molecules

MDPI

Article

Einstein Model of a Graph to Characterize Protein Folded/Unfolded States

Steve Tyler [1,†], Christophe Laforge [1], Adrien Guzzo [1,‡], Adrien Nicolaï [1], Gia G. Maisuradze [2] and Patrick Senet [1,*]

[1] Laboratoire Interdisciplinaire Carnot de Bourgogne, UMR CNRS 6303, Université de Bourgogne, 21078 Dijon CEDEX, France
[2] Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853, USA
* Correspondence: psenet@u-bourgogne.fr
† Current address: DEMR-ONERA/SONDRA-CentraleSupélec, Université Paris Saclay, 91190 Gif-sur-Yvette, France.
‡ Current address: INSERM U1903 CAPS, Université de Bourgogne, 21078 Dijon CEDEX, France.

Abstract: The folded structures of proteins can be accurately predicted by deep learning algorithms from their amino-acid sequences. By contrast, in spite of decades of research studies, the prediction of folding pathways and the unfolded and misfolded states of proteins, which are intimately related to diseases, remains challenging. A two-state (folded/unfolded) description of protein folding dynamics hides the complexity of the unfolded and misfolded microstates. Here, we focus on the development of simplified order parameters to decipher the complexity of disordered protein structures. First, we show that any connected, undirected, and simple graph can be associated with a linear chain of atoms in thermal equilibrium. This analogy provides an interpretation of the usual topological descriptors of a graph, namely the Kirchhoff index and Randić resistance, in terms of effective force constants of a linear chain. We derive an exact relation between the Kirchhoff index and the average shortest path length for a linear graph and define the free energies of a graph using an Einstein model. Second, we represent the three-dimensional protein structures by connected, undirected, and simple graphs. As a proof of concept, we compute the topological descriptors and the graph free energies for an all-atom molecular dynamics trajectory of folding/unfolding events of the proteins Trp-cage and HP-36 and for the ensemble of experimental NMR models of Trp-cage. The present work shows that the local, nonlocal, and global force constants and free energies of a graph are promising tools to quantify unfolded/disordered protein states and folding/unfolding dynamics. In particular, they allow the detection of transient misfolded rigid states.

Keywords: protein folding; intrinsically disordered proteins; graph theory; Kirchhoff index; Wiener index; molecular dynamics

Citation: Tyler, S.; Laforge, C.; Guzzo, A.; Nicolaï, A.; Maisuradze, G.G.; Senet, P. Einstein Model of a Graph to Characterize Protein Folded/Unfolded States. *Molecules* **2023**, *28*, 6659. https://doi.org/10.3390/molecules28186659

Academic Editor: Takeshi Kikuchi

Received: 15 August 2023
Revised: 11 September 2023
Accepted: 14 September 2023
Published: 16 September 2023

1. Introduction

In spite of significant advances in experimental [1–7], theoretical [8–19], and computational research [20–29], many questions related to protein folding remain unanswered [19]. In particular, a complete understanding of preferred folding pathways and misfolding and protein aggregation, which are related to neurodegenerative diseases, still remains a challenge. So far, none of these problems can be tackled by current deep-learning and other recent successful computational approaches of protein folding [25], as these methods relate two ensembles of end structures, the linear sequences of amino acids (completely unfolded unstable structures) and the protein folded geometries extracted from an experimental database, and have no reliable information on the ensemble of intermediate structures. It is worth noting that the inclusion of sequence databases in these methods only improves the results by a few percent, emphasizing the importance of information arising from the physical laws: the database of equilibrium experimental structures.

According to Anfinsen's principle, the equilibrium protein structure (native state in vivo) is the global minimum of the protein free energy in vitro [8], i.e., it is governed by the second law of thermodynamics. More precisely, the protein equilibrium structure is defined by the amino-acid sequence and the thermodynamical parameters of the environment (T, p, pH) [9]. The protein folding phenomenon is thus cast in thermodynamic terms as a phase transition between an unfolded state and a folded native state. From the Boltzmann law, we know that free energy always involves all microstates. The partition of microstates in folded and unfolded ensembles is at the heart of Landau's order parameter theory for which the free energy is expanded as a function of one macroscopic parameter that varies between the two phases. This two-state view is challenged by the nanoscopic nature of the macromolecules. Unlike a macroscopic system, there is no sharp unique melting temperature of the transition for a protein, as it depends on the physical property (order parameter) measured and its spatial localization [4,17,18]. Moreover, protein folding may occur through intermediate states or even via a continuum of states (barrierless folding) [6]. A predominant description of protein folding is thus the consideration of an expansion of free energy as a continuous function of order parameter(s): the protein free-energy landscape (FEL) [12]. As for glasses, the protein FEL has multiple local minima [14,30–32]. It evolves as a function of temperature, as often pictured as a funnel [12]. The protein FEL concept is essential to understanding the misfolding and aggregation of these heterogeneous polymers. A challenging problem is to define appropriate order parameters to describe the folded, misfolded, unfolded, and intrinsically disordered ensembles of protein structures. The nonfolded state of proteins is not necessarily random, nor does it resemble a Gaussian chain model, and must be characterized. For example, we showed recently that the α-synuclein monomer, a prototypical intrinsically disordered protein involved in Parkinson's disease, occurs in two distinct disordered states by using an FEL representation based on two order parameters [33]. Therefore, there is still a need to develop useful representations (order parameters) of the folding process based on fundamental laws. The present work aims to develop and test order parameters derived from graph theory [34,35] to contribute to the characterization of protein-disordered ensembles. The theoretical concepts developed in the present work will be tested for protein structures extracted from all-atom molecular dynamics (MD) simulations. Small- and medium-size proteins have been successfully folded using physical laws by all-atom MD simulations [22].

To associate a graph with a protein structure, we represent the amino acids by just a set of points (vertices of the protein graph) together with lines (edges of the graph) joining pairs of these points according to some rules (see Section 3). We select the C^{α} atom in the protein structure as a vertex representing each amino acid in the graph. For example, the model protein studied here, TRP-cage, will be represented by a graph of 20 vertices. The graph derived from atom positions in a 3D protein structure is hereby called a protein graph (PG). Unlike 3D models, where each C^{α} atom has a defined position in space and links between the C^{α} represent pseudo bonds, the relative positions of the vertices and of the shape and length of lines representing edges of a 2D representation of the PG have a priori no significance. The selection of a specific representation of a PG in 2D will depend on some additional descriptors which will serve to cluster the vertices in groups to reveal hidden information in structural or dynamical properties.

The applications of graphs and simplified three-dimensional networks to analyze protein structures and functions have been widely developed [36–39]. It was established early that graphs representing protein structures share the characteristics of small-world networks [40–43]. Critical amino acids [44], conserved amino-acids networks [45] in proteins, and signal propagation within the macromolecule were identified by using graphs [42]. Network models were applied to study protein flexibility [46,47], protein unfolding [48], and protein folding pathways [49–51]. The complex network of folding pathways can be represented by a graph where each vertex is a microstate or ensemble of microstates, and the edges represent the transitions between them [49].

Here, we show that two topological descriptors of a PG, the Kirchhoff index and the average shortest path between two vertices, can be used to cluster folded and disordered protein structures. Using a linear chain model and statistical physics, we demonstrate that the Kirchhoff index has the physical meaning of the inverse global force constant of the network, and we introduce the local force constant of a vertex, which can be related to Einstein's seminal model of crystal heat capacity. The free-energy models of the PG are thus defined based on Einstein's hypothesis and normal mode analysis. As a proof of concept, the present order parameters are used to analyze an all-atom molecular dynamics folding/unfolding trajectory and the ensemble of experimental NMR models of the fast-folder Trp-cage protein. To test the robustness of the findings, the analysis of the topological parameters was repeated for an all-atom molecular dynamics folding/unfolding trajectory of the fast-folder HP-36 protein.

This paper is organized as follows. In Section 2.1, we present the theory based on the analogy between a PG and a 1D chain with harmonic spring force constants. An analytical relation between the Kirchhoff index and the average shortest path length of a graph is derived for a fully unfolded protein. The definitions of the local, nonlocal, and global force constants of a PG and their relation with the Randić resistance of a graph, as well as the definition of the free energies of a PG, are given. In Section 2.2, numerical results are presented for the MD trajectory of Trp-cage. The results for HP-36 are presented in the Section S2 of the Supplementary Materials, as they are similar to those presented in the main text for Trp-cage protein. Technical details on the numerical construction of the PG and on the MD are reported in Section 3. This paper concludes with Section 4.

2. Results and Discussion

2.1. Theory

2.1.1. Mechanical Interpretation of a Simple, Connected, and Undirected Graph

Here, we introduce the topological equivalence between a simple, connected, and undirected graph and a *linear* chain of atoms with interatomic harmonic potentials.

First, we consider the Hamilonian of the linear chain with n atoms in the harmonic approximation:

$$H = \epsilon_0 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi_{ij}u_i u_j \tag{1}$$

where ϕ is the force constant matrix (Hessian), u_i and u_j are the displacements of the ith and jth atoms of the chain, and ϵ_0 is the ground-state energy. Because the chain is linear, the displacements are scalar numbers, which can take positive or negative values. By construction, the chain is connected: there is no atom not linked to another.

Newton's equations of motion are:

$$(\forall i): \quad f_i = m_i\ddot{u}_i \tag{2}$$

where m_i is the mass of atom i. The force f_i is conservative:

$$f_i = -\frac{\partial H}{\partial u_i} = -\sum_{j=1}^{n}\phi_{ij}u_j \tag{3}$$

From Equation (3), one finds:

$$(\forall i): \phi_{ii} = -\sum_{\substack{j=1\\ j\neq i}}^{n}\phi_{ij} \tag{4}$$

because for a rigid translation, i.e., $u_i = U\ (\forall i)$, all the forces must be zero.

