Key interaction patterns in proteins revealed by cluster expansion of the partition function

Matteo Tajana[1], Antonio Trovato[2] and Guido Tiana[3†]

[1]Department of Physics, Università degli Studi di Milano, via Celoria 16, 20133 Milano, Italy.

[2]Department of Physics and Astronomy "G. Galilei", Università degli Studi di Padova and INFN, via Marzolo 8, 35121 Padova, Italy.

[3]Department of Physics and Center for Complexity and Biosystems, Università degli Studi di Milano and INFN, via Celoria 16, 20133 Milano, Italy.

[†]Corresponding author, email: guido.tiana@unimi.it

Abstract

The native conformation of structured proteins is stabilized by a complex network of interactions. We analyzed the elementary patterns that constitute such network and ranked them according to their importance in shaping protein sequence design. To achieve this goal, we employed a cluster expansion of the partition function in the space of sequences and evaluated numerically the statistical importance of each cluster. An important feature of this procedure is that it is applied to a dense, finite system. We found that patterns that contribute most to the partition function are cycles with even numbers of nodes, while cliques are typically detrimental. Each cluster also gives a contribute to the sequence entropy, which is a measure of the evolutionary designability of a fold. We compared the entropies associated with different interaction patterns to their abundances in the native structures of real proteins.

Keywords: Protein designability, interaction network, elementary patterns.

1 Introduction

The native state of proteins is stabilized by a complex set of heterogeneous interactions among their amino acids. Approximating these interactions as two–body and short–ranged, one can simplify the description of the native state in terms of a network whose nodes are the amino acids and whose edges are their mutual interactions (for a review, see ref. [1]).

The analysis of the interaction networks between amino acids of single–domain proteins is not trivial due to the fact that they are usually small, and thus the features that characterize quantitatively the associated networks vary on limited ranges. Nonetheless, such networks were found to be usually small–world [2] and display clusters of interactions [3], which are often related to the biological properties of the protein. Even small clusters of contacts can be highly relevant for protein kinetics [4], by nucleating the folding process [5] and for the thermodynamics stabilization of the native state [6].

To produce well–folding proteins, evolution was supposed to insert in the nodes of the network the twenty types of amino acids with the purpose of obtaining a particularly low energy in the native conformation [7], trying to minimize the frustration of the system [8]. In this respect, evolution can be regarded as a stochastic process in the space of sequences controlled by the energy of the native conformation, being the rest of the conformational spectrum sequence–independent [9]. Assuming that it reached a stationary state [10], evolution can then be described by a canonical ensemble associated to the space of sequences, the temperature T_s being a parameter that sets the evolutionary bias towards sequences having a low energy in the native states.

It was suggested that some protein conformations are more 'designable' than others, being able to act as native state of more sequences [11]. Several properties of the native conformation have been associated with an increased designability, including conformational symmetries [12–14], abundance of local interactions [15], of returning loops [16], of loops of specified size [17], or a large local entropy [18].

The goal of the present work is to study if there are specific patterns in the interaction network that can favour the evolution of the proteins displaying them. In principle, one could study the partition function of the protein in the space of sequences, a problem analogous to a Potts model on an irregular lattice defined by the contact map of the native conformation. However, this approach has two drawbacks. First, it can be used only for small systems because the number of protein sequences of size N to be summed in the partition function scales as 20^N, a very fast growth even for small–sized proteins. Moreover, calculating the partition function for a given protein gives only information on that protein, while we would like to learn what are the most important 'bricks' that drive sequence design in general.

For this purpose, we employed a standard tool of statistical mechanics, that is the cluster expansion [19], on the network of interactions that stabilize the native states of proteins. The terms of the cluster expansions are the 'bricks', which are independent on the overall contact map of the protein. There are some important differences with the standard cluster expansion of dilute gases. Most importantly, proteins, even large ones are small systems and the thermodynamic limit cannot be applied. Also, proteins are dense systems and we do not expect that the highest–degree terms of the expansion vanish. Our approach is to evaluate what are the most important terms of each degree and eventually to infer what are the elementary structures of the proteins that most contribute to the partition function. On the other hand, studying a finite system we do not need to worry about the convergence of the expansion.

In Sect. 5 we will compare the elementary structures with highest entropy in the space of sequences with those found most commonly in natural proteins.

2 The cluster expansion

The partition function in the space of sequences, for a fixed native conformation, is

$$\mathcal{Z} = \sum_{\{\boldsymbol{\sigma}\}} \exp\left[-\beta_s \sum_{i<j} E_{\sigma_i\sigma_j}\Delta_{ij}\right], \qquad (1)$$

where $\boldsymbol{\sigma}$ is the N-element vector that specifies a protein sequence, $\beta_s \equiv 1/T_s$, $E_{\sigma\pi}$ is the interaction energy between amino acids of type σ and π, and Δ_{ij} is the contact map of the protein, that is a binary matrix that specifies what amino acids is in spatial contact with what amino acid in the native conformation. Two amino acids are defined to be in contact if any two atoms belonging, respectively, to each of them have a distance lower than a threshold d_c and they are separated by at least other $i_{\min}$ residues along the chain.

Defining the Mayer functions $f_{ij} \equiv \mathrm{e}^{-\beta_s E_{\sigma_i\sigma_j}\Delta_{ij}} - 1$, the partition function can be rearranged as

$$\begin{aligned}\mathcal{Z} = \sum_{\{\boldsymbol{\sigma}\}} 1 + \sum_{\{\boldsymbol{\sigma}\}}\sum_{i<j} f_{ij} + \sum_{\{\boldsymbol{\sigma}\}}\sum_{i<j,k<l} f_{ij}f_{kl} \\ + \sum_{\{\boldsymbol{\sigma}\}}\sum_{i<j,k<l,m<n} f_{ij}f_{kl}f_{mn} + \cdots . \qquad (2)\end{aligned}$$

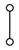
Due to the presence of the binary function Δ_{ij} in the Mayer's function, the different terms of this sum reflect the geometry of the interactions and can be represented as graphs like, for example

$$\overset{\bullet}{\underset{\bullet}{|}} = \sum_{\{\boldsymbol{\sigma}\}}\sum_{i<j} f_{ij} = n_{\overset{\bullet}{\underset{\bullet}{|}}} q^{N-2} \sum_{\sigma\pi}(\mathrm{e}^{-\beta_s E_{\sigma\pi}} - 1) \qquad (3)$$

$$= n_{\overset{\bullet}{\underset{\bullet}{|}}} q^{N} \times \overset{\circ}{\underset{\circ}{|}}, \qquad (4)$$

