SCIENTIFIC REPORTS

OPEN

Network measures for protein folding state discrimination

Giulia Menichetti[1], Piero Fariselli[2] & Daniel Remondini[1]

Received: 06 January 2016
Accepted: 24 June 2016
Published: 28 July 2016

Proteins fold using a two-state or multi-state kinetic mechanisms, but up to now there is not a first-principle model to explain this different behavior. We exploit the network properties of protein structures by introducing novel observables to address the problem of classifying the different types of folding kinetics. These observables display a plain physical meaning, in terms of vibrational modes, possible configurations compatible with the native protein structure, and folding cooperativity. The relevance of these observables is supported by a classification performance up to 90%, even with simple classifiers such as discriminant analysis.

Protein folding is one of the most studied biophysical problems[1], and despite the fact that protein folding is a straightforward biophysical process[2], up to now there is not a general agreement on how and why proteins fold[3]. Experimentally, protein folding kinetics is divided into two fundamental categories: Two-State (TS) folding and Multi-State (MS) folding. While Two-State kinetics can be considered as an "all-or-none" transition, Multi-State folding displays at least one or more intermediates. Measuring experimentally the type of protein kinetics is not an easy task[3], and computational studies can help unraveling relevant mechanisms[4]. The classification of proteins in these two major groups and the related prediction of folding rates have been widely debated in recent years[4–8]. Previous studies have focused on several different types of predictors[9–11], exploiting the main features of protein primary structures and protein contact map representations (for a review see ref. 3). The geometry of the native protein structure plays a relevant role to infer the value of the folding rate. For this task, different predictors have been proposed based on: structural topology measures such as contact order[12,13] and long range contact order[14], clustering coefficient, characteristic path length and assortativity coefficient[11], cliquishness[15], chain length and amino acid composition[9,10]. These observables or combinations of them were usually evaluated by means of binary logistic regression (BLR) and support vector machine (SVM). In particular, SVM classifiers map the data into a higher dimensional feature space, that is usually not easily interpretable in terms of the original variables. Most predictors do not usually perform in the same way both for Two-State and Multi-State proteins, causing unbalanced value of sensitivity and sensibility according to the target of the analysis.

In this paper we focus on the problem of the discrimination of protein folding state, and not on the real-valued prediction of the folding rates. Our aim is to find physics-based, and easy interpretable observables, that can be related to the folding state classification. We propose novel observables based on the network properties of the native structure only, and we show that, together with a clear physical interpretation, they also predict with high performance if a protein behaves as Two-State or Multi-State by using simple discriminant methods. As done before by other authors[16–18], we represent the protein 3D structure as a contact map between amino-acid residues (Protein Contact Network PCN, see Fig. 1 for an example). The PCN is the adjacency matrix of a graph, whose links represent the contacts between residues. Our assumption is that the native PCN contains a clue of the protein folding kinetics. In this respect, we introduce three observables that should take into account that Multi-State proteins must be trapped into one or more intermediate states. First of all, we make the hypothesis that MS protein should have more configurational microstates to explore than TS proteins, and we implemented a measure of Network Entropy to quantify this aspect onto a combination of the full PCN with contact potentials as weights for the existing PCN links. Second, from a modified version of the PCN, that keeps only long-range contacts but preserving network connectivity, we evaluated the spectrum of the Laplacian matrix, since it has been shown that its vibrational properties can be used to model experimental data[19,20]. Finally, in order to measure the folding cooperativity[21], we evaluate the fraction of sequence separation (diagonals of the full PCN) that do not contain residue contact pairs. The rationale of this measure is that the more diffuse is the cooperation (most of the diagonal participate) the less probable is to be trapped into intermediate states. In order to keep these observables

[1]Department of Physics and Astronomy and INFN Sez. Bologna, University of Bologna, Viale B. Pichat 6/2 40127 Bologna, Italy. [2]Department of Comparative Biomedicine and Food Science, University of Padova, Viale dell'Universitá 16 35020 Legnaro, Italy. Correspondence and requests for materials should be addressed to D.R. (email: daniel.remondini@unibo.it)