Second, we consider a graph $G = (V, E)$ with vertex set $V$and edge set E. The number of vertices in V is n. We assume the graph to be undirected, simple, and connected. A pair of vertices v_i and v_j has edge weight w_{ij}, which is defined to be 0 if there is no edge

between v_i and v_j. Because the graph is assumed undirected, $w_{ij} = w_{ji}$. The adjacency matrix A has elements:

$$A_{ij} = w_{ij} \text{ if } i \neq j \text{ and } A_{ii} = 0 \tag{5}$$

The degree d_i of the ith vertex is the sum of the weights of all edges having v_i as one end:

$$d_i = \sum_{j=1, j\neq i}^{n} w_{ij} = \sum_{j=1, j\neq i}^{n} A_{ij} \tag{6}$$

The Laplacian L of the graph is defined as usual by:

$$L_{ij} = -A_{ij} \text{ if } i \neq j \text{ and } L_{ii} = d_i \tag{7}$$

There is a complete equivalence between L and ϕ if we interpret the nondiagonal elements of the force constant matrix as proportional to minus the weights of the edges of a graph connecting the atoms, i.e.,

$$\phi_{ij} = -A_{ij}c = -w_{ij}c \text{ if } i \neq j \tag{8}$$

where c is an arbitrary force constant, which ensures the proper physical dimension of ϕ. Therefore, we have:

$$\phi = cL \tag{9}$$

For a particular case of an unweighted graph for which A is a binary matrix (all nonzero weights are equal to 1), the associated linear chain has atoms connected with the same spring force constant c. For the particular case of the PG, A is binary. The vertices of the PG represent all the C^{α} atoms of the protein. An edge within the PG represents a contact between two C^{α} atoms, i.e., they are at a distance in the 3D protein structure shorter than a cut-off radius (see Section 3 for the construction of the PG). The PG Laplacian is thus equivalent to the force constant matrix of a linear chain where the atoms are connected (according to A) by the same harmonic spring strength c.

The spectral properties of L are equivalent (to a constant factor) to the spectral properties of the Hessian (Equation (9)). For a chain of atoms having the same mass, i.e., $(\forall i) : m_i = m$, the eigenvalues of L are related to the vibrational modes of the chain. Indeed, assuming a harmonic solution at the frequency ω for the displacement of the ith atom, i.e., $u_i(t) = u_i(0)e^{i\omega t}$, then $\ddot{u}_i(t) = (-\omega^2)u_i(0)e^{i\omega t}$. Using this in Equation (3), we find the following usual eigenvalue equation:

$$\omega^2 u_i(0) = \sum_{j=1}^{n} \left[\frac{\phi_{ij}}{m}\right] u_j(0) \equiv \sum_{j=1}^{n} B_{ij} u_j(0) \tag{10}$$

The diagonalization of B gives the frequencies ω_l and eigenvectors e_l of the vibrational modes of the chain:

$$B_{ij} = \sum_{l=1}^{n} \omega_l^2 e_l(i) e_l(j) \tag{11}$$

where $e_l(i)$ is the component of the eigenvector of the ith atom in the lth vibrational mode. We sort the modes by increasing frequency $\omega_{l+1} > \omega_l$. The mode 1 corresponds to a translation with $\omega_1 = 0$.

From Equations (9) and (10), the eigenvectors of L are identical to those of B, and thus the spectral decomposition of L is:

$$L_{ij} = \sum_{l=1}^{n} \lambda_l e_l(i) e_l(j) \tag{12}$$

with eigenvalues given by:

$$\lambda_l = \frac{\omega_l^2}{\Omega^2} \tag{13}$$

with $\Omega^2 = c/m$.

One has $\lambda_1 = 0$ as expected for a connected graph. The lowest nonzero eigenvalue λ_2 is named the algebraic connectivity, and the eigenvector e_2 is the Fiedler vector, which can be used to partition the graph. The corresponding vibrational mode 2 of the atomic chain is the mode with the largest wavelength, i.e., the components of its eigenvector are those that fluctuate less along the chain, i.e., less varying among the vertices in a graph.

2.1.2. Thermostatistical Interpretation of Topological Descriptors of a Simple, Connected, and Undirected Graph

We define three descriptors of the thermal fluctuations of a linear chain of atoms with harmonic interatomic potentials and show their equivalence with topological descriptors of a simple, connected, and undirected graph.

The most general solution of the equations of motions is a linear combination of the eigenmodes:

$$u_i(t) = \sum_{l=1}^{n} Q_l(t) \frac{e_l(i)}{\sqrt{m_i}} \tag{14}$$

where $Q_l(t)$ are the weights of the modes, the so-called normal coordinates. Using Equations (2) and (3), we have as usual for $\forall l \neq 1$:

$$\sqrt{\mu_l}\ddot{Q}_l(t) + \sqrt{\mu_l}\omega_l^2 Q_l(t) = 0 \tag{15}$$

with the solution $Q_l(t) = Q_l(0)cos(\omega_l t)$ and where μ_l is an arbitrary effective mass. For $l = 1$, $Q_l = constant$, and $\omega_1 = 0$. In the microcanonical ensemble (n, V, E) for which the energy E is constant, i.e., $Q_l(0)$, it is easy to show that

$$\langle Q_l(t) Q_{l'}(t) \rangle = \delta_{ll'} \frac{Q_l^2(0)}{2} \tag{16}$$

where $\langle \ldots \rangle = lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau \ldots$ is a time average. In the canonical ensemble (n, V, T), the energy of the microstates are $\frac{P_l^2}{2\mu_l} + \frac{\omega_l^2 Q_l^2}{2}$, with P_l being the momentum associated with mode l. Using the equipartition theorem for the chain in thermal equilibrium, we have $\forall l \neq 1$:

$$\left\langle Q_l^2 \right\rangle_T = \frac{k_B T}{\omega_l^2} \tag{17}$$

where $\langle \ldots \rangle_T$ is the average over all microstates in the canonical ensemble, T is the absolute temperature, and k_B is the Boltzmann constant. Because the normal coordinates are statistically independent in the harmonic approximation (as stated by Equation (15), there is no coupling between the modes), thus we have:

$$\langle Q_l Q_{l'} \rangle_T = \delta_{ll'} Q_l^2 \tag{18}$$

and finally, from Equation (14),

$$\left\langle u_i^2 \right\rangle_T = k_B T \left[\sum_{l=2}^{n} \frac{|e_l(i)|^2}{m_i \omega_l^2} \right] = \frac{k_B T}{\hat{k}_i} \tag{19}$$

where we have introduced an effective local force constant (local stiffness) $\hat{k}_i$. Equation (19) has a clear physical meaning: it represents the statistical fluctuations of the displacement of the ith atom in a local harmonic potential with a curvature $\hat{k}_i$. Summing the atom thermal

fluctuations of the entire chain defines an effective global force constant (global stiffness) $\hat{K}$:

$$\sum_{i=1}^{n}\left\langle u_i^2\right\rangle_T = k_B T\left[\sum_{l=2}^{n}\sum_{i=1}^{n}\frac{|e_l(i)|^2}{m_i\omega_l^2}\right] = \frac{k_B T}{\hat{K}} \tag{20}$$

Equation (20) represents the entire chain as fluctuating in a harmonic ground-state potential of curvature $\hat{K}$. A third effective nonlocal force constant (nonlocal stiffness) can be defined for the thermal fluctuations of a pair of atoms relative to each other:

$$\left\langle (u_i - u_j)^2\right\rangle_T = k_B T\left[\sum_{l=2}^{n}\frac{\left|\frac{e_l(i)}{\sqrt{m_i}} - \frac{e_l(j)}{\sqrt{m_j}}\right|^2}{\omega_l^2}\right] = \frac{k_B T}{\hat{K}_{ij}} \tag{21}$$

Equation (21) represents the fluctuations of each pair of atoms ij as if they were in a harmonic potential with curvature $\hat{K}_{ij}$. The global force constant can also be defined and measured for actual protein structures, where it is related to the dynamical transition of proteins [52].

The relation between the force constants and the topological descriptors of the corresponding graph is deduced from Equations (19) to (21) by using $m_i = m\ \forall i$, the normalization of eigenvectors $\sum_{i=1}^{n}|e_l(i)|^2 = 1$, and Equation (13) for the Laplacian eigenvalues:

$$\left\langle u_i^2\right\rangle_T = \frac{\gamma}{k_i} \tag{22}$$

$$\sum_{i=1}^{n}\left\langle u_i^2\right\rangle_T = \frac{\gamma}{K} \tag{23}$$

$$\left\langle \left(u_i^2 - u_j^2\right)^2\right\rangle_T = \frac{\gamma}{K_{ij}} \tag{24}$$

where $\gamma \equiv \frac{k_B T}{c}$ and with the dimensionless local (k_i), global (K), and nonlocal (K_{ij}) force constants of the graph defined by

$$\frac{1}{k_i} = \sum_{l=2}^{n}\frac{|e_l(i)|^2}{\lambda_l} \tag{25}$$

$$\frac{1}{K} = \sum_{l=2}^{n}\frac{1}{\lambda_l} \tag{26}$$

$$\frac{1}{K_{ij}} = \sum_{l=2}^{n}\frac{|e_l(i) - e_l(j)|^2}{\lambda_l} \tag{27}$$

The three force constants are related to each other through three sum rules obtained by using the normalization condition $\sum_{i=1}^{n}|e_l(i)|^2 = 1$:

$$\frac{1}{K} = \frac{1}{2n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{K_{ij}} \tag{28}$$

$$\frac{1}{K} = \sum_{i=1}^{n}\frac{1}{k_i} \tag{29}$$

$$\frac{1}{k_i} = \frac{1}{n}\left[\sum_{j=1}^{n}\frac{1}{K_{ij}} - \frac{1}{K}\right] \tag{30}$$

The formulation of the topological descriptors in the context of a chain in thermal equilibrium provides an interesting physical interpretation of the known topological descriptors of a graph as follows. The robustness of a graph is an important concept to measure the

quality of a physical network represented by a graph, as in for example a telecommunication system. One measure of the robustness is related to the effective (Randić) resistances of a graph [53]. For a pair of vertices (i, j), this quantity, noted Ω_{ij}, is defined through the Moore–Penrose inverse of the Laplacian, L^{-1}:

$$\Omega_{ij} = L_{ii}^{-1} + L_{jj}^{-1} - 2L_{ij}^{-1} \tag{31}$$

From the spectral decomposition of the Moore–Penrose inverse of the Laplacian, i.e., $L^{-1} = \sum_{i=1}^{n} e_l(i)e_l(j)/\lambda_l$, one immediately find that the nonlocal force constant K_{ij} and the Randić resistance Ω_{ij} are simply inversely related:

$$\Omega_{ij} = \frac{1}{K_{ij}} \tag{32}$$

Therefore, all the other force constants of a graph can be formulated in terms of Ω_{ij} because of the sum rules (Equations (28)–(30)). As shown in Randić's seminal paper, if a connected graph is associated with an electric network with resistances equal to the inverse of the weight w_{ij} between two nodes, Ω_{ij} represents the effective resistance between the nodes i and j if a voltage difference is applied between these two nodes. By analogy, if the connected graph is associated with a linear chain with interatomic force constants equal to the weights w_{ij} (normalized by a constant c), K_{ij} represents the effective force constant between the atoms i and j if a couple of forces are applied to this pair of nodes, as explicitly demonstrated in Section S4 of the Supplementary Materials. For a linear chain, the Randić resistance Ω_{ij} of an atom pair is also exactly its compliance C_{ij} [54] and equal to $1/K_{ij}$. For a three-dimensional elastic network, the compliance is a 3×3 tensor representing the elastic response of an atom pair to a couple of forces. A scalar compliance of a pair of nodes of a three-dimensional elastic network, similar to the Randić resistance or nonlocal force constant, can be computed by applying a couple of forces to the atoms in the direction of the vector that joins them [54]. As demonstrated in Section S4 of the Supplementary Materials, this scalar compliance [54] can be related analytically to the tensorial Randić resistance of the atom pair (equal to the inverse of the nonlocal force constant matrix).