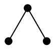
where $n_{\overset{\bullet}{\underset{\bullet}{|}}} = \sum_{i<j}\Delta_{ij}$ is the number of times that this specific graph appears in the protein, q is the

number of types of amino acids and



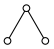
$$\equiv \frac{1}{q^2}\sum_{\sigma\pi} \mathrm{e}^{-\beta_s E_{\sigma\pi}} - 1 \tag{5}$$

is a shorthand notation for the term containing the average over the types of amino acids. Graphs with empty circles will be called 'interaction graphs' because they depend only on the interaction matrix $E_{\sigma\pi}$ and not on the overall structure of the protein. Similarly,

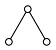


$$= \sum_{\{\boldsymbol{\sigma}\}} \sum_{i<j<k} (f_{ij}f_{jk} + f_{ik}f_{jk} + f_{ij}f_{ik}) \tag{6}$$

$$= n_{\wedge} q^N \times \tag{7}$$

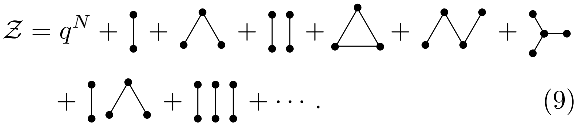


where



$$\equiv \frac{1}{q^3}\sum_{\sigma\pi\rho} \mathrm{e}^{-\beta(E_{\sigma\pi}+E_{\pi\rho})} - 2\frac{1}{q^2}\sum_{\sigma\pi}\mathrm{e}^{-\beta_s E_{\sigma\pi}} + 1, \tag{8}$$

and $n_{\wedge} = \sum_{i<j<k}(\Delta_{ij}\Delta_{jk} + \Delta_{ik}\Delta_{jk} + \Delta_{ij}\Delta_{ik})$, and so on for the other graphs. Using this formalism, the partition function reads



$$\mathcal{Z} = q^N + \cdots \tag{9}$$

For proteins, the number of terms is finite and the highest–order term can be, at most, of order of the number of contacts. Anyway, the strategy we will follow is that of selecting *a priori* a limited set Γ of connected graphs to investigate that are common to realistic protein conformations and study their contribution to the partition function. Vice versa, the summation of all possible graphs that compose the interaction network of a protein is equivalent to summing its partition function, is computationally hard and yields results that can be hardly generalized to other conformations.

There are two properties that are useful in this expansion. First, in each term the 'energetic' and the 'geometric' parts of the graph factorize, and consequently all of them are in the form

$$\gamma = n_\gamma q^N \tilde{\gamma}, \tag{10}$$

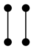
where γ denotes the kind of graph, n_γ is the number of times that the specific graph appears in the native conformation of the protein and $\tilde{\gamma}$ is the corresponding interaction graph, that is

$$\tilde{\gamma} \equiv \frac{1}{q^N}\sum_{\{\boldsymbol{\sigma}\}} \prod_{(i,j)\in\text{edges}} \left(\mathrm{e}^{-\beta_s E_{\sigma_i\sigma_j}} - 1\right) \tag{11}$$

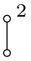
Moreover, if the native conformation of a protein displays disjoint interaction patterns, the partition function factorizes into their connected components.

An unhandy feature of Eq. (9) is that it contains not only connected graphs describing the interaction patterns contained in the native structure, but also all possible disjoint combinations of them. If the protein were an infinite system, with all residue pairs in contact with each other, one could use the connected–graph theorem [19] and the partition function in the grand-canonical ensemble could be written as the exponential of the sum of contributions from connected graphs only. Since we are focusing on finite–size properties, we followed another strategy.

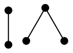
The value of a disjoint interaction graph is just the product of its components. The problem is to count the number of occurrences of the corresponding pattern in the protein. For example,



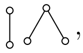
$$= q^N\left[\binom{n_{\mid}}{2} - n_{\wedge}\right]\cdot \tag{12}$$

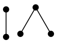


where the binomial coefficient counts the number of ways one can extract two links from the native contact map with $n_{\mid}$ edges and $n_{\wedge}$ removes from that count the number of instances in which the two links are sharing a common node, thus not contributing to Eq. (12). Similarly,

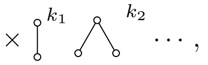


$$= q^N\left[\binom{n_{\mid}}{1}\binom{n_{\wedge}}{1} - n_{\wedge\!\wedge} - n_{\succ} - n_{\triangle} - n_{\wedge}n_{\mid\wedge/\wedge}\right] \times \tag{13}$$



where the negative terms enumerate all possible cases in which the connected components and can be found as superimposing elements in the native contact map, that is sharing common nodes and/or edges. In particular, $n_{\mid\wedge/\wedge} = 2$ counts in how many ways the disjoint components of the graph  can be superimposed to form the graph . As a matter of fact, all terms of Eq. (9) can be written in a form analogous to Eq. (12),

$$\mathcal{Z} = q^N \sum_{k_1=0}^{n_{\mid}} \sum_{k_2=0}^{n_{\wedge}} \cdots \left[\binom{n_{\mid}}{k_1} \binom{n_{\wedge}}{k_2} \cdots - \nu_{k_1,k_2,\ldots} \right]$$



(14)

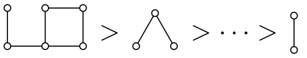
where a given choice of $k_1, k_2, \ldots$ corresponds to an in general disjoint graph, and $\nu_{k_1,k_2,\ldots}$ is the number of ways this graph can be formed by extracting its disjoint components from the native pattern of interactions in such a way that at least some of those components superimpose with each other. For example, $\nu_{2,0,0,\ldots} = n_{\wedge}$, $\nu_{1,1,0,\ldots} = n_{\wedge\!/} + n_{\triangle} + n_{\succ} + n_{\wedge} n_{\mid\wedge/\wedge}$ (while $\nu_{0,0,0,\ldots} = 0$ for the graph with no links, and $\nu_{1,0,\ldots} = \nu_{0,\ldots,0,1,0,\ldots} = 0$ for all connected graphs).

Importantly, the sums in Eq. (14) are taken only for connected graphs in our preselected set Γ; otherwise, Eq. (14) would be as intractable and ungeneralizable as Eq. (1). Note also that in general several prefactors $[\ldots]$ in Eq. (14) can be zero.