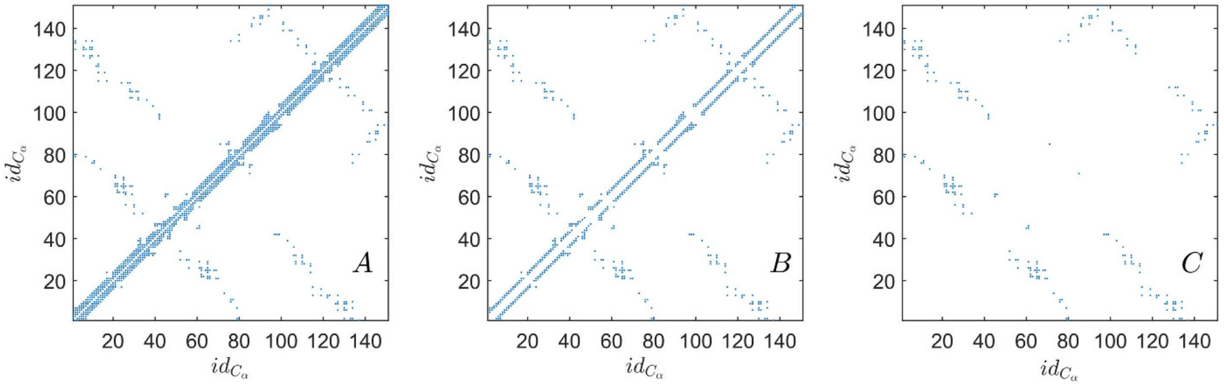

Figure 1. An example of different PCN representations for protein 1*A6N*, with MS folding kinetics and 151 C_α residues. In Panel A the whole PCN is displayed. In Panel B, once calculated the number of diagonals $b = 4$ needed to break the protein network in more than one component, 3 backbone diagonals were removed. In Panel C the related Long-range Interaction Network (LIN) is shown[11,14] in which no backbone diagonal is present.

(a) Classification performances					
%	λ_N	λ_{N-1}	λ_{N-2}	R_0	S_R
λ_N	76.6 ± 1.3	74.7 ± 1.4	70.7 ± 1.8	85.2 ± 1.4	78.4 ± 1.2
λ_{N-1}		76.7 ± 1.4	70.7 ± 1.4	**88.3** ± 1.1	75.9 ± 2.3
λ_{N-2}			77.6 ± 1.1	87.3 ± 1.4	76.5 ± 1.9
R_0				75.9 ± 1.2	84.5 ± 1.3
S_R					80.4 ± 1.8
(b) Matthews correlation coefficient *MCC*					
λ_N	0.57 ± 0.02	0.53 ± 0.02	0.46 ± 0.03	0.69 ± 0.03	0.60 ± 0.02
λ_{N-1}		0.58 ± 0.02	0.46 ± 0.03	**0.76** ± 0.02	0.57 ± 0.04
λ_{N-2}			0.59 ± 0.02	0.74 ± 0.03	0.56 ± 0.04
R_0				0.52 ± 0.02	0.67 ± 0.03
S_R					0.59 ± 0.04

Table 1. Classification performances of the newly defined observables and Matthews correlation coefficient MCC, based on quadratic discriminant analysis. The tables show the performances of couples of observables, with the performance of the single observables along the diagonal; the best overall performance is bold-typed. The results presented are the average values of 10-fold cross-validation over 10000 instances and their standard deviation.

as independent as possible from the protein size, they were accordingly rescaled by a function of residue chain length. In this paper we show that these observables perform very well even with a simple discriminant classifier, that allows to give a intuitive biophysical interpretation to our results.

Results

As previously mentioned, we introduced three main groups of observables (see Methods), i.e.

- Network Entropy-based ratio S_R, a measure related to the possible configurations consistent with the native protein structure PCN combined with the contact potential matrix P (see Methods).
- Laplacian-based observables λ_N, λ_{N-1}, λ_{N-2}, related to the highest-frequency vibrational modes of a modified PCN, in which the backbone is removed.
- Inter-residue link density R_0, measuring the fraction of sequence separation (diagonals of the full PCN) that do not contain residue contact pairs.

All the defined observables were considered as features for classification of TS and MS proteins by Fisher discriminant analysis (see Methods).

