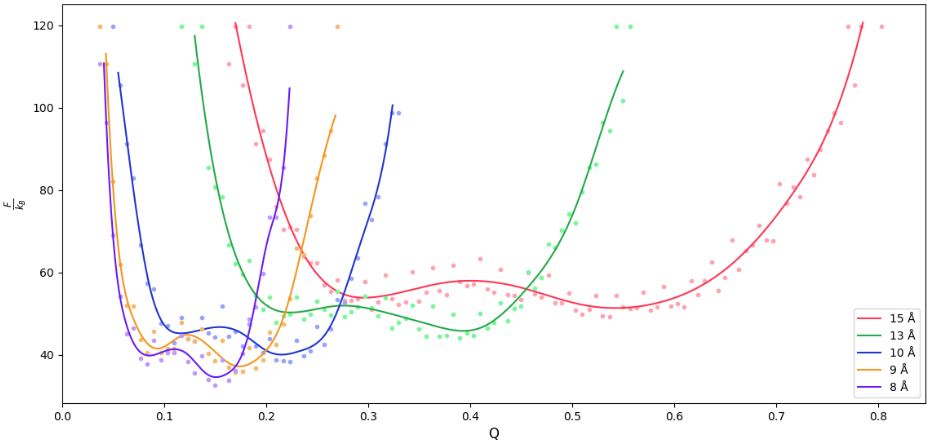


Figure 16: Profile of the free energy landscape as a function of the native-contact fraction for the ensemble of configurations for k_BT=13. The figure shows the results obtained from different contact maps with critical distances between 8Å (purple) to 15Å (pink).

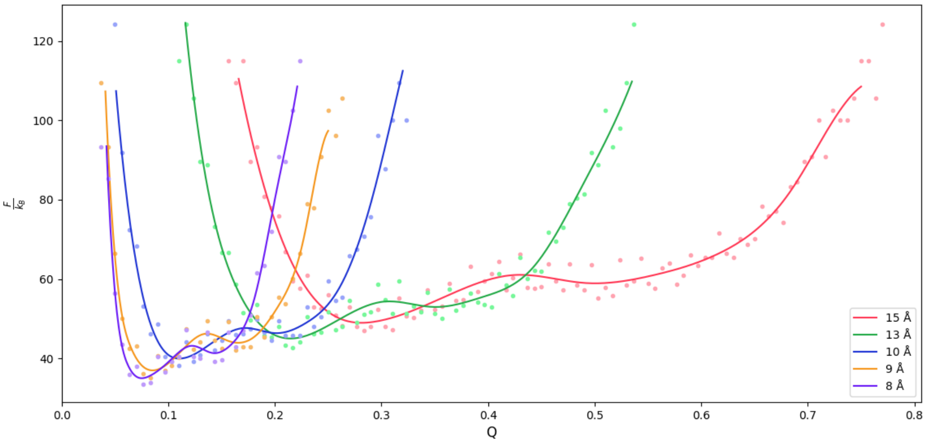


Figure 17: Profile of the free energy landscape as a function of the native-contact fraction for the ensemble of configurations for k_BT=13.5. The figure shows the results obtained from different contact maps with critical distances between 8Å (purple) to 15Å (pink).