One can evaluate $\mathcal{Z}$ in mean–field theory by applying an approximation similar to saddle–point to the sums in Eq. (14). The largest term in each binomial sum is that associated with $k_\gamma = n_\gamma \tilde{\gamma}/(\tilde{\gamma}+1)$. If $\tilde{\gamma} \gg 1$ then the dominant term is $k_\gamma \approx n_\gamma$, corresponding to the disjoint graphs with the maximum number of connected subgraphs belonging to Γ. For this graph, the term ν cannot be neglected because the probability of superimposing nodes is also maximum. The dominant term is then given keeping the subgraphs associated with the largest $\tilde{\gamma}$ whose nodes do not overlap. We call Γ^* this set and $\tilde{n}_\gamma$ the number of subgraphs γ present in the dominant term (in general $\tilde{n}_\gamma < n_\gamma$ due to the no-overlap constraint). In this case the prefactor of the subgraphs in Eq. (14) is of the order of 1 and thus

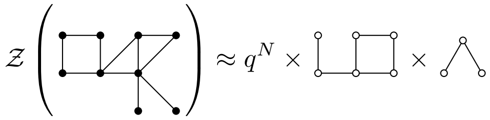
$$\mathcal{Z} \approx q^N \prod_{\gamma \in \Gamma^*} \tilde{\gamma}^{\tilde{n}_\gamma}. \qquad (15)$$

Note that Γ^* and $\tilde{n}_\gamma$ depend on the given native state contact map. A simple way to calculate them is to order the $\tilde{\gamma} \in \Gamma$ from the largest, identifying in this descending order which of them are present in the contact map of the protein, and exclude those whose nodes overlap with the selected ones. For example, assuming that



(16)

belong to Γ, one obtains



(17)

A way equivalent to that of Eq. (15) to estimate the importance of the graphs is to consider the corresponding free energy, that in the case $\tilde{\gamma} \gg 1$ reads

$$-T_s \log \mathcal{Z} = -T_s N \log q - T_s \sum_{\gamma \in \Gamma^*} \tilde{n}_\gamma \log \tilde{\gamma}. \quad (18)$$

Note that Eq. (18) defines an approximate free energy, due to using only connected graphs from Γ in the cluster expansion and to the saddle point approximation discussed above.

The quantities one needs to study numerically are then the $\log \tilde{\gamma}$ for the different types of connected graphs in Γ.

3 Numerical evaluation of the contribution of connected graphs

The partition function depends not only on the native contact map, but also on the interaction matrix $E_{\sigma\pi}$ and on the evolutionary temperature T_s.

3.1 The system

As interaction matrix we employed that of Table VI of the Miyazawa–Jernigan (MJ) work [20], obtained from the statistics of contacts in the Protein Data Bank. Although these potentials are a crude simplification of the actual interactions between amino acids [21], they contain the basic features that are needed in a coarse–grained description of proteins as that given by

Eq. (1), like attraction between opposite charges and among hydrophobic residues, repulsion of similar charges, etc. We have modified the MJ matrix removing the term associated with the interaction between cysteines (substituting it with the matrix average) because disulphide bonds display physical mechanisms, different from those of the other pairs, whose consideraion goes beyond the scope of the present work. As controls, we used random matrices in which the elements of the MJ matrix are reshuffled, loosing the correlations present in the original MJ, and random matrices with the same Gaussian distribution of elements as the original MJ ($\overline{E} = 0.02$ and $\sigma_E = 0.30$). Moreover, we also tested the effect of random matrices, either reshuffled from the MJ or extracted from a Gaussian distribution, in which all the diagonal elements are set to $\overline{E}$.

To set a realistic value for T_s we explored the thermodynamic properties of the sequence space of three small proteins (the villin headpiece [pdb code: 1bpi] of 36 residues, protein G [1pgb] of 56 residues, erabutoxin B [3ebx] of 62 residues) in the canonical ensemble at varying evolutionary temperatures with an adaptive simulated tempering algorithm [22]. We calculated numerically (Fig. 1) the average energy $\langle E \rangle$ of the protein as a function of T_s for the three proteins interacting with the MJ matrix and with randomly–reshuffled matrices. As expected [23], at low temperatures, $\langle E \rangle$ is a linear function of T_s, while at high temperatures $\langle E \rangle \sim 1/T_s$ which is typical of the random energy model. The crossover between the two behaviors is slightly system–dependent and takes place between $T_s^{c1} = 0.5$ and $= 1.0$. The statistical properties of natural sequences are compatible with the low–temperature regime [24]; we then focused our study on $T_s = 0.25$, at which all curves are in the linear regime.

3.2 Contribution of the connected graphs

Due to the factorization of all graphs γ into a geometric factor n_γ and an interaction term $\tilde{\gamma}$, it is interesting to identify the largest values $\tilde{\gamma}$ that yield the main contributions to the partition function. They can be either positive or negative (cf. Eqs. (5) and (8)) according to whether the typical contribution is attractive or repulsive, respectively. In principle, an even number of negative terms can result in a large positive contribution to the partition function (14). However, this would be a stabilization mechanism hardly to be generalized. Therefore, in our cluster expansion strategy we will consider only connected graphs with a large *positive* contribution as possible subsets of Γ^*.

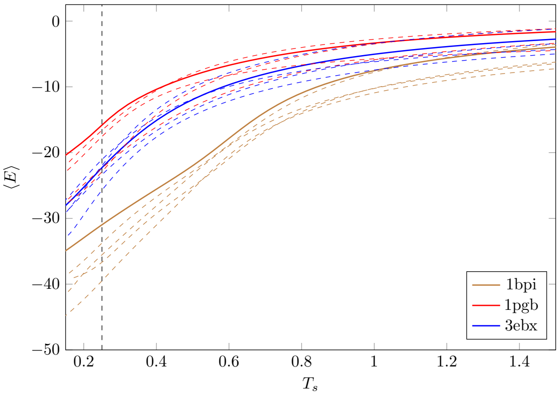


Fig. 1 The average energy $\langle E \rangle$ as a function of evolutionary temperature T_s for three proteins (in the legenda, their pdb codes) interacting with the MJ interaction matrix (solid curves) and for randomly–reshuffled matrices (dashed curves). The vertical bar indicates the temperature we chose, that is in the low–temperature regime for all systems.

The values for all connected graphs with up to 4 links and for some graphs with a larger number of links, computed for the MJ matrix, (blue bars in Fig. 2) are compared with the average contributions of random matrices. If the native conformation is made of $n_{\underset{\circ}{|}}$ independent contacts, the approximate free energy (18) is straightforwardly proportional to the logarithm of $\overset{\circ}{\underset{\circ}{|}} = 0.97$ multiplied by $n_{\underset{\circ}{|}}$. If these contacts are assembled in more complex patterns, this value of the free energy is corrected by including the corresponding graphs in Γ^*.

Note that, in the simple "diluted" example of $n_{\underset{\circ}{|}}$ independent contacts, one can explicitly sum the binomials in Eq. (14), so that the *exact* result is $\log \mathcal{Z} = N \log q + n_{\underset{\circ}{|}} \log(\overset{\circ}{\underset{\circ}{|}} + 1)$, as it should. The cluster expansion strategy that we propose fails in this simple example, but is in fact meant for dense connected graphs.