Classification with combinations of the novel observables is very high and balanced between the two classes, with performances up to 88% and Matthews Correlation Coefficient $MCC = 0.76$ (see Methods). Previous results, with a more complex classifier (nonlinear Support Vector Machine[13]) that does not allow a simple interpretation of the results as in our case, were around 80%. The details of all performances of the analyzed signatures, both with single observables and their combinations in couples, are shown in Table 1. We also considered higher-dimensional signatures (with combinations of 3 and 4 observables) but the performance was

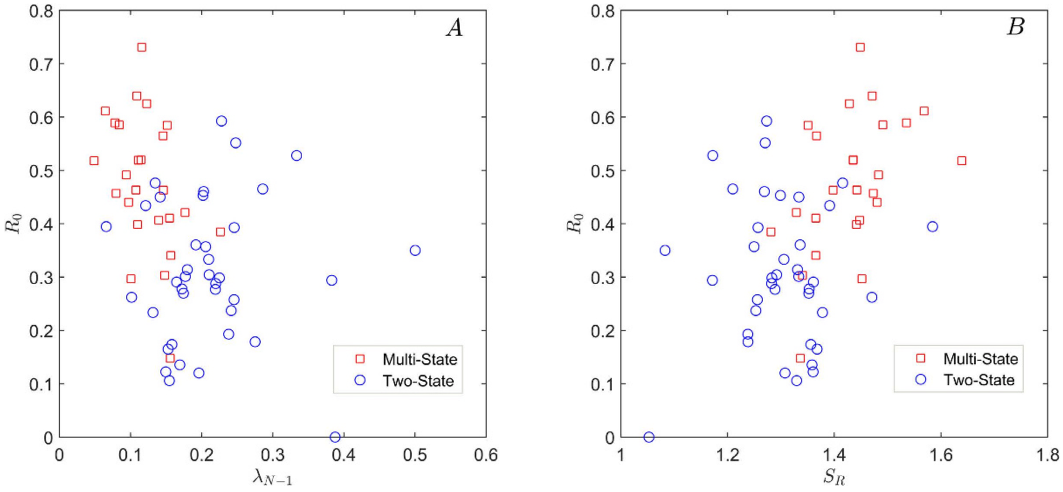

Figure 2. Scatterplots of two top-ranking classification couples of observables. Panel A: λ_{N-1} and R_0 (classification = 88.3% ± 1.1%, MCC = 0.76 ± 0.02). Panel B: S_R and R_0 (classification = 84.5% ± 1.3%, MCC = 0.67 ± 0.03).

not significantly increased. The combinations of the largest Laplacian eigenvalues λ_N, λ_{N-1}, λ_{N-2} with the link density R_0 produce the best performances of classification, with a top-score value given by the couple (λ_{N-1}, R_0), with 88.33% ± 1.10% correctly classified proteins, and a highly homogeneous performance on both classes (87.20% ± 2.21% MS and 89.07% ± 1.01% TS correctly classified proteins, $MCC = 0.76 \pm 0.02$). Considering the full PCN matrix, the classification performance of the Laplacian eigenvalue was reduced (see Methods). The entropy ratio S_R is the best single classifier (80.36% ± 1.81% correctly classified proteins, $MCC = 0.59 \pm 0.04$), and it also has a very high performance in combination with R_0 (84.50% ± 1.30% correctly classified proteins, $MCC = 0.67 \pm 0.03$). Classifying without cross-validation, i.e. using the entire set of (λ_{N-1}, R_0) features, we obtain a performance of 90.48%, with 92.00% for MS proteins and 89.47% for TS proteins. In Fig. 2 we show the scatter-plot for two top-scoring couples: (λ_{N-1}, R_0) and (S_R, R_0). It appears that MS and TS proteins are almost linearly separated in this parameter space, and this may allow a simple interpretation in terms of the observables, as stated in the Discussion section.

Since in previous studies[9] it has been shown that the chain length N_C is a good classifier of folding classes, we rescaled our observables in order to keep them as much independent as possible from protein length. Moreover, as a comparison for classification performance, we used N_C as a variable for discrimination. In our dataset, the N_C parameter correctly classifies 78.23% ± 1.52 proteins, with a large unbalance between correctly classified proteins from the two classes: 57.61% ± 3.04 for MS proteins and 91.80% ± 1.29 for TS proteins ($MCC = 0.54 \pm 0.03$). We also evaluated the classification power of some measures typically used in literature, i.e. the average hydrophobicity value $\langle h \rangle$[22], the contact order CO[12,13] and the long range contact order LRCO[14]. In the considered dataset the couple (N_C, $\langle h \rangle$) correctly classifies 73.71% ± 2.15 proteins, with $MCC = 0.44 \pm 0.05$. Both the structural topology measures perform poorly: CO correctly classifies 68.42% ± 0.65 proteins with $MCC = 0.36 \pm 0.01$ while LRCO guesses right 54.38% ± 2.41 proteins with $MCC = 0.14 \pm 0.05$. The best performing couple of observables (N_C and LRCO) reaches a performance of 80% ± 1, with a high heterogeneity of performance for TS and MS proteins ($MCC = 0.57 \pm 0.02$). A complete summary of the performances of classical measures (both singularly and in couples) can be found in Table 2. We also performed Discriminant Analysis with all possible combinations of new and classical observables, but the results did not outperform the performances obtained with the new observables only (CO and S_R: 81.4% ± 1.8, $MCC = 0.61 \pm 0.04$).