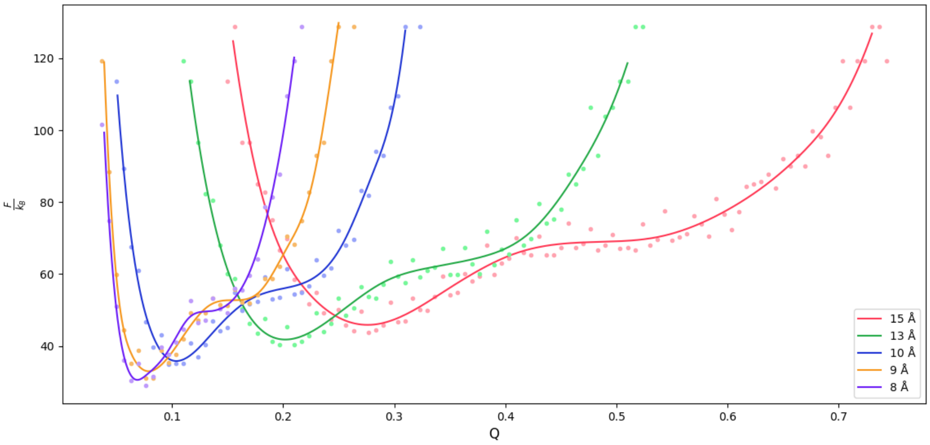


Figure 18: Profile of the free energy landscape as a function of the native-contact fraction for the ensemble of configurations for k_BT=14. The figure shows the results obtained from different contact maps with critical distances between 8Å (purple) to 15Å (pink).

Another interesting result is that both, in the representation of P(Q) and in that of F(Q) it is observed how the width of the function increases with the critical distance, and the function also moves to the right. This is because the higher the critical distance, the higher the number of contacts.

6.3. RMSD as reaction coordinate

Q is not the only parameter that can be used for the study of the protein folding process and there are other alternatives. Perhaps one of the most frequently used (along with Q) is the root-mean-square distance (RMSD). The RMSD quantifies the proximity or distance of a certain structure with another (which could be the native state) based on the existing spatial differences (within a cartesian space) between two monomers that make up both structures [35]. This measure is defined as follows:

$$RMSD = \frac{1}{N}\sqrt{\sum_{ij}\left(\overrightarrow{r_{ij}} - \overrightarrow{r_{ij}}^{T}\right)^2} \tag{11}$$

Being N the number of monomers that make up the polymer, $\overrightarrow{r_{ij}}$ the distance between a pair of monomers in the sampled structure and $\overrightarrow{r_{ij}}^{T}$ the distance between that pair of monomers in the target structure.

According to this, it is expected that two protein chains will be more similar as the RMSD is lower, although this does not have to coincide with reality. For example, suppose two conformations (the native state and the one to compare) that have two α-helices, in such a way that one of the helices is very similar in both cases but the other one is very different. In this case, the RMSD between both conformations could be high, suggesting that there is not structural similarity. Precisely in a case like this, Q will continue to show similarities, so it is usual to study both parameters together.

The procedure for obtaining the free energy profile for each temperature in function of the RMSD (the reaction coordinate in this case) is similar to the one followed in the previous section. First, we calculated the RMSD between the protein in its native structure and each one of the configurations for the three temperatures analysed ($k_B\mathrm{T} = 13$, $k_B\mathrm{T} = 13.5$ and $k_B\mathrm{T} = 14$). After this, we obtained the normalized values of the RMSD and their probability. Finally, the free energy was calculated following the equation (8), that for this case results:

$$F(RMSD) = -k_B T\, ln\, (P(RMSD)) \tag{12}$$

As in the previous section we calculated $\frac{\mathrm{F(RMSD)}}{\mathrm{k_B}}$, obtaining the following results:

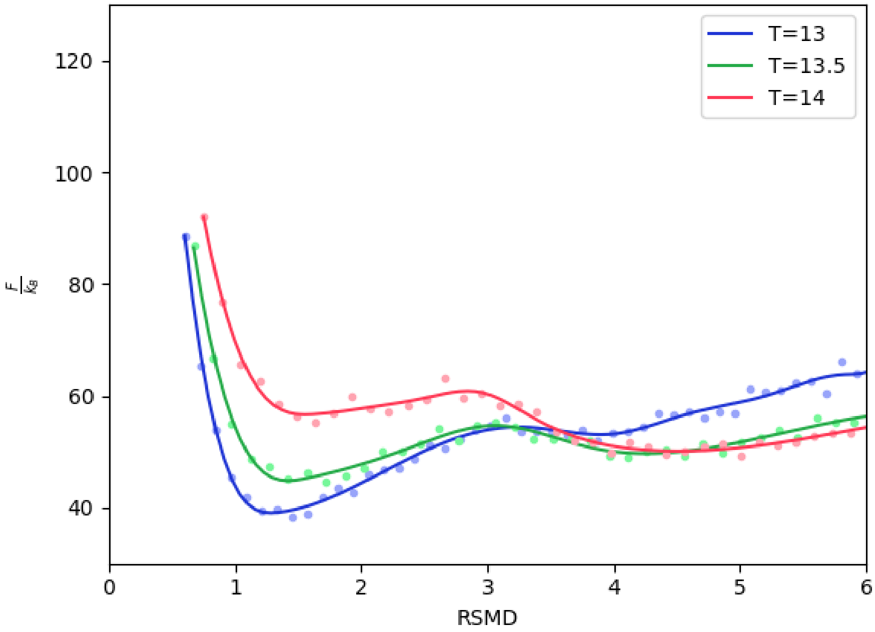


Figure 19: Profile of the free energy landscape as a function of RMSD for the ensemble of configurations for k_BT=13 (blue), k_BT=13.5 (green) and k_BT=14 (pink).

Even though these results seem to be the opposite to the ones obtained using Q as the reaction coordinate (section 6.2, figure 12), they are quite similar. In this case, we have two minimums and one maximum too, but now the first minimum corresponds to the folded state and the second one to the unfolded state. This is because, the more similar the two protein chains compared (native state and the obtained configurations) are, the smaller is the value of the RMSD. In general, the first minimum is located near RMSD=1, with a F/k_B value similar to the obtained using Q as the reaction coordinate. For $k_B\mathrm{T} = 13$, the first minimum is much bigger, indicating that the different configurations of the protein at this temperature are similar to the protein on its native state. Therefore, the protein at this temperature is folded. For $k_B\mathrm{T} = 13.5$, both minimums have a similar size, denoting the phase transition. Finally, for $k_B\mathrm{T} = 14$ the second minimum is much bigger, showing that the configurations are more different to the protein on its native state. These results coincide with the ones observed in the previous section.

7. Conclusions

Abstract

Se ha observado que el comportamiento de la proteína estudiada en el proceso de plegamiento consta de dos estados bien diferenciados: uno de baja energía (estado plegado) y otro de alta energía (estado desplegado). La temperatura para la cual las probabilidades de obtener alguno de estos estados son iguales se denomina temperatura de plegamiento y se suele estimar como la temperatura a la cual la capacidad calorífica alcanza su valor máximo. Usando como coordenada de reacción la fracción de contactos nativos se ha obtenido que dicha temperatura se encuentra aproximadamente entre k_BT = 13 y k_BT = 13.5. Por otra parte, cabe destacar que, tal y como era de esperar, la variación de la distancia crítica afecta a los resultados obtenidos empleando Q como coordenada de reacción. Además, al tomar la RMSD como coordenada de reacción, las curvas obtenidas para la energía libre son similares a las obtenidas para Q. Esto implica que ambas opciones son buenas coordenadas de reacción para estudiar el plegamiento de las proteínas.

It is known that the folding/unfolding process of the protein in the funneled energy landscape is determined by the temperature. High temperatures, determine a predominance of high-energy structures, but as the temperature falls the lower energy will predominate. Under a certain transition temperature, such systems will fall into a low-energy funnel, and may remain trapped in it. Precisely at this time, when a sufficiently large core of the structure can be formed, the rest of the low-energy structure is rapidly reached [36]. For small proteins, as in our case, it has been shown that folding process presents this type of cooperative behavior, observing two differentiated states, one, corresponding to the unfolded protein, and another that describes the folded protein.

From the Parallel Tempering Montecarlo sampling a set of 10000 structures for each of three temperature scenarios were obtained. Since, we can define the folding temperature (T_f), as the temperature at which the free energy reaches two similar minima over the reaction coordinate, that is the temperature at which the probability for the folded configuration is equal to the probability for the unfolded configuration [37] in terms of a reaction coordinate (R). This reaction coordinate expresses the similarity with the native structure. Experimentalists usually estimate this transition temperature as the temperature at which the heat capacity attains its peak value [38]. In this sense, we must highlight that Hernández-Rojas & Gómez Llorente [1] obtained for the same protein chain that we have used (in the canonical ensemble too) a peak value of heat capacity between k_BT = 13 and k_BT = 14. The results that we have obtained reveal that for k_BT = 13 the protein is still folded. However, for this temperature we observed two clear minimums, as for k_BT = 13.5. This indicates that the folding temperature (T_f) for our protein model is located between both temperatures. We also observed that for k_BT = 14 the protein is already folded.