Since the protein is a dense system, the value of $\tilde{\gamma}$ tends to increase with the number of links. The graphs displaying the largest relative contribution to the partition function for the MJ matrix are the cycles with even number of links (Fig. 3). The largest contribution to this graphs comes from arrangements of pairs of residues of the kind $\alpha\beta\alpha\beta\ldots$, where the $E_{\alpha\beta}$ belong to the low–energy tail of the distribution of energies. These arrangements are equally likely for the random matrices; however, for purely random matrices whose elements are extracted from the Gaussian distribution with the same first two moments as the MJ interactions, these arrangements remain the favoured ones only when constraining the diagonal matrix terms (see discussion below). On the other hand, the reshuffled MJ matrices display similar values for the even cycles, both with and without the constraint on diagonal terms.

The fact that the contributions of even cycles with the MJ matrix are larger than those with the randomly reshuffled matrices indicates that the MJ matrix displays specific correlations between residues, like the case of charged and of hydrophobic residues.

On the other hand, cycles with an odd number of links cannot benefit from simply looping the elements with low $E_{\alpha\beta}$, but their value is controlled by the existence (by chance or by specific correlations in the interaction matrix) of a residue kind γ such that also $E_{\alpha\gamma}$ and $E_{\beta\gamma}$ are attractive enough. The fact that the contributions for odd cycles (triangles and pentagons) interacting with the reshuffled matrix, with the constraint on diagonal terms, are smaller than those interacting with the MJ matrix suggests the presence of anti–correlations in the interactions between natural amino acids (i.e., if $E_{\alpha\beta} \ll 0$ then the same is not true for $E_{\alpha\gamma} + E_{\beta\gamma}$).

Also the values for paths (i.e., unlooped chains of linked nodes) and stars (i.e., nodes linked only to a central node) yield a positive, non–negligible contribution which is similar between paths and stars with the same number of links. In all cases the graphs associated with the MJ matrix are larger than the randomly–reshuffled ones.

The different control models interacting with random energies display specific features (Fig. 2). The graphs calculated with matrices randomly reshuffled from the MJ do not change appreciably their behavior whether the diagonal elements are randomly reshuffled like all the other elements (red bars) or they are set equal to a constant value equal to the average $\overline{E}$ of the matrix. On the contrary, in the case of matrices obtained by a random extraction of energies from a Gaussian distribution (gray bars), one can obtain average values that are very large and not self–averaging (cf. the error bars), corresponding to the realizations of the matrix in which a particularly negative element appears on the diagonal. In this case, the value of all graphs become much larger then in the typical case because of the multiple appearance of the corresponding amino acids. In fact, matrices in which off–diagonal elements are generated randomly and diagonal elements are set to $\overline{E}$ (cf. purple bars) do not display this effect.

The overall scaling of the different free contributions with the number of links (Fig. 3), suggests that a sensible way of identifying Γ^* in the contact map of a protein is first to search for even cycles, then for stars and paths.

Cliques (i.e., fully–connected sets of nodes) made of four or more nodes yield a negative contribution (cf. Fig. 2). This appears as a consequence of the frustration of the system; according to the MJ matrix it is not possible to allocate more than three amino acids in such a way that all pairs are attractive. Cliques interacting with random matrices without a constrain on diagonal terms (red and gray bars in Fig. 2) display large positive averages, but they are not self–averaging. In fact, the average arises from a distribution with a long tail (cf. e.g. the inset of Fig. 2), producing a standard deviation that is similar to the average. The reason is that there is a non–negligible probability that some realizations of the random matrix display blocks of negative elements (e.g., a 2×2 block with $E_{\alpha\beta}$, $E_{\alpha\alpha}$, $E_{\beta\beta} \ll -T_s$). The larger the block, the (exponentially) larger the contribution to the clique but the lower the probability that the specific realization of the matrix can achieve such large values. This results in a wide distribution of values for cliques and thus in its non–self–averaging character. Diagonal elements are essential for this mechanism, as shown by the fact that random matrices with ‘typical’ elements $\overline{E}$ on the diagonal cannot display such large values for cliques. Cliques are the graphs that can build the largest number of links with a given number of nodes and thus can occasionally produce very large values.

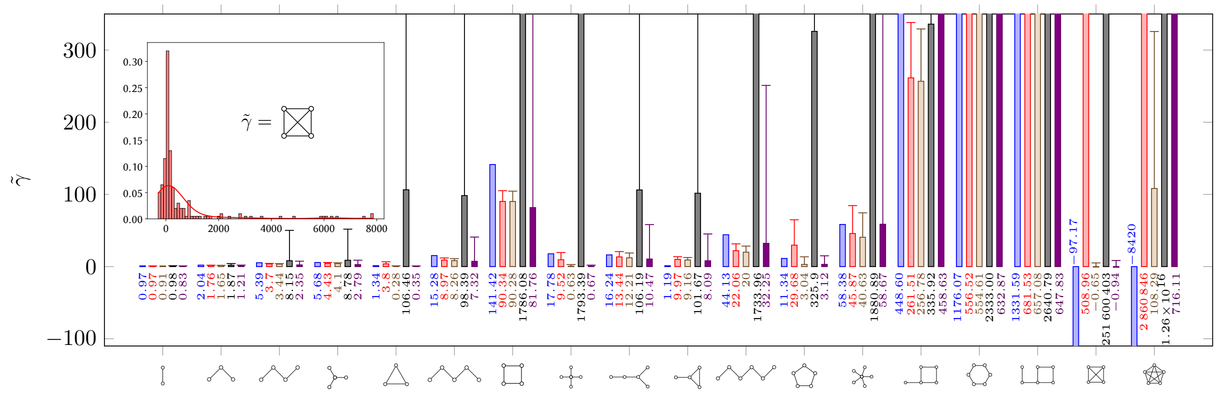


Fig. 2 The values of $\tilde{\gamma}$ for graphs interacting with the modified MJ matrix (blue bars), with randomly-reshuffled matrices (red bars), with randomly reshuffled matrices setting all diagonal elements equal to $\overline{E}$ (brown bars), with random matrices with average $\overline{E}$ and standard deviation σ_E (gray bars) and random matrices setting all diagonal elements equal to $\overline{E}$ (purple bars). For random matrices is indicated the average over 100 realizations and the thin bars indicates the standard deviation. In the inset, the distribution of values obtained for the 4–link clique interacting with the reshuffled matrices.

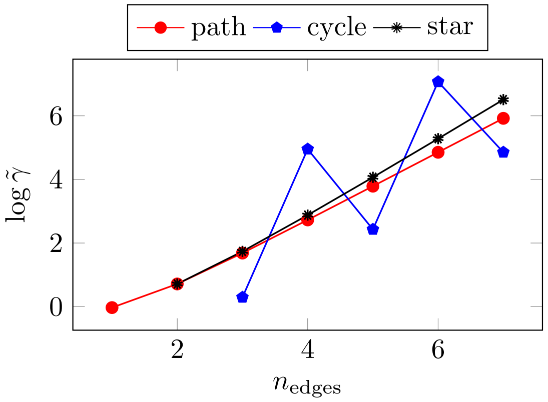


Fig. 3 The same data for the modified MJ matrix, shown as blue bars in Fig. 2 for paths, cycles and stars, displayed in semi–logarithmic scale as a function of the number of edges.