We also performed the same analysis on a different and more recent dataset[7] (http://kineticdb.protres.ru/db/index.pl), containing 85 proteins for which the PDB structure is available. In this dataset we found the same top-performing variables: the combination of R_0 with S_R had a 80.9% ± 1.3% ratio of correctly classified protein with $MCC = 0.58 \pm 0.03$, while R_0 for λ_{N-2} we found 79.7% ± 1.1% and $MCC = 0.59 \pm 0.02$.

Discussion

Trying to deduce properties of the proteins from their structure is still an open challenge: in this paper we propose novel observables, based on the network properties of PCN, that allow the discrimination between proteins with a different folding dynamics (TS proteins that present only two configurations (folded/unfolded) and MS proteins with a richer landscape of stable and metastable states) by looking only at information on their native state structure. Since previous work used the number of protein residues as a discriminating variable, but this may have some bias as we will discuss later, we only considered size-independent observables, by rescaling their value with an appropriate function of protein size. Thanks to this processing our analysis resulted more performing and robust, in particular in the "grey region" of short MS and long TS proteins. The performance of these observables is achieved by simple Fisher Discriminant Analysis, that allows a plain physical interpretation, in terms of vibrational modes, possible configurations compatible with the native protein structure, and folding cooperativity.

(a) Classification performances				
%	N_C	$\langle h\rangle$	CO	$LRCO$
N_C	78.2 ± 1.5	73.7 ± 2.1	78.4 ± 1.7	**80.0** ± 1.1
$\langle h\rangle$		57.3 ± 1.4	72.6 ± 2.8	62.7 ± 1.9
CO			68.4 ± 0.7	70.4 ± 2.1
$LRCO$				54.4 ± 2.4
(b) Matthews correlation coefficient *MCC*				
N_C	0.54 ± 0.03	0.44 ± 0.05	0.54 ± 0.04	**0.57** ± 0.02
$\langle h\rangle$		0.17 ± 0.03	0.44 ± 0.06	0.22 ± 0.04
CO			0.36 ± 0.01	0.44 ± 0.04
$LRCO$				0.14 ± 0.05

Table 2. Classification performances of the observables used in literature and Matthews correlation coefficient MCC. The tables show the performances of couples of observables, with the performance of the single observables along the diagonal; the best overall performance is bold-typed. The results presented are the average values of 10-fold cross-validation over 10000 instances and their standard deviation.

The observable R_0 simply counts the density of inter-residue distances in which there are no contacts, but nonetheless it results to be very powerful for this classification purposes though being independent on protein size. This means that the information contained in the PCN bands (the diagonals of the related adjacency matrix containing links between d–neighboring nodes) is very relevant, and possibly more complex measures could be developed based on this information. Another class of observables that we have introduced is based on the eigenvalues of the Laplacian operator applied to PCNs. In analogy with the physical Laplacian operator (acting on Euclidean space) the eigenvalues and eigenvectors can be put in relation with the main vibrational modes of the network and their respective frequencies. We remark that for Laplacian observables it was important to emphasize the role of long-range residue contacts, that effectively characterize the protein 3D structure, by removing the protein backbone with a procedure that preserves the network connectivity as a unique component: based on PCN properties, this filtering of non-relevant links is protein-specific, differently from the more general definition of long-range interaction commonly used, with a unique threshold for all proteins. Our Discriminant Analysis showed that in general TS and MS proteins are better classified by larger Laplacian eigenvalues, corresponding to high-frequency vibrating modes, at difference with small eigenvalues such as the Fiedler number. From the best performing couple of observables, λ_{N-1} and R_0 see Fig. 2, we deduce that TS proteins have larger values of fast-vibrating frequencies, and a larger number of inter-residue contacts (as can be seen in Fig. 2A). An interesting remark is that the vibrating modes associated with large eigenvalues tend to be more localized in specific residue chain regions (such as focusing modes in optics and whispering modes in acoustics[23]). It seems thus that the vibrating dynamics associated with specific regions of the residue may have a relevant role in these folding processes. The other observable we introduced, based on the concept of network ensemble, depends on an estimation of the size of network ensembles (from a canonical Statistical Mechanics point of view) that share common constraints (in our case the degree and the strength sequence of the PCN). As expected by the physical meaning of entropy, that counts the number of "microstates" corresponding to a "macrostate" characterized by some fixed constraints, TS proteins show a smaller value of S_R, that can be interpreted as a smaller number of topological configurations available to the related networks. A high number of available PCN states, given a fixed degree and strength sequence, is thus very likely associated with MS proteins, with more intermediate states during the folding process.