During the analysis of the protein folding process using Q as reaction coordinate, the results were clearly affected by the variation of the critical distance. This was to be expected since the definition of Q itself depends on the chosen critical distance. We have tried different values of the critical distance from 8 Å to 15 Å, being able to observe that, both for P(Q) and for the free energy, the function moves to the right (higher values of Q) with the increase of the critical distance. We also observed that the function widens as the critical distance increases and that the peaks of both functions are more pronounced. This is because for greater critical distances, the number of contacts rises, and so the value of Q. However, we realised that the critical distance values must be chosen with some care. This is because, for very small distances, it was observed that one of the peaks disappeared and, therefore, the information about one of the states was lost. On the other hand, the results obtained for the free energy using the RMSD as the reaction coordinate are very similar to the ones obtained using Q, with minimum energy values at similar levels. This shows that both methods are good for analysing the folding of, at least, small proteins like the one studied in this work.

References

[1] HERNANDEZ-ROJAS, J., & LLORENTE, J. G. (2008*). Microcanonical versus canonical analysis of protein folding.* Physical review letters, 100(25), 258104.

[2] HERNANDEZ-ROJAS, J., & LLORENTE, J. G. Private communication.

[3] CLEMENTI, C., MARITAN, A., & BANAVAR, J. R. (1998). *Folding, design, and determination of interaction potentials using off-lattice dynamics of model heteropolymers.* Physical review letters, 81(15), 3287.

[4] BOERSMA, C. (January 14, 2021). *Amino Acids and Their Production during the Photolysis of Astroohysically Relevant Ices.* The Astrophysics & Astrochemistry Lab at NASA Ames Research Center. Recovered on June 10, 2021 from http://www.astrochem.org/sci/Amino_Acids.php

[5] LUBRIZOL LIFE SCIENCE. (2021). *Protein Structure: Primary, Secondary, Tertiary, Quaternary Structures.* HEALTH CDMO Division. https://lubrizolcdmo.com/technical-briefs/protein-structure/

[6] CREIGHTON T. E. (1990). *Protein folding.* The Biochemical journal, 270(1), 1–16.

[7] GHÉLIS, C. (2012). *Protein folding.* Academic Press.

[8] FINKELSTEIN, A. V. (2018). *50+ years of protein folding.* Biochemistry (Moscow), 83(1), S3-S18.

[9] LEVINTHAL, C. (1968). *Are there pathways for protein folding?.* Journal de chimie physique, 65, 44-45.

[10] DILL, K. A., & MACCALLUM, J. L. (2012). *The protein-folding problem, 50 years on.* Science, 338(6110), 1042-1046.

[11] DILL, K. A. (1990). *Dominant forces in protein folding.* Biochemistry, 29(31), 7133-7155.

[12] RAVAL, A., PIANA, S., EASTWOOD, M.P., DROR, R.O. AND SHAW, D.E. (2012). *Refinement of protein structure homology models via long, all-atom molecular dynamics simulations.* Proteins, 80: 2071-2079.

[13] LEOPOLD, P. E., MONTAL, M., & ONUCHIC, J. N. (1992). *Protein folding funnels: a kinetic approach to the sequence-structure relationship.* Proceedings of the National Academy of Sciences, 89(18), 8721-8725.

[14] ENGLANDER, S. W., MAYNE, L., & KRISHNA, M. M. (2007). *Protein folding and misfolding: mechanism and principles.* Quarterly reviews of biophysics, 40(4), 287.

[15] VOELZ, V. A., & DILL, K. A. (2007). *Exploring zipping and assembly as a protein folding principle.* Proteins: Structure, Function, and Bioinformatics, 66(4), 877-888.