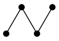
3.3 Dependence on the temperature

The relative weight of different graphs changes with T_s in a non–straightforward way (Fig. 4). Increasing the temperature from $T_s = 0.25$ one can observe several crossings between the curves associated with the different graphs. Above $T_s^{c1} \approx 0.6$, when the statistical properties of sequences space are described by the random energy model, the term associated with a single link becomes dominant for a large range of temperatures and, in general, simple small graphs become larger than more complex ones.

All graph contributions go to zero at large evolutionary temperature ($\beta_s \to 0$) because each Mayer function vanishes. Keeping only the first–order term in the high–temperature expansion of each factor of Eq. (11), one obtains

$$\lim_{\beta\to 0}\tilde{\gamma} = \beta^k \frac{(-1)^k}{q^N}\sum_{\{\sigma\}} E_{\sigma_1\sigma_2}E_{\sigma_3\sigma_4}\cdots E_{\sigma_{2k-1}\sigma_{2k}}, \tag{19}$$

where the relations among the indexes of the elements of the vector σ reflect the structure of the graph and k is the number of edges. Thus, one can obtain the k–point correlation function of the MJ matrix from the high–temperature limit of the graphs.

This correlation function for the MJ matrix can be either positive or negative. Its sign determines whether $\tilde{\gamma}$ approaches zero from the positive or the negative side. For example, in the MJ matrix the two–point correlation function is positive while the three–point correlation function  is negative.

3.4 Average Energy and Entropy

Since $-T_s \log\tilde{\gamma}$ is the contribution of the different graphs of Γ^* to the free energy of sequences, one can split them into an energetic and an entropic contribution. The average energy is $\langle E\rangle = -\partial\log\mathcal{Z}/\partial\beta_s$, and thus the contributions to it of the different graphs are $-\partial\log\tilde{\gamma}/\partial\beta_s$.

The contribution of the different graphs belonging to Γ^* to the entropy can be obtained applying the thermodynamic relation $S = \beta_s\langle E\rangle +$

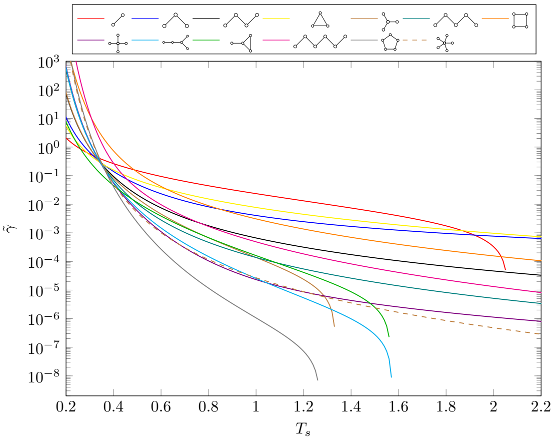


Fig. 4 The dependence of some graph cluster expansion contributions with MJ interactions on the evolutionary temperature T_s in semi–log scale. Some curves are displayed only partially because they become negative.

$\log \mathcal{Z}$ to Eq. (18), thus obtaining

$$S = N \log q + \tilde{n}_{|} \left(1 - \beta_s \frac{\partial}{\partial \beta_s}\right) \log \left[\overset{\circ}{\underset{\circ}{|}} \right] + \tilde{n}_{\wedge} \left(1 - \beta_s \frac{\partial}{\partial \beta_s}\right) \log \left[{}_{\circ}\!\overset{\circ}{\wedge}\!{}_{\circ} \right] + \cdots . \quad (20)$$

Each term of the sum should be read as an entropic cost with respect to the infinite–temperature entropy $N \log q$ (cf. Fig. 5, where the graphs are ordered according to their contribution to the free energy). Importantly, the entropic contribution of graphs is comparable to their energetic contribution, thus the sequence probability results from a non–trivial balance between stabilization of a given pattern and its degeneracy in sequence space. The contributions from different graphs to energy, entropy and free energy are roughly proportional to each other. Notable exceptions to this pattern are the squares and the 6–paths that yield a energy contribution comparable to the other graphs with the same number of edges but with a lower entropic cost.

4 Patterns in small systems

In the case of small networks of few amino acids, the approximations that lead to Eq. (18), the main result of our cluster expansion strategy, would not be justified. In this case, the partition function can be summed directly and the energy and entropy can be calculated exactly as $\langle E \rangle = -\partial \log \mathcal{Z} / \partial \beta$ and $S = \partial (T \log \mathcal{Z}) / \partial T$, respectively. In this context we will consider densities dividing extensive quantities by the number of nodes. In case of disjoint patterns the partition function is the product of the connected patterns and the free energy is the sum of the associated free energies.

The densities of energy and entropy calculated for some small systems are displayed with cyan and blue bars, respectively, in Fig. 6. The sum of the two bars gives, for each system, the density of free energy. The largest density of free energy is associated, also in the case of small systems, to even cycles. On the other hand, cliques are now not as penalized as in the case of larger systems.

A problem associated with this calculation is that the partition function includes, and may be dominated by, sequences with unrealistic concentration of the twenty amino acids. Specifically, the dependence of the stability of proteins on the native energy only [9] requires fixed concentrations. When studying the role of small graphs in a large protein, this is expected to be a minor problem, because the amino acids surrounding the graph of interest act as a reservoir of amino acids. For small systems this can become a problem.

We then studied small systems with different interaction patterns, by adding twenty chemical potentials μ_σ to set the average concentration of amino acids to their natural values [25]. The associated partition function is

$$Z = \sum_{\boldsymbol{\sigma}} \exp \left[-\beta_s \left(\sum_{i<j} E_{\sigma_i \sigma_j} \Delta_{ij} - \sum_i \mu_{\sigma_i} n_{\sigma_i} \right) \right], \quad (21)$$

where n_σ is the number of residues of kind σ in the sequence $\boldsymbol{\sigma}$.