Another usual measure of the robustness of a graph is the Kirchhoff index of a graph, Kf, defined by the sum of the eigenvalues of the Moore–Penrose inverse of the Laplacian and is simply the inverse of the global force constant:

$$Kf = \frac{1}{K} \tag{33}$$

The Kf is proportional to the average Randić resistance of the graph. Indeed, from Equation (28), $Kf = 1/K = (n-1)\langle\Omega_{ij}\rangle/4$. For a linear chain, $\langle\Omega_{ij}\rangle$ is also its average compliance.

In the present thermal statistical interpretation of the topological descriptors, each descriptor has the same physical meaning as the stiffness of a specific harmonic potential.

2.1.3. Relation between the Global Force Constant and the Average Shortest Path Length: Analytical Results

The path length between two vertices is defined as the sum of the weights of edges constituting the path. For a binary adjacency matrix, the length of a path between two vertices is the number of edges of the path connecting them. For a path $\alpha(i, j)$ between the vertices i and j, the length $l_{\alpha(i,j)}$ is

$$l_{\alpha(i,j)} = \sum_{\substack{pairs(r,s) \\ \in \alpha(i,j)}} w_{rs} = -\frac{1}{c} \sum_{\substack{pairs(r,s) \\ \in \alpha(i,j)}} \phi_{rs} \tag{34}$$

where (r, s) is an edge of the path $\alpha(i, j)$. The shortest path length between two vertices i and j is an important topological descriptor. We use the notation l_{ij}^0:

$$\min\left\{l_{\alpha(i,j)}\right\}_{\alpha(i,j)} = l_{ij}^0 \tag{35}$$

The average over all shortest path lengths between all pairs of vertices of the graph, $\left\langle l^0 \right\rangle$, is another well-studied topological descriptor of the graph robustness and is defined by

$$\left\langle l^0 \right\rangle = \frac{1}{n(n-1)} \sum_{r=1}^{n} \sum_{\substack{s=1 \\ s \neq r}}^{n} l_{rs}^0 \tag{36}$$

In Equation (36), the double sum is the so-called graph Wiener index.

As both the Kirchhoff index and the average shortest path length describe the robustness of a network in the literature, a natural question to ask is if and how they are related. An analytical answer can be found for a binary adjacency matrix A of a graph for which the first ($i = 1$) and last ($i = n$) vertices have degree 1, and all the others have degree 2. This graph corresponds to a linear chain with spring force constants between nearest-neighbor atoms only, and the PG is one of a completely unfolded protein (straight polypeptide). For this graph, the average over all shortest path lengths is simply:

$$\left\langle l^0 \right\rangle = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |j - i| \tag{37}$$

From Equation (37) and simple but tedious algebra (see Section S1 of the Supplementary Materials), one finds

$$\left\langle l^0 \right\rangle = \frac{n+1}{3} \tag{38}$$

The spectral properties of the Laplacian of this graph can be found analytically by analogy with the corresponding linear chain of atoms where $\phi_{ij} = -c$ between nearest-neighbor atoms only. The vibrational frequencies and eigenvectors of such a chain are well known [55] from which we find the eigenvalues of the Laplacian:

$$\lambda_l = \frac{\omega_l^2}{\Omega^2} = 4 \, \sin^2\left(\frac{(l-1)\pi}{2n}\right) \tag{39}$$

with $l = 1, 2, 3, \ldots n$. Using Equations (26) and (39), the global force constant for this graph is given by

$$\frac{1}{K} = \frac{1}{4} \sum_{l=1}^{n-1} \frac{1}{\sin^2\left(\frac{l\pi}{2n}\right)} = \frac{n^2 - 1}{6} \tag{40}$$

where the last equality is found as follows. Using $2sin^2\left(\frac{l\pi}{2n}\right) = 1 - cos\left(\frac{l\pi}{n}\right)$, we have

$$\frac{1}{K} = \frac{1}{2} \sum_{l=1}^{n-1} \frac{1}{1 - \cos\left(\frac{l\pi}{n}\right)} \tag{41}$$

We observe that $x_l \equiv \cos\left(\frac{l\pi}{n}\right)$ are the roots of the derivative of the Chebyshev polynomial of the first kind $T_n(x) = \cos(n\theta)$ with $x = \cos(\theta)$. The derivative of $T_n(x)$ is a polynomial of degree $n - 1 : P_{n-1}(x) \equiv \frac{dT_n(x)}{dx} \equiv T_n'(x)$ [56,57]. Using the chain rule for the derivatives, $P_{n-1}(x) = \frac{n \sin(n\theta)}{\sin(\theta)} = nU_{n-1}(x)$, where $U_n(x)$ is the Chebyshev polynomial of the second

kind [57]. Then, the sums in the right-hand-side of Equation (41) can be found by using the general rule

$$\frac{d\ln(P_{n-1}(x))}{dx} = \sum_{l=1}^{n-1} \frac{1}{x - x_l} \tag{42}$$

where x_l are the roots of P_{n-1}. From Equation (42), one has

$$\frac{1}{2}\sum_{l=1}^{n-1} \frac{1}{1-\cos\left(\frac{l\pi}{n}\right)} = \frac{1}{2}\left[\frac{P'(1)}{P(1)}\right] \tag{43}$$

Using l'Hôspital's rule, one easily finds $P(1) = n^2$, $P'(1) = \frac{n^4-n^2}{3}$. Inserting these expressions into Equation (43) leads to the announced result in Equation (40). Combining Equations (38) and (40), we derive the analytical relation between K and $\langle l^0 \rangle$ for this particular graph:

$$\frac{1}{K} = \left\langle l^0 \right\rangle \left(\frac{3}{2}\left\langle l^0 \right\rangle - 1\right) \tag{44}$$

which shows an inverse relation between the global force constant (1/Kirchhoff index) and the average of the shortest path length.

Equation (44) defines the lowest possible value in the diagram (K, l) of a PG. Indeed, for any PG, there is always a path from vertex i to vertex j with a length equal to $|i - j|$ because the corresponding C^α atoms form pseudo-bonds at a distance smaller than the cut-off radius defining a contact (see Section 3). Any contact between the C^α of amino acids not adjacent in the protein sequence in the 3D protein structure will add an edge that either does not change the length $|i - j|$ or reduces the length $|i - j|$. Therefore, the average shortest length given by Equation (37) is the largest possible *among all PGs having the same number of vertices (n amino acids)*. Consequently, the smallest value of K is given by Equation (40) and is proportional to n^{-2} for large n.

It is worth noting that the graph having the smallest average shortest path length is a complete graph where all vertices are related to all the others by a single edge for which $\langle l^0 \rangle = 1$. This graph is unrealistic for a macromolecule. The eigenvalues of the Laplacian of a complete graph are well known: $\lambda_1 = 0$ and $\lambda_i = n \; \forall i \neq 1$. Then, we have for such an extreme case:

$$\frac{1}{K} = 1 - \frac{1}{n} \tag{45}$$

For a large n, the K of a complete graph converges to its minimum mathematical value of 1.

2.1.4. Einstein's Model of a Graph

We build an Einstein model [55] of a simple, connected, and undirected graph by using the mechanical analogy described in Section 2.1.2. The Einstein model is applied here to protein structures recorded every picosecond. On this short timescale, each structure can be considered as fluctuating harmonically on a frozen energy landscape both in the folded and unfolded states. Experimentally, the fluctuations of the vibrational modes of a protein as a function of time can be measured by single-molecule Raman spectroscopy [58]. Following the Einstein model hypothesis, one assumes that each atom i (for $i = 1, \ldots n$) of the linear chain with identical masses and force constants (c) has a position fluctuating in a local harmonic potential with a local force constant $\hat{k}_i$ (Equation (19)) with a local frequency $\hat{\omega}_i^2 \equiv (\hat{k}_i/m)$. Unlike the original Einstein model, one assumes a frequency difference for each atom. The energy of each atom is given by

$$\hat{E}_i = \hat{E}_{oi} + (q + \frac{1}{2})\hbar\hat{\omega}_i \tag{46}$$

where q is the number of vibrational quanta, and $\hat{E}_{oi}$ is the potential energy minimum of atom i. Assuming thermal equilibrium at a temperature T in the (NVT) ensemble, the atom internal energy is [55]

$$\hat{U}_i = \hat{E}_{oi} + \frac{k_B T}{2} \hat{z}_i \coth\left(\frac{\hat{z}_i}{2}\right) \tag{47}$$

with $\hat{z}_i \equiv \frac{\hbar \hat{\omega}_i}{k_B T}$. The atom entropy is

$$\hat{S}_i = k_B \left[\hat{z}_i \left(\frac{1}{\exp(\hat{z}_i) - 1} \right) - \ln(1 - \exp(-\hat{z}_i)) \right] \tag{48}$$

The classical limit of enthalpy and entropy are found at high temperatures by expanding the exponential around $\hat{z}_i \ll 1$:

$$\lim_{\hat{z}_i \ll 1} \hat{U}_i = \hat{E}_{oi} + k_B T \tag{49}$$

$$\lim_{\hat{z}_i \ll 1} \hat{S}_i = k_B [1 - \ln(\hat{z}_i)] \tag{50}$$

The classical limit of the free energy is simply

$$\lim_{\hat{z}_i \ll 1} \hat{F}_i = \hat{E}_{oi} + k_B T \ln(\hat{z}_i) \tag{51}$$

The constant $\hat{k}_i$ (local frequency $\hat{\omega}_i$) is associated with a conformation of the chain, i.e., a PG built from the three-dimensional structure of the protein. Assuming some reference conformation of the chain with a free energy $\hat{F}_i(0)$ corresponding to a local force constant $\hat{k}_i(0)$ and frequency $\hat{\omega}_i(0)$, the free-energy difference $\Delta \hat{F}_i = \hat{F}_i - \hat{F}_i(0)$ in the classical limit is

$$\lim_{\hat{z}_i \ll 1} \Delta \hat{F}_i = \Delta \hat{E}_i + k_B T \ln\left(\frac{\hat{z}_i}{\hat{z}_i(0)} \right) \tag{52}$$

where the first term is the difference between potential energy minima

$$\Delta \hat{E}_i = \hat{E}_{oi} - \hat{E}_{oi}(0) \tag{53}$$

Equation (52) can be simplified

$$\lim_{\hat{z}_i \ll 1} \Delta \hat{F}_i = \Delta \hat{E}_i - \frac{k_B T}{2} \ln\left(\frac{\hat{k}_i(0)}{\hat{k}_i} \right) = \Delta \hat{E}_i - T \Delta \hat{S}_i \tag{54}$$

where $\Delta \hat{S}_i$ is the entropy variation, which is the only term depending on the local force constants.