[16] WHITFORD, P. C., NOEL, J. K., GOSAVI, S., SCHUG, A., SANBONMATSU, K. Y., & ONUCHIC, J. N. (2009). *An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields.* Proteins: Structure, Function, and Bioinformatics, 75(2), 430-441.

[17] PERLMUTTER, J. D., DRASLER, W. J., XIE, W., GAO, J., POPOT, J. L., & SACHS, J. N. (2011). *All-atom and coarse-grained molecular dynamics simulations of a membrane protein stabilizing polymer.* Langmuir, 27(17), 10523-10537.

[18] SOCCI, N. D., & ONUCHIC, J. N. (1994). *Folding kinetics of proteinlike heteropolymers.* The Journal of chemical physics, 101(2), 1519-1528.

[19] KOUZA, M. (2013). *Numerical Simulation of Folding and Unfolding of Proteins.* arXiv preprint arXiv:1308.2380.

[20] DILL, K. A., OZKAN, S. B., WEIKL, T. R., CHODERA, J. D., & VOELZ, V. A. (2007). *The protein folding problem: when will it be solved?.* Current opinion in structural biology, 17(3), 342-346.

[21] ONUCHIC, J. N., LUTHEY-SCHULTEN, Z., & WOLYNES, P. G. (1997). *Theory of protein folding: the energy landscape perspective.* Annual review of physical chemistry, 48(1), 545-600.

[22] WALES, D. J., & DOYE, J. P. (1997). *Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms.* The Journal of Physical Chemistry A, 101(28), 5111-5116.

[23] VENDRUSCOLO, M., KUSSELL, E., & DOMANY, E. (1997). *Recovery of protein structure from contact maps.* Folding and Design, 2(5), 295-306.

[24] MIRNY, L., & DOMANY, E. (1996). *Protein fold recognition and dynamics in the space of contact maps.* Proteins: Structure, Function, and Bioinformatics, 26(4), 391-410.

[25] EARL, D. J., & DEEM, M. W. (2005). *Parallel tempering: Theory, applications, and new perspectives.* Physical Chemistry Chemical Physics, 7(23), 3910-3916.

[26] MITSUTAKE, A., & OKAMOTO, Y. (2000). *Replica-exchange simulated tempering method for simulations of frustrated systems.* Chemical Physics Letters, 332(1-2), 131-138.

[27] NIGRA, P., CARIGNANO, M. A., & KAIS, S. (2001). *Study of phase changes of the water octamer using parallel tempering and multihistogram methods.* The Journal of Chemical Physics, 115(6), 2621-2628.

[28] BOLHUIS, P. G. (2009). *Two-state protein folding kinetics through all-atom molecular dynamics based sampling.* Frontiers in bioscience (Landmark edition), 14, 2801-28.

[29] VAN GUNSTEREN, W. F., DAURA, X., & MARK, A. E. (2002). *Computation of free energy.* Helvetica Chimica Acta, 85(10), 3113-3129.

[30] WOLYNES, P. G. (2005). *Recent successes of the energy landscape theory of protein folding and function.* Quarterly reviews of biophysics, 38(4), 405.

[31] NYMEYER, H., GARCÍA, A. E., & ONUCHIC, J. N. (1998). *Folding funnels and frustration in off-lattice minimalist protein landscapes.* Proceedings of the National Academy of Sciences, 95(11), 5921-5928.

[32] KUBELKA, J., HENRY, E. R., CELLMER, T., HOFRICHTER, J., & EATON, W. A. (2008). *Chemical, physical, and theoretical kinetics of an ultrafast folding protein.* Proceedings of the National Academy of Sciences, 105(48), 18655-18662.

[33] CLEMENTI, C., & PLOTKIN, S. S. (2004). *The effects of nonnative interactions on protein folding rates: theory and simulation.* Protein Science, 13(7), 1750-1766.

[34] BEST, R. B., HUMMER, G., & EATON, W. A. (2013). *Native contacts determine protein folding mechanisms in atomistic simulations.* Proceedings of the National Academy of Sciences, 110(44), 17874-17879.