To obtain explicit equations for the chemical potentials, one can approximate the average number of amino acids for a given system with that associated with the same number of independent contacts. One can begin evaluating μ_σ in the high–temperature limit, that gives

$$\mu_\sigma = \frac{1}{q(2q-1)-1} \left[\sum_{\rho\pi} E_{\rho\pi} - \sum_{\pi} E_{\sigma\pi} - \frac{1}{\beta} \log \left(\frac{2 n_\sigma}{N} \right) \right] . \quad (22)$$

However, the average concentrations $\langle n_\sigma \rangle = \beta_s^{-1} \partial \log Z / \partial \mu_\sigma$ calculated in the high–temperature approximations are not in good agreement with the correct ones (cf. red bars in

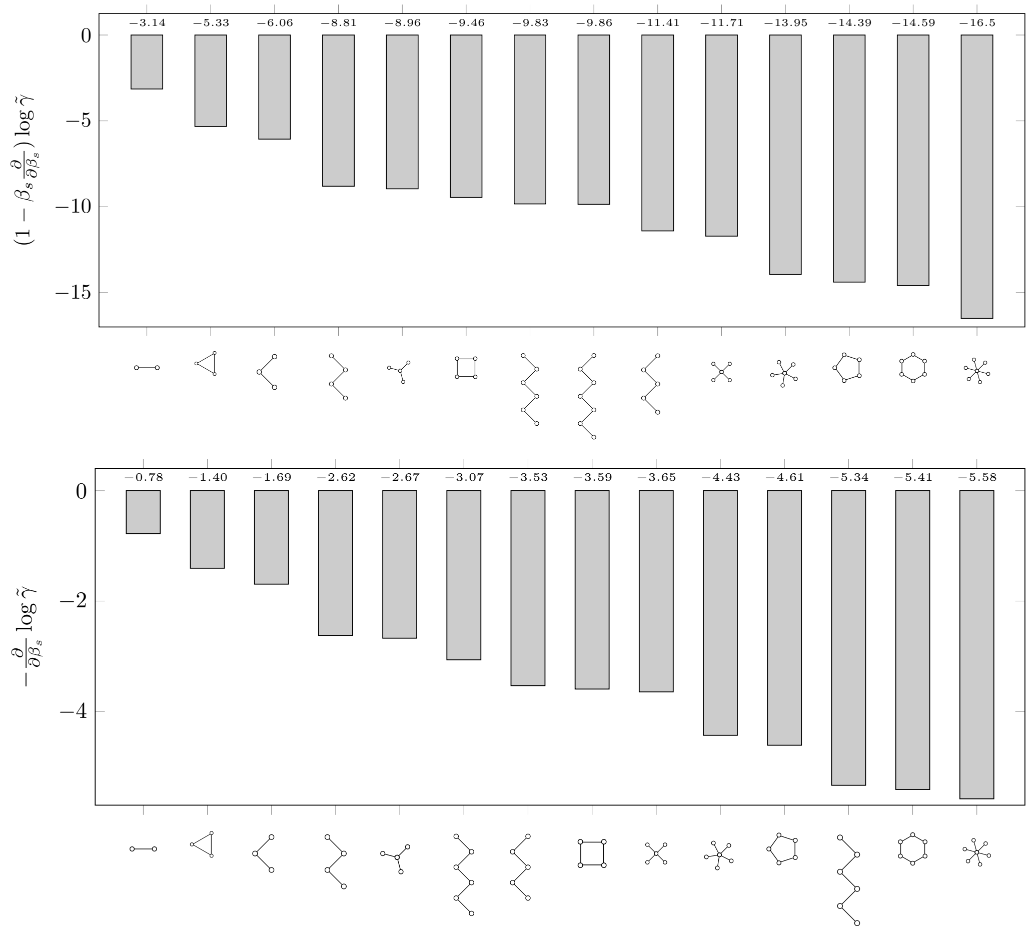


Fig. 5 The cluster expansion contribution of different graphs to the entropy in sequence space (upper panel) and to the energy (lower panel) for the MJ matrix at $T_s = 0.25$.

the inset of Fig. 6) and we corrected them in an iterative way applying 100 times

$$\mu_\sigma = \frac{1}{\beta} \log \left(p_\sigma \frac{\sum_{\rho\pi} \mathrm{e}^{-\beta(E_{\rho\pi} - \mu_\rho - \mu_\pi)}}{\sum_\pi \mathrm{e}^{-\beta(E_{\sigma\pi} - \mu_\pi)}} \right). \quad (23)$$

In this way, the average concentrations are well estimated even for non–trivial interaction patterns (cf. the inset of Fig. 6). Not unexpectedly, in the grand–canonical ensemble the potential energies are slightly less negative than in the canonical ensemble (with the exception of the single link) and the entropies slightly more positive. In fact, the constraints in the composition of the protein eliminate states where few types of amino acids are repeated in a regular way. Overall, the free energies in the grand–canonical ensemble do not depart drastically from the canonical ones (cf. Fig. 6).

5 Analysis of patterns in real proteins

Since the entropy (Eq. (20)) is a measure of the abundance of protein sequences displaying specific conformational patterns, we analyzed the contact maps of a set of natural proteins and compared the number of occurrences of the patterns (Fig. 5) with the corresponding entropy. We calculated the contact map of a set of 565 non–redundant proteins obtained from the pdb (< 25% sequence similarity).

To define the contact map, one has first to define what a contact is. We define two amino

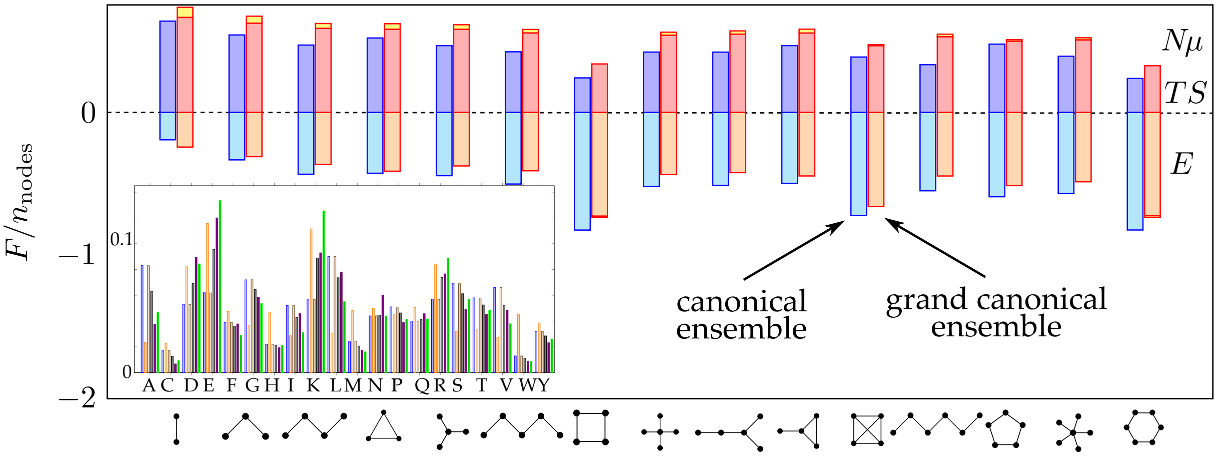


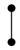
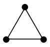
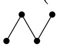
Fig. 6 The density of energy (cyan) and entropy (blue) calculated for some small patterns of interaction in the canonical ensemble. The density of energy (orange), entropy (red) and chemical potential (yellow) in the grand–canonical ensemble. In the inset, the natural fraction of the twenty types of amino acids (blue bars), the averages calculated in the high–temperature approximation from the grand–canonical partition function (orange) and those obtained from the iterative estimation of the chemical potentials (brown for an isolated contact, gray for a 2-edge path, purple for a 3-node clique and green for a 3-edge star.

acids to be in contact if any two atoms belonging to each of them, respectively, are closer than a distance d^* and farther than three residues along the sequence. We have chosen for d^* the value 3.7Å, at which the mean number of contacts in the dataset displays a sharp increase (cf. Fig. 7a). This value is also close to that at which the size of the giant component of the network has a sharp increase and is consistently lower than the percolation threshold (cf. Fig. 7b).