We further make the hypothesis that each atom oscillates independently (as in the Einstein model). Therefore, for n amino acids, we have,

$$\lim_{\hat{z}_i \ll 1} \Delta \hat{F} = \sum_{i=1}^{n} \Delta \hat{E}_i - \frac{k_B T}{2} \sum_{i=1}^{n} \ln\left(\frac{\hat{k}_i(0)}{\hat{k}_i} \right) = \sum_{i=1}^{n} \Delta \hat{E}_i - \frac{k_B T}{2} \ln\left[\prod_{i=1}^{n} \frac{\hat{k}_i(0)}{\hat{k}_i} \right] \tag{55}$$

Thus, from Equation (55), we define the free energy ΔF_{local} of a PG by

$$\Delta F_{local} = \frac{1}{2} \left\{ \epsilon \sum_{i=1}^{n} (d_i - d_i(0)) - \ln\left[\prod_{i=1}^{n} \frac{k_i(0)}{k_i} \right] \right\} \tag{56}$$

where the dimensionless parameter $\epsilon < 0$ is unknown and controls the enthalpic (potential energy) contribution to the graph free energy. The reference conformation can be the graph examined in the previous section for which K is given by Equation (40) (completely un-

folded chain) or any other reference, for example, the PG built from the native experimental structure of the protein as in the numerical applications of Section 2.2.

Another formula for the free energy of a graph, $\Delta F_{nonlocal}$, can be built similarly:

$$\Delta F_{nonlocal} = \frac{1}{2}\left\{\epsilon \sum_{i=1}^{n}(d_i - d_i(0)) - \frac{1}{2}\ln\left[\prod_{i=1}^{n}\prod_{j=1}^{n}\frac{K_{ij}(0)}{K_{ij}}\right]\right\} \tag{57}$$

and finally a coarse-grained expression, ΔF_{global}, is defined:

$$\Delta F_{global} = \frac{1}{2}\left\{\epsilon \sum_{i=1}^{n}(d_i - d_i(0)) - \ln\left[\frac{K(0)}{K}\right]\right\} \tag{58}$$

An important property of the local and nonlocal dimensionless graph free energies of a PG is that the entropic contribution is dominated by the smallest force constants of the graph. For a PG, ΔF is of course $\Delta\hat{F}/k_BT$, but Equations (56)–(58) can be applied to any graph where the temperature has no meaning, such as, for example, a communication network.

An alternative to the Einstein model is the graph free energy built from the collective modes of the chain (Equation (15)). Each mode is associated with a collective frequency ω_l (Equation (11)). According to Equation (51), the collective graph free energy is thus defined as:

$$\Delta F_{collective} = \frac{1}{2}\left\{\epsilon \sum_{i=1}^{n}(d_i - d_i(0)) - \ln\left[\prod_{l=2}^{n}\frac{\lambda_l(0)}{\lambda_l}\right]\right\} \tag{59}$$

where $\lambda_l(0)$ are the eigenvalues of the Laplacian of the reference conformation (Equation (13)). For the PG of a completely unfolded chain, they are given by Equation (39). If we neglect the degree term, given an ensemble of graphs with the same number of vertices, the one that has the lowest free energy is the one with the smallest *product* of the eigenvalues of its Laplacian.

2.2. Topological Analysis of Folding/Unfolding MD Trajectory of Trp-Cage

2.2.1. Two-State Definition

We evaluate and investigate the application of graph force constants and free energies presented in Section 2.1 to folding/unfolding. As a proof of concept, we present here numerical applications for one MD trajectory of the mini-protein: Trp-cage [59]. The Trp-cage is a well-known toy model to study protein folding. This 20-amino-acid peptide is a C-terminal fragment of exendin-4. This construct folds within 4 microseconds in water at physiological pH and exhibits a tightly folded tertiary structure in solution. It consists of a short helix, a 3/10 helix, and a C-terminal poly-proline that packs against a Trp in the alpha helix [59]. The MD trajectory is 500 ns in duration and consists of snapshots calculated on every picosecond when the temperature is 380 K. More details of the MD trajectory are given in Section 3. The strategy is to build the PG of each snapshot and compute the parameters K and $< l_0 >$. In this way, we capture the topological information of the protein structures during the folding/unfolding dynamics.

A two-state (folded/unfolded) description of protein folding dynamics hides the complexity of unfolded and misfolded microstates [18]. To decipher the complexity behind these two macrostates, we need first to define them. Many usual global order parameters can be used to partition protein structures in folded and unfolded ensembles. Here, we use the fraction of the native contacts $\xi(t)$ computed for each snapshot at time t in the MD trajectories (see Section 3). At time $t = 0$ by construction, $\xi(0) = 1$ and fluctuates below 1 at 380 K (above the unfolding temperature) in the MD trajectory of Trp-cage, as shown in Figure 1. From this figure, we divide the snapshots into a folded state $\xi \geq 0.6$ and an unfolded state $\xi < 0.6$. Based on this criterion, we identify an interesting region 100 ns $< t <$ 400 ns where the behavior of descriptors obtained from graph representations can be studied. It contains a folding transition in the first half and an unfolding transition in the next half. It is important to note here that for a protein to function, apart from the

kinetic criterion of it folding to its native structure, it should also populate its native state for a significant fraction of time which can be mentioned as the *thermodynamic criterion*. Hence, even if we can observe more instances where the fraction of native contacts is above 0.6, the above-mentioned time interval becomes the most important since it pushes the structure to situations where the thermodynamic criterion is favored.

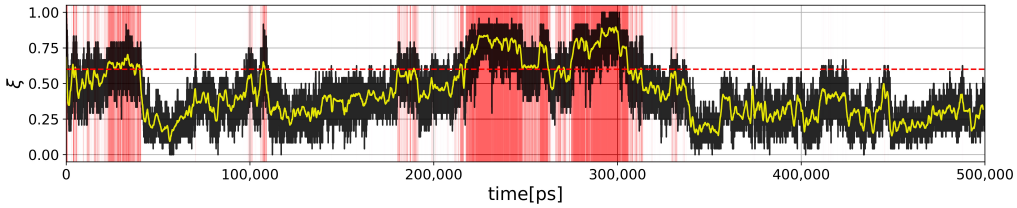


Figure 1. MD trajectory of Trp-cage at 380 K. Time t in red $(\forall t) : \xi(t) > 0.6$. The yellow curve is computed for a moving mean with a window size of 1 ns.

2.2.2. Force Constants and Shortest Path Length

First, we computed the global force constant K of the PG as a function of time, as shown in Figure 2.

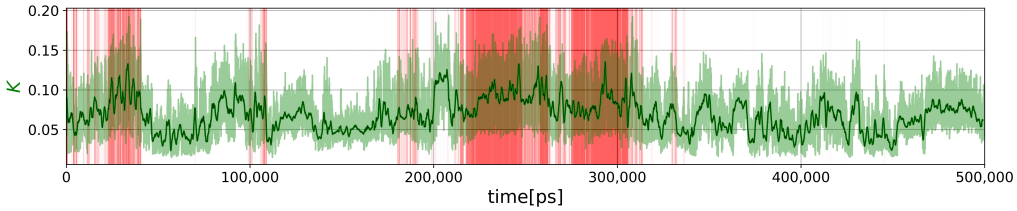


Figure 2. Evolution of the global force constant K for the MD trajectory shown in Figure 1. The bold green curve is computed for a moving mean with a window size of 1 ns.

A visual inspection of the curves shows that the global force constant is somehow related to the fraction of native contacts, but as K is more fluctuating than ξ, the Pearson correlation coefficient is not large: 0.4684. As intuitively expected, the time average value of the global force constant of folded structures ($\xi \geq 0.6$) $<K_{folded}> = 0.0882$ is significantly larger than its value for unfolded structures ($\xi < 0.6$), $<K_{unfolded}> = 0.0631$. According to Equation (40), the smallest possible value for Trp-cage is $K = 0.0150$, and the maximum hypothetical value is 1.0526 (Equation (45)). From Figure 2, the minimum and maximum values observed in the MD trajectory are $K = 0.0150$ and $K = 0.1940$, respectively. Although the folded protein is expected to be more rigid than an unfolded polymer chain, disordered or misfolded structures are also expected to be rigid. For example, in the time window 201 ns–208 ns, structures with $\xi \approx 0.4$–0.5 have $K \approx 0.12$ much larger than $<K_{folded}>$ and twice the value of $<K_{unfolded}>$. Thus, the descriptor K contains more information on the unfolded state than the global ξ order parameter. The two-dimensional probability density function (PDF) of the (ξ, K) values computed from the trajectory is represented in Figure 3a and revealed the existence of two unfolded substates at $(\xi \approx 0.3, K \approx 0.062)$ and $(\xi \approx 0.4, K \approx 0.04)$ and two folded substates at $(\xi \approx 0.8, K \approx 0.052)$ and $(\xi \approx 0.8, K \approx 0.100)$.

The time variation of $\langle l^0 \rangle$ is shown in Figure 4, where it is compared to K. The minimum and maximum values observed in the MD trajectory of $\langle l^0 \rangle$ are 2.0263 and 7.0, respectively. The maximum observed value corresponds to a completely unfolded chain, as predicted by Equation (38). The means of $\langle l^0 \rangle$ computed for folded ($\xi \geq 0.6$) and unfolded structures ($\xi < 0.6$) are 2.8465 and 3.2751, respectively. As expected, the paths in the folded PG are shorter on average. As for K, $\langle l^0 \rangle$ is not significantly correlated with the nativeness characterized by ξ (shown in Figure 1), as the Pearson correlation coefficient is only -0.3920. The variations of $\langle l^0 \rangle$ thus provide additional information on the different protein substates, as shown by the three local minima of the $(\xi, \langle l^0 \rangle)$ PDF in Figure 3b.