[35] SCHLITTER, J., SWEGAT, W., & MÜLDERS, T. (2001). *Distance-type reaction coordinates for modelling activated processes.* Molecular modeling annual, 7(6), 171-177.

[36] LIFSHITZ, E.M., PITAEVSKII, L.P. "Physical Kinetics." SYKES, J.B., FRANKLIN, R.N. trans., *Course of Theoretical Physics*, Vol. 10. Oxford: Pergamon Press, 1981: 427-438.

[37] GUTIN, A.M., ABKEVICH, V.I. & SHAKHNOVICH, E.I. (1995). *Is burst hydrophobic collapse necessary for protein folding?* Biochemistry 34, 3066-3076.

[38] TRAVASSO, R. D., FAÍSCA, P. F., & REY, A. (2010). *The protein folding transition state: Insights from kinetics and thermodynamics.* The Journal of chemical physics, 133(12), 09B612.

Appendix A

Table 1: sequence of amino acids and position of each one in the three-dimensional space.

Position	Monomer type	X (Å)	Y (Å)	Z (Å)
1	N	16,668150	3,652860	71,255260
2	C	13,836850	2,503840	73,501940
3	C	12,505450	1,917730	76,978450
4	C	13,521510	0,919690	80,460980
5	C	16,395540	0,401930	82,855190
6	B	18,607290	2,861180	84,719400
7	B	17,626320	6,208500	86,223930
8	B	13,912780	5,897270	86,964830
9	O	11,005160	3,482440	86,580070
10	N	9,842840	-0,129100	86,385620
11	B	11,254460	-3,615830	85,853980
12	B	11,754110	-5,520540	82,604910
13	N	9,083470	-4,456480	80,121850
14	C	6,904100	-1,361020	79,860450
15	O	6,470460	2,285170	80,775910
16	O	7,896320	5,783980	80,754280
17	C	11,134440	7,719030	80,805440
18	O	14,854730	7,915890	80,241620
19	O	17,934540	6,423780	78,663520
20	O	19,911780	3,435820	77,477940
21	B	19,936430	-0,171900	76,343830
22	B	17,912410	-3,362810	76,562420
23	N	14,525400	-5,053810	76,693000
24	B	11,847420	-4,521570	74,059340
25	N	8,674300	-2,695650	73,063420
26	N	6,713470	0,552710	73,049030
27	N	6,429200	4,179840	74,127650
28	N	8,511160	7,354170	74,101000
29	N	11,788470	9,270160	74,013130
30	B	15,504760	9,687980	73,346130

Table 2: σ_{ij} values for each pair of monomers

Monomer i	**Monomer j**	$\boldsymbol{\sigma_{ij}}$ (Å)
1	3	6,459669
1	4	8,968332
1	5	10,723694
1	6	12,150345
1	7	13,566126
1	8	14,351284
1	9	14,550256
1	10	15,151319
1	11	15,290811
1	12	13,699298
1	13	12,636085
1	14	12,400391
1	15	12,443398
1	16	11,637801
1	17	10,457029
1	18	8,993110
1	19	7,164500
1	20	6,269936
1	21	6,397982
1	22	7,851561
1	23	9,261808
1	24	8,811124
1	25	9,199219
1	26	9,381222
1	27	9,433086
1	28	8,347799
1	29	7,062508
1	30	5,797551
2	4	6,319625
2	5	8,802203
2	6	10,844335
2	7	12,260228
2	8	12,349604
2	9	11,932294
2	10	12,218543
2	11	12,466528
2	12	10,943459
2	13	9,526843
2	14	9,041794
2	15	9,184423
2	16	8,811659
2	17	8,317759
2	18	7,716512
2	19	6,824756
2	20	6,510762
2	21	6,454990
2	22	6,876631
2	23	7,280411
2	24	6,533475
2	25	6,531401
2	26	6,577634
2	27	6,760464
2	28	6,428533
2	29	6,310836
2	30	6,575462
3	5	6,380866
3	6	8,797947
3	7	10,151897
3	8	9,659760
3	9	8,776824
3	10	8,901282
3	11	9,374366
3	12	8,320975
3	13	7,023201
3	14	6,340085
3	15	6,366768
3	16	6,333349
3	17	6,313414
3	18	6,412645
3	19	6,442412
3	20	6,719392
3	21	6,859456
3	22	6,697645
3	23	6,437984
3	24	6,315914
3	25	6,381751
3	26	6,365814
3	27	6,328137
3	28	6,555943
3	29	7,107034
3	30	8,098306
4	6	6,151176
4	7	7,873493
4	8	7,321240
4	9	6,345137