The naive counting of small patterns of contacts in the database of proteins displays some qualitative features predicted from the cluster expansion (Fig. 5, with counting of pattens in log scale). First, as discussed in Sect. 3.2, cliques are not likely. We can find in real proteins fewer small cliques than other kinds of graph and there are not cliques of degree higher than four. Small paths displaying large entropy (Fig. 7c) are quite abundant in real proteins. Moreover, squares and 3–edge stars are abundant and display rather large entropy. On the contrary, pentagons are found in real proteins but display a rather low entropy from the cluster expansion. Also, paths with increasing length are more and more present in real proteins, although their entropy roughly decreases with their length in Fig. 5.

However, it should be considered that the different graphs enumerated in the protein data set are not mutually exclusive. For example, paths are present in essentially all other graphs; in fact, they are the most abundant, increasing combinatorially with the number of edges. On the other hand, the estimation of the free energy in Eq. (18) requires the knowledge of the number $\tilde{n}_\gamma$ of graphs found as disjoint components, which is not the overall number n_γ of graphs in the native contact map.

The calculation of $\tilde{n}_\gamma$ is not straightforward because their definition requires to specify the order in which they have to be enumerated (cf. Eq. (16)). Following the prescription of Eq. (18), we defined a set Γ^* of graphs ordered according to the associated value of $\tilde{\gamma}$ and enumerated the graphs in that order, removing the nodes after they have been counted once (from left to right in Fig. 7d).

The simplest graphs  and are the most abundant ones and correspond to a large entropy contribution (upper panel in Fig. 5). The graphs  and  are predicted to display large entropy but they are suppressed by our node removal strategy (they are not displayed in Fig. 7d). The graphs and are correctly predicted as the most abundant ones after those already described. For the 4 most abundant graphs, remarkably, the counts found in natural proteins are larger than for random decoys (red bars in Fig. 5), obtained from the collapse of a protein model in which all atoms interact with the same energy [18]. Finally, our strategy to avoid double counting of nodes allows (roughly) to find a similar behaviour as a function of length for the paths abundances in real proteins (Fig. 7d) and the cluster expansion entropy (Fig. 5, upper panel), including the reduced entropy loss for 6–paths.

One can argue that the absence of triangles is due to the fact that they are not compatible with the formation of hydrogen bonds, that is one of the most important stabilizing interactions in protein. On the contrary, squares and paths can be stabilized by hydrogen bonds.

It is also known that proteins are rich of secondary structures, most notably α–helices and β–sheets. From the point of view of the graphs as defined above, both kind of structures are products of paths of various length, in the case of β–sheets sometimes decorated with squares (cf. Fig. 8).

Overall, paths and squares seem particularly abundant in proteins because associated with large entropy in sequence space.

6 Conclusions

A cluster expansion is a suitable tool to study the partition function of proteins. In the case of the partition function in the space of sequences of proteins with a specified native state, the clusters have a very clear meaning of patterns of interactions defined by the geometry of the native conformation.

Proteins are dense, finite systems. For this reason, the cluster expansion has to be handled differently than in simpler diluted systems, like non–ideal gases. Here, one cannot expect that the terms of higher degree become negligible. This is not a problem for the convergence of the expansion, because it contains a finite set of terms, but makes the evaluation of the importance of the different clusters tricky. Finding the most important cluster corresponds to a huge and complicated graph spanning most of the amino acids of the protein would not be particularly useful. It would be as cumbersome as summing the full partition function and the results for a protein could be hardly generalized to other proteins. For this reason, we studied classes of clusters (e.g., paths, polygons, etc.), comparing their relative importance. We also developed a sensible strategy to identify small important clusters in a connected network of interactions, solving the problem of avoiding a double–counting of the nodes.

Using a standard parametrization for the contact energies between amino acids, we found that clusters that contribute most to the partition functions are cycles with even numbers of nodes, while cliques are detrimental. These contributions can be partitioned into an energetic and an entropic part. Low–energy clusters are important for protein stability, high–entropy clusters for designability of the fold. As a rule, these two features are correlated. Small paths are highly entropic but poorly stabilizing. Stars and polygons are stabilizing but poorly entropic. Important exceptions are the square and the 6–path, that are both stabilizing and entropic.

We compared these results with the countings of different small contact patterns in real proteins. Overall, the kinds of patterns that are overrepresented in real proteins tend to agree with those predicted to have large entropy. In particular, paths and squares are patterns particularly abundant in α–helices and β–structures, respectively, that are the basic secondary structures of proteins.

Of course, one should consider that the thermodynamic stability is an epistatic constraint, but it is not the only factor that determines the evolutionary fitness of a protein. The formation of active sites, the kinetic accessibility, the need of avoiding aggregation are other requirements that could explain the differences between high–entropy graphs predicted by the model and those found in real proteins. Our findings can thus be regarded as the baseline on which evolution adds more stringent requirements.

Author contribution: MT performed the numerical calculations and helped to develop the theoretica calculations; AT and GT developed the theoretical calculations; GT wrote the manuscript with the help of MT and AT.

All data and codes used in this work can be downloaded from https://github.com/guidotiana/cluster-expansion .

References

[1] L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, and A. Giuliani. Protein contact networks:

An emerging paradigm in chemistry. Chem. Rev., **113** 1598–1613 (2013)

[2] M Vendruscolo, N V Dokholyan, E Paci, and M Karplus. Small-world view of the amino acids that play a key role in protein folding. Phys. Rev. E, **65** 061910 (2002)

[3] N. Kannan and S. Vishveshwara. Identification of side-chain clusters in protein structures by a graph spectral method. J. Mol. Biol., **292**, 441–464 (1999)

[4] M Vendruscolo, E Paci, C M Dobson, and M Karplus. Three key residues form a critical contact network in a protein folding transition state. Nature, **409** 641–645 (2001)

[5] V I Abkevich, A M Gutin, and E I Shakhnovich. Specific nucleus as the transition state for protein folding: evidence from the lattice model. Biochemistry, **33** 10026–10036 (1994)

[6] Guido Tiana, Fabio Simona, Giacomo M S De Mori, Ricardo A Broglia, and Giorgio Colombo. Understanding the determinants of stability and folding of small globular proteins from their energetics. Protein Sci., **13** 113–124 (2004) ISSN 0961-8368.