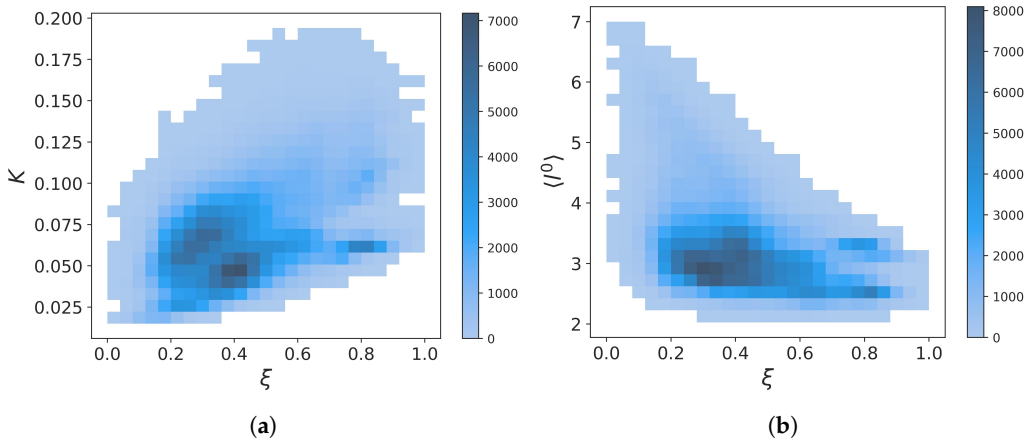


Figure 3. Panels (**a**,**b**) represent respectively the PDF of (ξ, K) values and $(\xi, \langle l^0 \rangle)$ computed from the trajectory shown in Figure 1.

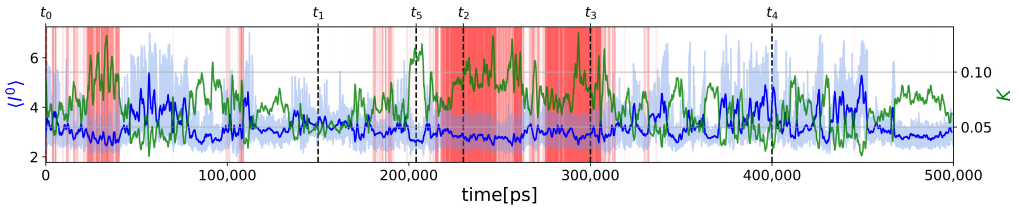


Figure 4. Comparison between the average shortest path length (blue) and global force constant (green) for the MD trajectory shown in Figure 1. The bold green curve is computed for a moving mean with a window size of 1 ns. Times t_0, t_1, t_2, t_3, t_4, and t_5 discussed in the text are indicated.

Figure 4 clearly shows that $\langle l^0 \rangle \propto 1/K$ as for a completely unfolded polymer chain (Equation (44)) (Pearson correlation coefficient is -0.8398). However, except for extremal values of $\langle l^0 \rangle$, a given average shortest path length corresponds to a range of values for K, as can be seen from Figure 5a. This can be explained because an intermediate protein size corresponds to a large number of possible conformations with different K values. For example, we show three selected structures $s1$, $s2$, and $s3$ (named by increasing K value) with the same value $\langle l^0 \rangle = 3$ in Figures 5c–e, respectively. They correspond to graphs with different robustness. In particular, the structures $s1$ and $s2$ have N-term and C-term which remain flexible, unlike the $s3$ structure. The nonuniqueness of the relation between K and $\langle l^0 \rangle$ explains why the PDF of the $(\xi, \langle l^0 \rangle)$ values computed from the trajectory (Figure 3b) shows only one substate in the unfolded region, whereas the PDF of (ξ, K) (Figure 3a) has two substates.

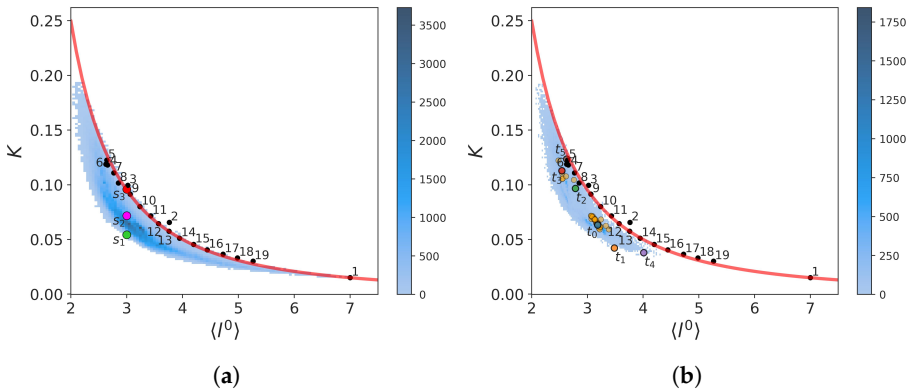


Figure 5. *Cont.*

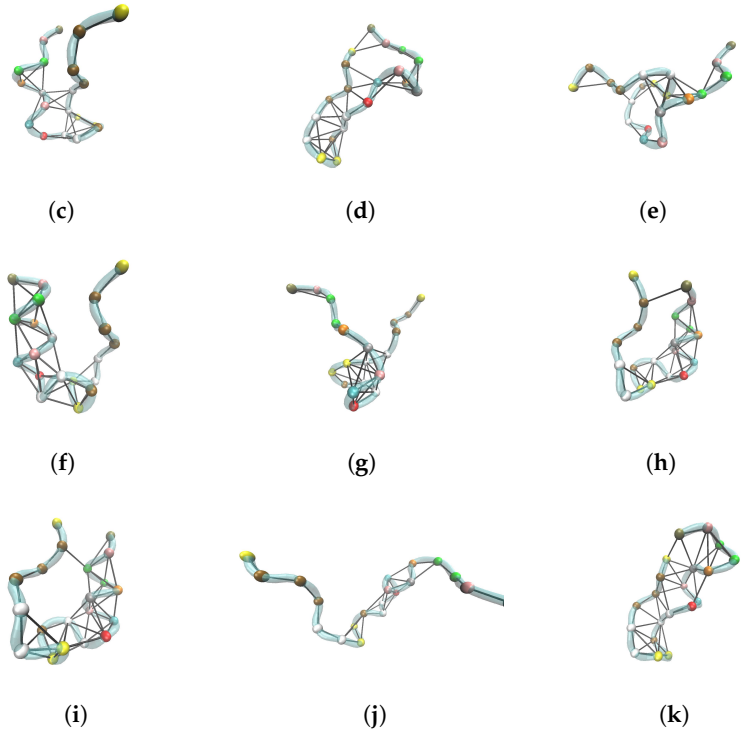


Figure 5. Relationship between K and l computed for the MD trajectory in Figure 1. Panel (**a**) PDF of $(K,\langle l^0\rangle)$ (blue dots) and pairs of values $(K,\langle l^0\rangle)$ for three selected snapshots named s_1 (green dot), s_2 (pink dot), and s_3 (red dot) with the same value of $\langle l^0\rangle$ as discussed in the text. Red line is the result of application of Equation (44). Black dots are the results of model chains with regular long distance spring force constants of different lengths named $(20, j = 1, 2, 3...)$ in the main text. Panel (**b**) PDF of $(K,\langle l^0\rangle)$ from all snapshots with $\xi > 0.6$ (blue). Red line and black dots are as in Panel (**a**). Orange dots are the $(K,\langle l^0\rangle)$ values of the experimental NMR models of Trp-cage (PDB ID: 1l1Y). Colors dots correspond to the values computed for the snapshots at times t_0 to t_5 indicated at Figure 4. Panels (**c**–**k**) are three-dimensional representations of the structures s_1, s_2, s_3 in Panel (**a**) and of the structures at times t_0, t_1, t_2, t_3, t_4, t_5, respectively. The spheres are the positions of the C^α atoms, and the tube represents the backbone. The black lines are the contacts considered to build the PG.

A striking property of the $(K, \langle l^0\rangle)$ plot is that the ensemble of points draws nearly continuous lower and upper limits. These maximum and minimum values must be degenerated for $\langle l^0\rangle = 7$ which corresponds to a completely unfolded chain of n amino acids according to Equation (38). Indeed, the value of K predicted by Equation (44), shown by the black dot with label 1 in Figure 5a, agrees with the MD result. Although Equation (44) was derived for an unfolded chain, we applied it to predict a value of K for each value of $\langle l^0\rangle$ observed in the MD trajectory. Surprisingly, it predicts nearly perfectly the upper limit of K for all the values of $\langle l^0\rangle$, as shown in Figure 5a. This unexpected result seems at first glance in contradiction with the fact that for a chain of length n, Equation (44) predicts the absolute possible *minimum* value of K as explained in Section 2.1.3.

This apparent contradiction is explained as follows. At each value of $\langle l^0\rangle$ of the PG of Trp-cage (with $n = 20$ amino acids), we can associate the PG of a completely unfolded shorter protein chain with $n < 20$ amino acids. For example, the value $\langle l^0\rangle = 3.66$ is the average shortest path length of the PG of an unfolded chain with $n = 10$ vertices according to Equation (38). This unfolded shorter chain can be built from the unfolded chain of $n = 20$ amino acids by eliminating every other amino acid and by connecting the remaining ones by first nearest-neighbor contacts. Therefore, a good approximation of this shorter chain ($n = 10$) by the PG of Trp-cage ($n = 20$) is a structure having contacts only between second nearest-neighbor C^α atoms in addition to contacts representing the peptide bonds between the first nearest-neighbor atoms. We name this model $(20, 2)$. The

value of K of $(20, 2)$ should be close to the minimum value of K for a chain with $n = 10$ amino acids predicted by Equation (44). The value of K of the $(20, 2)$ chain is shown by the black dot 2 in the $(K, \langle l^0 \rangle)$ plot and is indeed very close to the analytical prediction. We have built a series of models of completely unfolded chains $(20, j)$ with contacts only between the third ($j = 3$), fourth ($j = 4$), fifth ($j = 5$), etc., nearest neighbors represented by the black dots numbered, 3, 4, 5, ..., respectively. These points follow the predictions of Equation (44) perfectly confirming the reasoning. It can also be seen in Figure 5d that the structure $s3$, close to the upper limit for the value $\langle l^0 \rangle = 3$, corresponds approximately to a three-dimensional structure having third-neighbor contacts only. From a topological point of view, the PG of $s3$ is equivalent to the PG of the $(20, 3)$ structure having a value of K close to a chain with $n = 8$. This reasoning explains the predictions of Equation (44) but not why $s3$ is an upper limit for a chain of $n = 20$ for that value of $\langle l^0 \rangle = 3$. This can be understood qualitatively because a PG where each vertex is connected similarly as in $(20, j)$ structures corresponds to a PG where there is no vertex with a low degree, i.e., no weak local force constant which would significantly lower K as stated by Equation (30). On the opposite end, as we can see in Figure 5c, the s1 structure with a low K has end amino acids connected with only peptide bonds and thus has low local force constants. Although we can figure out the reason for the lower bound in the $(K, \langle l^0 \rangle)$, at the time of writing, we have not found an analytical formula to predict it.