Monomer i	Monomer j	σ_{ij} (Å)
4	10	6,304612
4	11	6,611680
4	12	6,259525
4	13	6,224455
4	14	6,264247
4	15	6,405204
4	16	6,616126
4	17	6,420835
4	18	6,333195
4	19	6,463996
4	20	6,652970
4	21	6,811878
4	22	6,458037
4	23	6,348492
4	24	7,604255
4	25	8,490584
4	26	8,953679
4	27	8,945144
4	28	9,205800
4	29	9,520446
4	30	10,206060
5	7	6,078752
5	8	6,497300
5	9	6,437747
5	10	6,628829
5	11	6,389905
5	12	6,706277
5	13	8,178849
5	14	8,978248
5	15	9,155509
5	16	9,107767
5	17	8,207473
5	18	7,210332
5	19	6,670939
5	20	6,341061
5	21	6,579771
5	22	6,671086
5	23	7,507421
5	24	9,818347
5	25	11,412024
5	26	12,242166
5	27	12,245274
5	28	12,169555
5	29	11,884926
5	30	11,872983
6	8	5,366623
6	9	6,984267
6	10	8,361423
6	11	8,773504
6	12	9,821414
6	13	11,445710
6	14	11,878332
6	15	11,358945
6	16	10,471527
6	17	8,667614
6	18	6,904784
6	19	6,293697
6	20	6,577982
6	21	7,957192
6	22	9,140135
6	23	10,656575
6	24	13,002622
6	25	14,491444
6	26	14,970465
6	27	14,423881
6	28	13,663596
6	29	12,694737
6	30	12,168263
7	9	6,382911
7	10	8,931762
7	11	10,432688
7	12	12,122815
7	13	13,324192
7	14	12,976394
7	15	11,584181
7	16	9,928238
7	17	7,656545
7	18	6,096909
7	19	6,733100
7	20	8,400689
7	21	10,603926
7	22	12,093921
7	23	13,416815
7	24	15,323399
7	25	16,236055
7	26	16,048495

Monomer i	**Monomer j**	$\boldsymbol{\sigma_{ij}}$ (Å)
7	27	14,807409
7	28	13,571064
7	29	12,394045
7	30	12,062194
8	10	6,494267
8	11	8,857711
8	12	11,061487
8	13	11,856276
8	14	10,974781
8	15	9,193565
8	16	7,684604
8	17	6,240095
8	18	6,326183
8	19	8,211169
8	20	10,215955
8	21	12,087676
8	22	12,902725
8	23	13,397617
8	24	14,874786
8	25	15,282305
8	26	14,748265
8	27	13,342463
8	28	12,512991
8	29	12,094125
8	30	12,689426
9	11	6,360978
9	12	8,793872
9	13	9,266242
9	14	8,216572
9	15	6,647664
9	16	6,219124
9	17	6,388738
9	18	7,702187
9	19	9,712706
9	20	11,332063
9	21	12,488025
9	22	12,448003
9	23	12,081851
9	24	13,240867
9	25	13,396452
9	26	12,914421
9	27	11,853366
9	28	11,862106
9	29	12,360218
9	30	13,634852
10	12	6,106566
10	13	6,807795
10	14	6,459571
10	15	6,227830
10	16	7,470796
10	17	8,655911
10	18	10,057534
10	19	11,528605
10	20	12,384920
10	21	12,651308
10	22	11,711434
10	23	10,592334
10	24	11,766335
10	25	12,122645
10	26	12,215154
10	27	11,978674
10	28	12,871808
10	29	13,955374
10	30	15,394376
11	13	5,506009
11	14	6,889048
11	15	8,145400
11	16	9,970023
11	17	11,049805
11	18	11,862994
11	19	12,497988
11	20	12,447561
11	21	11,858714
11	22	10,228348
11	23	8,813412
11	24	10,512604
11	25	11,630190
11	26	12,639352
11	27	13,252383
11	28	14,521054
11	29	15,597742
11	30	16,707165
12	14	6,191971
12	15	8,551015
12	16	10,747515
12	17	11,905912