[7] E I Shakhnovich and A M Gutin. Engineering of stable and fast-folding sequences of model proteins. Proc. Natl. Acad. Sci. U. S. A., **90** 7195–7199 (1993)

[8] J D Bryngelson and P G Wolynes. Spin glasses and the statistical mechanics of protein folding. Proc. Natl. Acad. Sci. U. S. A., **84** 7524–7528 (1987)

[9] E I Shakhnovich and A M Gutin. A new approach to the design of stable proteins. Protein Eng., **6** 793–800 (1993)

[10] B Rost. Protein structures sustain evolutionary drift. Fold. Des., **2** S19–S24 (1997)

[11] Hao Li, Robert Helling, Chao Tang, and Ned Wingreen. Emergence of preferred structures in a simple model of protein folding. Science, **273**, 666–669, (1996) ISSN 00368075.

[12] Sridhar Govindarajan and R. A. Goldstein. Why are some proteins structures so common? Proc. Natl. Acad. Sci., **93** 3341–3345, (1996)

[13] Peter G. Wolynes. Symmetry and the energy landscapes of biomolecules. Proc. Natl. Acad. Sci. U. S. A., **93** 14249–14255, (1996)

[14] Amos Maritan, Cristian Micheletti, and Jayanth R. Banavar. Role of secondary motifs in fast folding polymers: A dynamical variational principle. Phys. Rev. Lett., **84**, 3009–3012, (2000) ISSN 00319007.

[15] K W Plaxco, K T Simons, and D Baker. Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol., **277**, 985–994 (1998)

[16] Jeremy L England and Eugene I Shakhnovich. Structural determinant of protein designability. Phys. Rev. Lett., **90** 218101 (2003)

[17] I N Berezovsky, A Yu Grosberg, and E N Trifonov. Closed loops of nearly standard size: common basic element of protein structure. FEBS Lett., **466**, 283–286 (2000)

[18] M. Negri, G. Tiana, and R. Zecchina. Native state of natural proteins optimizes local entropy. Phys. Rev. E, **104** 064117 (2021) ISSN 2470-0045.

[19] K Huang. *Statistical Mechanics.* J. Wiley and Sons., 1987.

[20] S Miyazawa and R Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules, **18**, 534–552 (1985)

[21] Paul D. Thomas and Ken A. Dill. Statistical potentials extracted from protein structures: How accurate are they? J. Mol. Biol., **257**, 457–469 (1996)

[22] G Tiana and L Sutto. Equilibrium properties of realistic random heteropolymers and their relevance for globular and naturally unfolded

proteins. Phys. Rev. E, **84** 61910 (2011)

[23] S Ramanathan and E I Shakhnovich. Statistical mechanics of proteins with evolutionary selected sequences. Phys. Rev. E, **50** 1303–1312 (1994)

[24] Giulia Franco, Matteo Cagiada, Giovanni Bussi, and Guido Tiana. Statistical mechanical properties of sequence space determine the efficiency of the various algorithms to predict interaction energies and native contacts from protein coevolution. Phys. Biol., **16** 046007 (2019) ISSN 1478-3975.

[25] T. E. Creighton. *Proteins: Structures and Molecular Properties.* W. H. Freeman and Co., 1992.

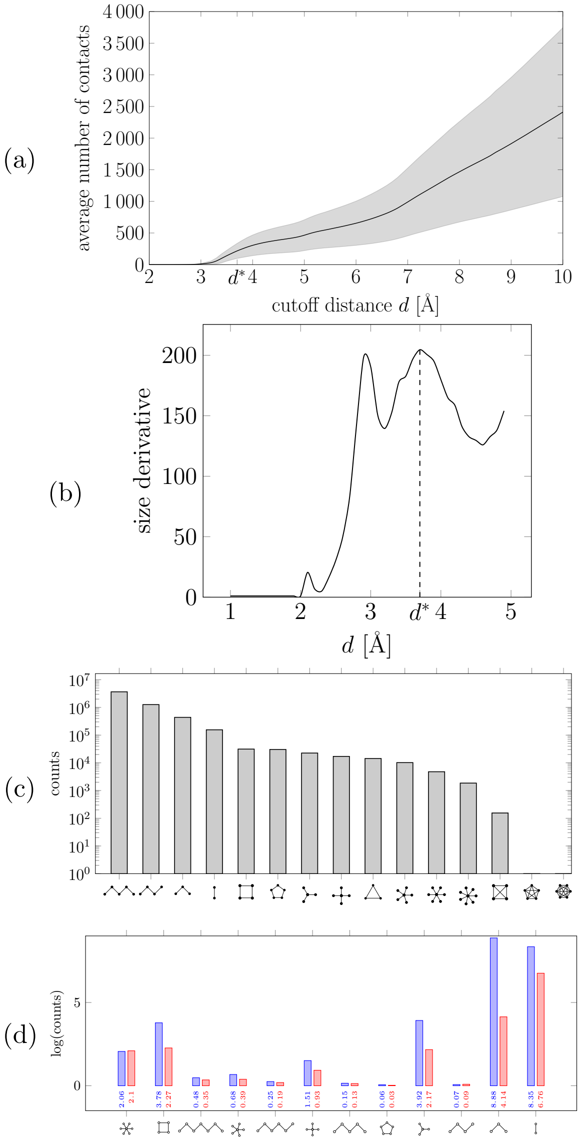


Fig. 7 (a) Average number of contacts in the protein dataset as a function of the threshold distance d^*. (b) Numerical derivative of the size of the graph giant component as a function of d^*. (c) The number of graphs found in a non–redundant dataset of natural proteins, (d) the average number of graphs per protein (blue bars), enumerated without double–counting nodes in the order displayed along the horizontal axis from left to right. In red, the same quantity calculated on a set of random decoys.

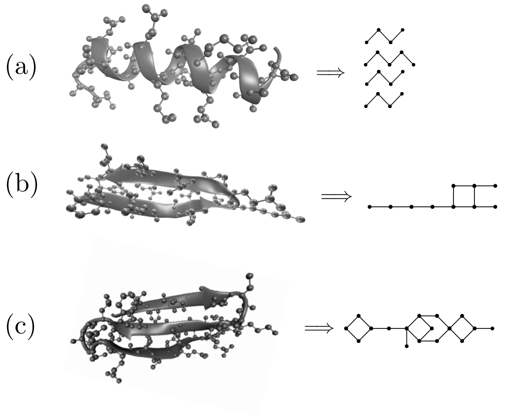


Fig. 8 The graphs arising from (a) the α–helix of 1pgb protein, (b) the first β–hairpin of protein 1h4x and (c) the first 3-β–sheet of protein 3mmy.