It is worth comparing the $(K, \langle l^0 \rangle)$ plot extracted for the MD trajectory to the one computed from the 38 experimental NMR models of Trp-cage (PDB code: 1L2Y), as shown in Figure 5b. Surprisingly, the NMR data reveal two distinct groups separated by a gap along the axis $\langle l^0 \rangle$. The first and second groups are in the regions $2.5 < \langle l^0 \rangle < 2.75$ and $3.05 < \langle l^0 \rangle < 3.38$, respectively. The first group corresponds to more robust structures with $K \approx 0.10$–0.12, whereas the second group has softer structures with $K \approx 0.06$–0.07. The native NMR structure used as a reference in the present work (marked t_0 for which $\xi = 1$ and $K = 0.0632$) is in the second group. Averaging the values of K and of $\langle l^0 \rangle$ of the NMR models gives 0.075 and 3.06, respectively. This is in good agreement with the average values of these quantities computed for folded snapshots ($\xi \geq 0.6$ relative to the model chosen at t_0), which are, respectively, 0.0882 and 3.2751, as mentioned above. The existence of two substates in the native state of Trp-cage was discussed above and are visible in the PDF (K, ξ) (Figure 3a) with a major substate identified as the softest second experimental group and a minor state as the first hardest one. The PDF $(\langle l^0 \rangle, \xi)$ (Figure 3b) also shows the two groups but not with the correct weight as many unfolded structures populated the region of the softer group. Indeed, we recall that the average value $\langle l^0 \rangle$ computed from unfolded snapshots ($\xi < 0.6$ relative to the model chosen at t_0) is 2.8465.

The topological descriptors K and $\langle l^0 \rangle$ are global properties of the different protein microstates represented by PGs. A more detailed topological description of these microstates is the sequence of their local force constants k_i. To illustrate how these sequences vary in the two folding/unfolding transitions (defined here by crossing the limit $\xi = 0.6$ in Figure 1), we selected four representative snapshots in the MD trajectory at $t_1 = 150$ ns ($\xi = 0.3333, K = 0.0422$), $t_2 = 230$ ns ($\xi = 0.8333, K = 0.0967$), $t_3 = 300$ ns ($\xi = 0.9167, K = 0.1128$), and $t_4 = 400$ ns ($\xi = 0.3333, K = 0.0380$), as indicated in Figure 4. The snapshots at different times are shown in Figure 5f–k. As we can see in Figure 5b, both structures in the folded state at t_2 and t_3 are in the first experimental group (hardest structures). We also selected an unfolded structure at $t_5 = 204$ ns ($\xi = 0.4583$) corresponding to a snapshot with high rigidity, i.e., $K = 0.1236$.

A representation of sequences of k_i is shown in Figure 6. The sequence of k_i of the folded structures at times t_2 and t_3 are similar. All k_i at these times are larger or equal to the values of k_i at t_0 (reference native state). However, this is not sufficient to explain why the native structure has a global force constant nearly twice as small as these two. In fact, the very low k_i of PRO18, PRO19, and SER20 at t_0 decrease K significantly (and thus increase the entropy) more for the native structure than for the structures at t_2 and t_3. The sums of inverse k_i (Equation (30)) from ASN1 to PRO17 at times t_0, t_2, t_3 give a value of K equal

to 0.1267 [0.0632], 0.1518 [0.0967] and 0.1727 [0.1128] to compare with the values for the complete chain recalled in brackets. The unfolded structures at t_1 and t_4 have nearly all their k_i smaller than the ones at t_0, but their low K global force constant is mainly due to the very low values of k_i at the N-term and C-term regions. Indeed, calculations of the sums of inverse k_i of only ASN1, LEU2, TYR3, PRO17, PRO18, PRO19, and SER20 at times t_1 and t_4 give values of K equal to 0.0620 [0.0422] and 0.0547 [0.0380] which are relatively close to the K values of the complete chain recalled in brackets. Low k_i at times t_0 to t_4 are due to vertices with a low degree, as shown in the representations of the snapshots at different times in Figure 5. On the contrary, the structure at t_5 has no small k_i in the N-term and C-term regions, which explains the strong rigidity of this unfolded state. As can be seen in the representation in Figure 5k, the PG of this snapshot has no vertex with a low degree. In addition, this PG is close to the model structure $(20, 5)$ (Figure 3b) and indeed has long-distance contacts. The contributions of residues at the C-term region (PRO17, PRO18, PRO19, and SER20) to K explain the large difference of rigidity between the structures at t_0 and t_5. Indeed, the calculation of K for ASN1 to ARG16 for t_0 and t_5 give similar global force constants: 0.1475 [0.0632] and 0.1481 [0.1236], respectively (values for the complete chain are in brackets).

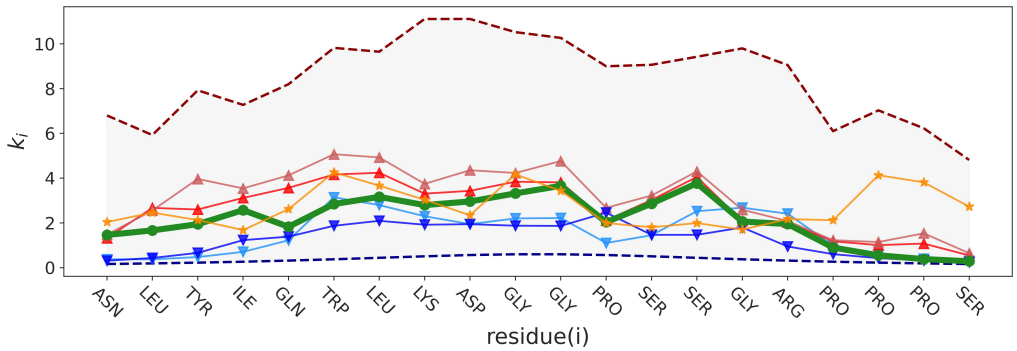


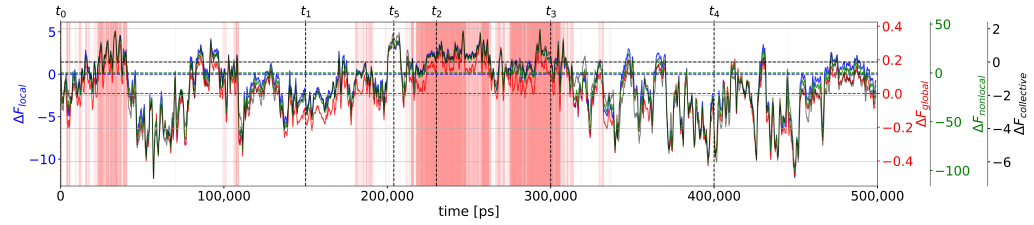
Figure 6. Distribution of the local force constants at times t_0 (bold green), t_1 (light blue), t_2 (red), t_3 (brown), t_4 (dark blue) and t_5 (orange) indicated in Figure 4 and discussed in the text. The gray area limited by dashed lines represents the range of values observed in the MD trajectory.

Metastable states competing with the native structure can be related to residual frustration. Frustration in condensed matter physics means that the system cannot simultaneously minimize the competing interactions between its different parts [60]. Proteins evolved in order to minimize frustration, which shapes a funnel free-energy landscape [60]. Although the study of the relations between residual frustration and topological descriptors (global, local, and nonlocal force constants) is far beyond the scope of the present study, we can make some qualitative observations. We computed the local residual frustration configurational index of each amino acid for the native structure of Trp-cage (PDB ID: 1L2Y, model 1) using the protein frustratometer 2 program [61]. The program predicts that amino acids in the N-term (from 1 to 9) are about 20% highly frustrated. This might be qualitatively related to the folded states at t_2 and t_3 for which contact between the N-term and C-term stabilizes these non-native folded configurations, as can be seen in Figure 5e,f, respectively. The difference between the sequence of k_i of these two configurations with the one of the native structure is also larger in the N-term, as shown in Figure 6.

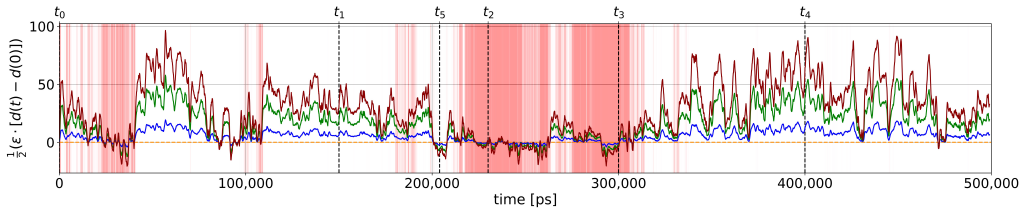
2.2.3. Calculation of Free Energies Using the Einstein Model

In the graph free-energy formula derived from the Einstein model, the force constant term is purely entropic (Equation (54)). This contribution is parameter free. The enthalpic part (first term in Equation (56)) depends on an energy scale defined by the single parameter ϵ. As we do not have information on ϵ, we treat it here as a variable. First, we compare the entropic contribution (i.e., for $\epsilon = 0$) of the local (Equation (56)), nonlocal (Equation (57)), global (Equation (58)), and collective (Equation (59)) models of the graph free energy in

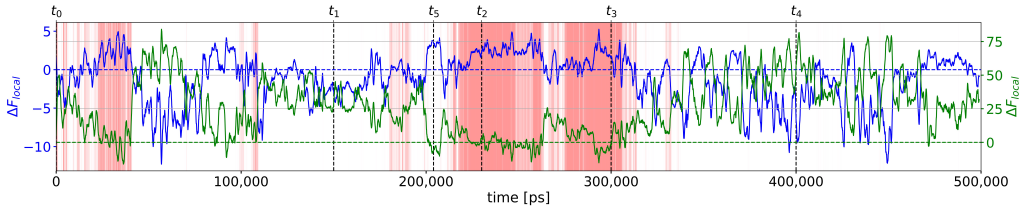
Figure 7a. The local, nonlocal, and collective models agree remarkably with each other with only a change in scale. The coarse-grained global model has the smallest scale and is also very similar to the other models, with a high Pearson correlation coefficient of 0.97 compared to the local model for example. In all models, the entropy change is positive in the folded parts of the MD trajectory, as expected, since the folding reduces possible structural fluctuations. In unfolded parts of the trajectory, the entropy change is mostly negative as expected. There are a few exceptions, for example, times around t_5. The time parts with positive entropy indicate unfolded very rigid structures. The calculation of the entropic term is thus a means to identify misfolded structures in the trajectory. In Figure 7b, we represent an enthalpic term for different values of ϵ. This term is positive in the unfolded parts of the trajectory, as expected, as the unfolded structures have vertices with a lower degree (fewer contacts), with the structures around t_5 being an exception. The enthalpic term is small in the folded parts, which indicates that folded structures are on average as connected as the reference structure at t_0. The enthalpic term is only roughly anticorrelated with the entropic term (the Pearson correlation coefficient between the two terms for the local model is -0.31). We observe structures with a positive entropy (rigid) but with fewer contacts than in the reference folded structure at t_0 (such as, for example, in the region 80–90 ns). The examination of the enthalpic and entropic parts of the free-energy models permits one to characterize the different rigid misfolded structures. The addition of the two terms is represented for a value of $\epsilon = -5$ in Figure 7c. With this value of ϵ, the structures in time ranges where the folded structure is stable on average (marked red in Figure 1) have zero or negative free energies. The metastable rigid structure at time t_5 also has negative free energy, whereas most of the unfolded structures have positive free energy.