Monomer i	Monomer j	σ_{ij} (Å)
12	18	12,456321
12	19	12,484824
12	20	11,748150
12	21	10,354513
12	22	7,986056
12	23	5,900408
12	24	7,621763
12	25	9,247852
12	26	11,007390
12	27	12,394308
12	28	14,028071
12	29	15,232594
12	30	16,213943
13	15	6,454495
13	16	9,174074
13	17	10,994811
13	18	12,139562
13	19	12,554776
13	20	12,188867
13	21	10,951972
13	22	8,602602
13	23	5,848549
13	24	5,914136
13	25	6,472513
13	26	7,972108
13	27	9,627731
13	28	11,791619
13	29	13,580889
13	30	15,089629
14	16	6,450893
14	17	8,937436
14	18	10,854943
14	19	12,052581
14	20	12,531952
14	21	12,079555
14	22	10,434228
14	23	8,128995
14	24	7,349980
14	25	6,381478
14	26	6,302121
14	27	7,102876
14	28	9,390042
14	29	11,628598
14	30	13,740648
15	17	6,353859
15	18	8,966865
15	19	10,977643
15	20	12,340494
15	21	12,787310
15	22	11,975661
15	23	10,402747
15	24	9,755843
15	25	8,408536
15	26	7,054153
15	27	6,155958
15	28	7,655143
15	29	9,837154
15	30	12,295932
16	18	6,455060
16	19	9,100958
16	20	11,237797
16	21	12,538970
16	22	12,620475
16	23	11,884014
16	24	11,480833
16	25	10,220262
16	26	8,375732
16	27	6,243729
16	28	6,120619
16	29	7,585913
16	30	10,053571
17	19	6,416234
17	20	9,143000
17	21	11,196049
17	22	12,139307
17	23	12,316575
17	24	12,450804
17	25	11,756260
17	26	10,197794
17	27	7,944978
17	28	6,432725
17	29	6,238906
17	30	7,888019
18	20	6,459665
18	21	9,127011
18	22	10,868820

Monomer i	Monomer j	$\boldsymbol{\sigma_{ij}}$ (Å)
18	23	11,948109
18	24	12,639190
18	25	12,644888
18	26	11,667168
18	27	9,836440
18	28	7,872688
18	29	6,303639
18	30	6,367968
19	21	6,435109
19	22	8,880534
19	23	10,776195
19	24	11,893926
19	25	12,599988
19	26	12,330401
19	27	11,184344
19	28	9,363427
19	29	7,333791
19	30	5,989835
20	22	6,348409
20	23	8,961451
20	24	10,570702
20	25	12,065039
20	26	12,652690
20	27	12,368081
20	28	11,131984
20	29	9,412404
20	30	7,731104
21	23	6,473905
21	24	8,482781
21	25	10,711990
21	26	12,160868
21	27	12,778703
21	28	12,337657
21	29	11,290077
21	30	9,983538
22	24	5,961587
22	25	8,819910
22	26	11,004507
22	27	12,398556
22	28	12,855407
22	29	12,674004
22	30	12,122090
23	25	6,480724
23	26	9,149681
23	27	11,158901
23	28	12,477037
23	29	13,179003
23	30	13,456128
24	26	6,475526
24	27	9,108983
24	28	10,982468
24	29	12,289408
24	30	13,100425
25	27	6,493205
25	28	8,989831
25	29	11,035879
25	30	12,593683
26	28	6,329278
26	29	9,014681
26	30	11,284940
27	29	6,567283
27	30	9,460646
28	30	6,592430