(**a**)



(**b**)



(**c**)

Figure 7. Free-energy graph calculations for the trajectory of Figure 1. (**a**) Local (blue), nonlocal (green), global (red), and collective (black) free energy with $\epsilon = 0$. Horizontal dashed lines indicate the zero baselines of the free energies with the corresponding colors. (**b**) Enthalpy term of the free-energy graph

with $\epsilon = -1$ (blue), $\epsilon = -3$ (green), and $\epsilon = -5$ (dark red). Horizontal dashed line indicates the zero baseline. $d(t) \equiv \sum_i d_i(t)$ and $d(0) \equiv \sum_i d_i(0)$. (**c**) Local free energy with $\epsilon = 0$ (blue) and $\epsilon = -5$ (green). Folded regions are indicated by red vertical lines as in Figure 1. Horizontal dashed lines indicate the zero baselines of the free energies with the corresponding colors.

3. Materials and Methods

3.1. Contacts and Protein Graph (PG)

Although a PG might be built from the all-atom protein structure, we focus here on a coarse-grained representation of the protein main chain, which only has proven to be useful in describing protein folding [18]. Namely, we represent the protein's three-dimensional structure by the sole positions of its C^α atoms. Each vertex of the PG thus represents the C^α atom of an amino acid, and the vertices are ordered as in the amino-acid sequences from $i = 1$ to n, where n is the number of amino acids. An edge between two vertices is drawn if the distance between the two C^α atoms is a contact. A contact is defined as usual for two C^α atoms belonging to nonadjacent amino acids in the protein sequence and which are at a distance in the three-dimensional protein structure below a cut-off radius $R = 0.6$ nm. This typical value includes the peak of the first nearest neighbors of the C^α atoms in folded protein structures. In the present work, a PG is always connected because we add an edge between two C^α atoms, which are nearest neighbors in the amino-acid sequence. These additional edges represent the peptide bonds. The PG is simple, i.e., there is no edge connecting a single vertex (graph loop) or multiple edges between two vertices. We do not make any distinction between the different edges and assume their weight is equal to 1. PG with no contact corresponds to the straight unfolded chain examined in Section 2.1.3 and has the minimum number of edges, i.e., $n - 1$. We define also as usual the native contacts as the contacts present in the experimental folded structure (PDB ID: 1L2Y, model 1). Say $nc_{native}(t)$, the number of native contacts in the structure of the snapshot at time t in the MD trajectory, then we define the fraction of native contacts $\xi(t)$ as follows:

$$\xi(t) = \frac{nc_{native}(t)}{nc^*_{native}}. \tag{60}$$

where nc^*_{native} corresponds to the number of contacts in the experimental native structure. In the MD trajectories studied here, it is also equal to $nc_{native}(t = 0)$ because the initial structure is the experimental one (see Section 3.2). We consider the fraction of native contacts at time t to obtain a measure of the structure's nativeness as a function of time (see text Section 2.2).

3.2. Molecular Dynamics Trajectories

The MD trajectory of Trp-cage was generated in a previous unrelated work using an all-atom force field in explicit water at 380 K (above the folding transition temperature) [62]. This MD trajectory was chosen as it clearly shows folding/unfolding events. The MD trajectory is 500 ns in duration and consists of snapshots stored every picosecond (500,000 structures/protein). The initial structure at time $t = 0$ in the MD trajectory is an experimental native structure (PDB ID: 1L2Y, model 1). More details on the MD trajectory can be found in the original paper [62].

3.3. Statistics

All statistical calculations (averages, probability densities, Pearson correlation coefficients) were computed from raw data (not from moving average data). The number of bins for both axes in the PDF calculations is 25 for Figure 3 and 100 for Figure 5. The average shortest path length between two vertices was computed with the *average_shortest_path_length* function of the NetworkX Python library [63].

4. Conclusions

We emphasize here the main conclusions of the present study and its further extensions. We show that the (K, ξ), $(\langle l^0 \rangle, \xi)$ and $(K, \langle l^0 \rangle)$ plots are relevant representations to characterize the diversity of unfolded and folded microstates. The study of k_i and K as functions of time in an MD trajectory permits the detection of misfolded rigid structures among unfolded conformations. The application of these topological concepts is particularly relevant to characterize the conformations of intrinsically disordered proteins, e.g., α-synuclein [33,64] and will be investigated elsewhere. Topological descriptors and graph free-energy models introduced here permit the characterization of a single simulated or experimental structure at a time. The entropic part is only governed by the force constants computed for the PG associated with a single structure. However, the PG rigidity does not represent of course the full mechanical response of proteins. As a PG is equivalent to a linear chain, it misses the dihedral/rotational degrees of freedom of proteins, which contribute to folding/unfolding transitions [48]. The effect of solvent [65] is also implicit in PG analysis. These degrees of freedom are related to transitions between PGs. Moreover, in protein folding, the stability of structures within a time window must also be considered, i.e., an ensemble of the PGs. An extension of the present theory will include a study of these PG ensembles, their transitions, and the fluctuations of topological descriptors in the folded and unfolded states.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/molecules28186659/s1, Section S1: Mathematical details on the demonstration of Equation (38); Section S2: Topological analysis of folding/unfoldong MD trajectory of HP-36; Section S3: Supplementary figures for HP-36 (PDB ID: 1VII); Section S4: Generalized Randić theorem and relation with compliance

Author Contributions: Conceptualization, P.S.; methodology, S.T. and P.S.; software, S.T.; validation, S.T. and P.S.; formal analysis, S.T., C.L. and P.S.; investigation, S.T. and P.S.; resources, P.S.; data curation, S.T.; writing—original draft preparation, P.S.; writing—review and editing, S.T., C.L., A.G., A.N., G.G.M. and P.S.; visualization, S.T.; supervision, P.S.; project administration, P.S.; funding acquisition, P.S. All authors have read and agreed to the published version of the manuscript.

Funding: The work is part of the project NANO-NEURO-MED funded by the EIPHI Graduate School (Contract ANR-17-EURE-0002), the Conseil Régional de Bourgogne-Franche-Comté and the European Union through the PO FEDER-FSE Bourgogne 2021/2027 program. G.G.M. acknowledges support from the National Institutes of Health (GM-14312).

Data Availability Statement: Data files of the MD trajectory and of the numerical results presented are available on a simple request to the authors.

Acknowledgments: The simulations were performed using HPC resources from DSI-CCuB (Université de Bourgogne). The authors thank Luka Maisuradze for careful English editing of the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MD	Molecular Dynamics
PDF	Probability Density Function

References

1. Dyson, H.J.; Wright, P.E. Unfolded Proteins and Protein Folding Studied by NMR. *Chem. Rev.* **2004**, *104*, 3607–3622 [CrossRef]
2. Schuler, B.; Eaton, W.A. Protein folding studied by single-molecule FRET. *Curr. Opin. Struct. Biol.* **2008**, *18*, 16–26. [CrossRef]
3. Lai, J.K.; Kubelka, G.S.; Kubelka, J. Sequence, structure, and cooperativity in folding of elementary protein structural motifs. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 9890–9895. [CrossRef]
4. Sukenik, S.; Pogorelov, T.V.; Gruebele, M. Can Local Probes Go Global? A Joint Experiment–Simulation Analysis of λ6–85 Folding. *J. Phys. Chem. Lett.* **2016**, *7*, 1960–1965. [CrossRef]
5. Muñoz, V.; Campos, L.A.; Sadqi, M. Limited cooperativity in protein folding. *Curr. Opin. Struct. Biol.* **2016**, *36*, 58–66. [CrossRef]

6. Muñoz, V.; Cerminara, M. When fast is better: Protein folding fundamentals and mechanisms from ultrafast approaches. *Biochem. J.* **2016**, *473*, 2545–2559. [CrossRef]
7. Zhuravleva, A.; Korzhnev, D.M. Protein folding by NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **2017**, *100*, 52–77. [CrossRef]
8. Anfinsen, C.B. Principles that Govern the Folding of Protein Chains. *Science* **1973**, *181*, 223–230. [CrossRef]
9. Anfinsen, C.B.; Scheraga, H.A. Experimental and Theoretical Aspects of Protein Folding. In *Advances in Protein Chemistry*; Anfinsen, C.B., Edsall, J.T., Richards, F.M., Eds.; Academic Press: Cambridge, MA, USA, 1975; Volume 29, pp. 205–300. [CrossRef]
10. Dill, K.A.; Bromberg, S.; Yue, K.; Chan, H.S.; Ftebig, K.M.; Yee, D.P.; Thomas, P.D. Principles of protein folding—A perspective from simple exact models. *Protein Sci.* **1995**, *4*, 561–602. [CrossRef]
11. Muñoz, V.; Eaton, W.A. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 11311–11316. [CrossRef]
12. Onuchic, J.N.; Wolynes, P.G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75. [CrossRef]
13. Scheraga, H.A. From helix–coil transitions to protein folding. *Biopolymers* **2008**, *89*, 479–485. [CrossRef]
14. Senet, P.; Maisuradze, G.G.; Foulie, C.; Delarue, P.; Scheraga, H.A. How main-chains of proteins explore the free energy landscape in native states. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 19708–19713. [CrossRef]
15. Dill, K.A.; Ozkan, S.B., Shell, M.S.; Weikl, T.R. The Protein Folding Problem. *Annu. Rev. Biophys.* **2008**, *37*, 289–316. [CrossRef]
16. Henry, E.R.; Best, R.B.; Eaton, W.A. Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 17880–17885. [CrossRef]
17. Grassein, P.; Delarue, P.; Scheraga, H.A.; Maisuradze, G.G.; Senet, P. Statistical Model to Decipher Protein Folding/Unfolding at a Local Scale. *J. Phys. Chem. B* **2018**, *122*, 3540–3549. [CrossRef]
18. Grassein, P.; Delarue, P.; Nicolaï, A.; Neiers, F.; Scheraga, H.A.; Maisuradze, G.G.; Senet, P. Curvature and Torsion of Protein Main Chain as Local Order Parameters of Protein Unfolding. *J. Phys. Chem. B* **2020**, *124*, 4391–4398. [CrossRef]
19. Vila, J.A. Rethinking the protein folding problem from a new perspective. *Eur. Biophys. J.* **2023**, *52*, 189–193. [CrossRef]
20. Kussell, E.; Shimada, J.; Shakhnovich, E.I. A structure-based method for derivation of all-atom potentials for protein folding. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 5343–5348. [CrossRef]
21. Vila, J.A.; Ripoll, D.R.; Scheraga, H.A. Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 14812–14816. [CrossRef]
22. Lindorff-Larsen, K.; Piana, S.; Dror, R.O.; Shaw, D.E. How Fast-Folding Proteins Fold. *Science* **2011**, *334*, 517–520. [CrossRef]
23. Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C. Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent. *J. Am. Chem. Soc.* **2014**, *136*, 13959–13962. [CrossRef]
24. Shao, Q.; Zhu, W. How Well Can Implicit Solvent Simulations Explore Folding Pathways? A Quantitative Analysis of α-Helix Bundle Proteins. *J. Chem. Theory Comput.* **2017**, *13*, 6177–6190. [CrossRef] [PubMed]
25. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
26. Buel, G.R.; Walters, K.J. Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* **2022**, *29*, 1–2. [CrossRef]
27. Marx, V. Method of the Year: Protein structure prediction. *Nat. Methods* **2022**, *19*, 5–10. [CrossRef]
28. Callaway, E. AlphaFold's new rival? Meta AI predicts shape of 600 million proteins. *Nature* **2022**, *611*, 211–212. [CrossRef]
29. Jones, D.T.; Thornton, J.M. The impact of AlphaFold2 one year on. *Nat. Methods* **2022**, *19*, 15–20. [CrossRef]
30. Austin, R.H.; Beeson, K.W.; Eisenstein, L.; Frauenfelder, H.; Gunsalus, I.C. Dynamics of ligand binding to myoglobin. *Biochemistry* **1975**, *14*, 5355–5373. [CrossRef]
31. Frauenfelder, H.; Sligar, S.G.; Wolynes, P.G. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254*, 1598–1603. [CrossRef]
32. Nicolaï, A.; Petiot, N.; Grassein, P.; Delarue, P.; Neiers, F.; Senet, P. Free-Energy Landscape Analysis of Protein-Ligand Binding: The Case of Human Glutathione Transferase A1. *Appl. Sci.* **2022**, *12*, 8196. [CrossRef]
33. Guzzo, A.; Delarue, P.; Rojas, A.; Nicolaï, A.; Maisuradze, G.G.; Senet, P. Missense Mutations Modify the Conformational Ensemble of the α-Synuclein Monomer Which Exhibits a Two-Phase Characteristic. *Front. Mol. Biosci.* **2021**, *8*, 1104. [CrossRef] [PubMed]
34. Albert, R.; Barabási, A.L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **2002**, *74*, 47–97. [CrossRef]
35. Das, R.; Soylu, M. A key review on graph data science: The power of graphs in scientific studies. *Chemom. Intell. Lab. Syst.* **2023**, *240*, 104896. [CrossRef]
36. Vishveshwara, S.; Brinda, K.V.; Kannan, N. Protein structure: Insights from graph theory. *J. Theor. Comput. Chem.* **2002**, *1*, 187–211. [CrossRef]
37. Giuliani, A.; Krishnan, A.; Zbilut, J.P.; Tomita, M. Proteins As Networks: Usefulness of Graph Theory in Protein Science. *Curr. Protein Pept. Sci.* **2008**, *9*, 28–38. [CrossRef]
38. Atilgan, C.; Okan, O.B.; Atilgan, A.R. Network-Based Models as Tools Hinting at Nonevident Protein Functionality. *Annu. Rev. Biophys.* **2012**, *41*, 205–225. [CrossRef]
39. Kantelis, K.F.; Asteriou, V.; Papadimitriou-Tsantarliotou, A.; Petrou, A.; Angelis, L.; Nicopolitidis, P.; Papadimitriou, G.; Vizirianakis, I.S. Graph theory-based simulation tools for protein structure networks. *Simul. Model. Pract. Theory* **2022**, *121*, 102640. [CrossRef]
40. Scala, A.; Amaral, L.A.N.; Barthélémy, M. Small-world networks and the conformation space of a short lattice polymer chain. *Europhys. Lett.* **2001**, *55*, 594–600. [CrossRef]

41. Vendruscolo, M.; Dokholyan, N.V.; Paci, E.; Karplus, M. Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E* **2002**, *65*, 061910. [CrossRef]
42. Atilgan, A.R.; Akan, P.; Baysal, C. Small-World Communication of Residues and Significance for Protein Dynamics. *Biophys. J.* **2004**, *86*, 85–91. [CrossRef]
43. Bagler, G.; Sinha, S. Network properties of protein structures. *Phys. A Stat. Mech. Its Appl.* **2005**, *346*, 27–33. [CrossRef]
44. Srivastava, D.; Bagler, G.; Kumar, V. Graph Signal Processing on protein residue networks helps in studying its biophysical properties. *Phys. A Stat. Mech. Its Appl.* **2023**, *615*, 128603. [CrossRef]
45. Higman, V.A.; Greene, L.H. Elucidation of conserved long-range interaction networks in proteins and their significance in determining protein topology. *Phys. A Stat. Mech. Its Appl.* **2006**, *368*, 595–606. [CrossRef]
46. Jacobs, D.J.; Rader, A.; Kuhn, L.A.; Thorpe, M. Protein flexibility predictions using graph theory. *Proteins Struct. Funct. Bioinform.* **2001**, *44*, 150–165. [CrossRef] [PubMed]
47. Atilgan, A.R.; Durell, S.R.; Jernigan, R.L.; Demirel, M.C.; Keskin, O.; Bahar, I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **2001**, *80*, 505–515. [CrossRef]
48. Rader, A.J.; Hespenheide, B.M.; Kuhn, L.A.; Thorpe, M.F. Protein unfolding: Rigidity lost. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 3540–3545. [CrossRef] [PubMed]
49. Rao, F.; Caflisch, A. The Protein Folding Network. *J. Mol. Biol.* **2004**, *342*, 299–306. [CrossRef]
50. Yin, Y.; Maisuradze, G.G.; Liwo, A.; Scheraga, H.A. Hidden Protein Folding Pathways in Free-Energy Landscapes Uncovered by Network Analysis. *J. Chem. Theory Comput.* **2012**, *8*, 1176–1189. [CrossRef]
51. Jacobs, W.M.; Shakhnovich, E.I. Structure-Based Prediction of Protein-Folding Transition Paths. *Biophys. J.* **2016**, *111*, 925–936. [CrossRef]
52. Zaccai, G. How Soft Is a Protein? A Protein Dynamics Force Constant Measured by Neutron Scattering. *Science* **2000**, *288*, 1604–1607. [CrossRef] [PubMed]
53. Klein, D.J.; Randić, M. Resistance distance. *J. Math. Chem.* **1993**, *12*, 81–95. [CrossRef]
54. Scaramozzino, D.; Khade, P.M.; Jernigan, R.L.; Lacidogna, G.; Carpinteri, A. Structural compliance: A new metric for protein flexibility. *Proteins Struct. Funct. Bioinform.* **2020**, *88*, 1482–1492. [CrossRef] [PubMed]
55. Hill, T. *An Introduction to Statistical Thermodynamics*; Dover Publications, Inc.: New York, NY, USA, 1986.
56. Sum of the Reciprocal of Sine Squared. Published: Mathematics Stack Exchange. Available online: https://math.stackexchange.com/q/122933 (accessed on 5 August 2023).
57. Handscomb, D.C.; Mason, J.C. *Chebyshev Polynomials*; Chapman and Hall/CRC: New York, NY, USA, 2002. [CrossRef]
58. Dai, X.; Fu, W.; Chi, H.; Mesias, V.S.D.; Zhu, H.; Leung, C.W.; Liu, W.; Huang, J. Optical tweezers-controlled hotspot for sensitive and reproducible surface-enhanced Raman spectroscopy characterization of native protein structures. *Nat. Commun.* **2021**, *12*, 1292. [CrossRef] [PubMed]
59. Neidigh, J.W.; Fesinmeyer, R.M.; Andersen, N.H. Designing a 20-residue protein. *Nat. Struct. Biol.* **2002**, *9*, 425–430. [CrossRef] [PubMed]
60. Ferreiro, D.U.; Komives, E.A.; Wolynes, P.G. Frustration in biomolecules. *Q. Rev. Biophys.* **2014**, *47*, 285–363. [CrossRef] [PubMed]
61. Parra, R.G.; Schafer, N.P.; Radusky, L.G.; Tsai, M.Y.; Guzovsky, A.B.; Wolynes, P.G.; Ferreiro, D.U. Protein Frustratometer 2: A tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Res.* **2016**, *44*, W356–W360. [CrossRef] [PubMed]
62. Nicolaï, A.; Delarue, P.; Senet, P. Intrinsic Localized Modes in Proteins. *Sci. Rep.* **2015**, *5*, 18128. [CrossRef]
63. Hagberg, A.A.; Schult, D.A.; Swart, P.J. Exploring Network Structure, Dynamics, and Function using NetworkX. In Proceedings of the 7th Python in Science Conference, Pasadena, CA, USA, 21 August 2008; Varoquaux, G., Vaught, T., Millman, J., Eds.; Los Alamos National Lab. (LANL): Los Alamos, NM, USA, 2008; pp. 11–15.
64. Guzzo, A.; Delarue, P.; Rojas, A.; Nicolaï, A.; Maisuradze, G.G.; Senet, P. Wild-Type α-Synuclein and Variants Occur in Different Disordered Dimers and Pre-Fibrillar Conformations in Early Stage of Aggregation. *Front. Mol. Biosci.* **2022**, *9*, 910104. [CrossRef]
65. Frauenfelder, H.; Fenimore, P.W.; Chen, G.; McMahon, B.H. Protein folding is slaved to solvent motions. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 15469–15472. [CrossRef]