We remark that some known parameters used for TS-MS classification, such as the residue number N_C, are "extensive" variables that depend on protein size, thus their performances might be biased from the fact that many MS proteins are "long" and many TS proteins are "short" (e.g. considering protein size of the used database), and might not reflect real physical properties of the studied proteins. Since our analysis has a better performance of at least 10% with respect to these observables, this means that protein size is very relevant but not crucial to characterize MS and TS classes.

In conclusion, the high classification performance achieved, together with a direct physical interpretation, indicate that the newly introduced network-based observables can be relevant for a better comprehension of protein folding processes.

Methods

Experimental data.

The proteins to be classified as TS or MS are obtained from the manually-annotated dataset curated by Ivankov and Finkelstein[24]. The dataset consists of 63 proteins, 25 of which are classified as Multi-State and 38 as Two-State. The protein structures are taken from the Protein Data Bank (www.rcsb.org). We model the protein structure with its alpha carbon (C_α) trace. We collapse the entire protein structures into related contact matrices between the C_αs of the residues. Contact matrices represent a common way of modeling proteins, that guarantees a good representation of the complex relationship between structure and function of proteins, while cutting out the redundant information embedded in the whole 3D structure. Contact matrices are essentially networks in which the role of nodes is played by residues and edges or "contacts" depend upon a notion of "distance" between each couple of residues. The position of an entire amino acid is usually collapsed into the corresponding C_α and the ordering of nodes is physically justified by the primary structure of the protein, i.e. the protein backbone. The backbone is composed by residues that are in sequence and whose distance ranges 3–4 Å, the so-called "peptide bond". Once obtained the C_α spatial distribution, the contact matrix D is considered, where each element d_{ij} is the 3D Euclidean distance between the i^{th} and j^{th} residues. The protein contact network PCN is then obtained by choosing an upper threshold of 8 Å[25–28]:

$$PCN_{ij} = 1 \quad if \quad d_{ij} < 8\,\text{Å} \tag{1}$$

In order to build our Entropy-based observables, we retrieved also data regarding amino acidic interactions, such as hydropathy indexes[29,30] and contact potentials, namely, 20 × 20 matrices describing the interactions between the 20 side-chains[31–33]. Each element of the contact potential represents the interaction strength between a pair of amino acids at contact. In this paper we provide results only for the contact potential matrix M, as described in ref. 31, since the results obtained with other potentials[32,33] are very similar. For each protein, starting from the known residue sequence, we define the contact potential matrix P in which

$$P_{ij} = M(r_i, r_j) \tag{2}$$

where r_i is the amino acid residue corresponding to the i^{th} C_α, and the matrix P has the same size of the related PCN. Since P_{ij} can assume negative values, in order to have only positive weights (necessary for the calculation of network entropy) we shifted their values in each PCN to have the smallest weight equal to one. For the Entropy calculations only, we use the Hadamard (element-wise) product of PCN and P.

Entropy-based measure S_R. The first observable introduced is associated with the-so called Entropy of a network ensemble[34,35]. Network entropy is related to the logarithm of the number of typical networks (in our case the possible PCNs) that satisfy some given constraints based on node and link features of a real network instance (the studied protein). We hypothesize that the network structure of the protein native state retains information related to the protein folding process (such as the possible intermediate states that could be represented as non-native PCNs). It has been recently applied in a biological context, as a measure of the "parameter space" available to the cell (in terms of gene expression profile or clonal diversity) and it allowed to successfully characterize different cell states related to different cancer stages or to physiological ageing[36]. In our approach, each protein is considered as an undirected weighted network, in which we integrate the information on the topological structure given by protein contacts with the information on residue interactions given by $\{P_{ij}\}$ as weights for the existing PCN links (related to the contact potential matrix M described in ref. 31 and explained in details in the "Experimental Data" Section, Eq. 2).

For each protein, we calculated the Network Entropy S_{BS} for two different ensembles, with a different number of constraints: in the first ensemble, we only fix the strength sequence $\{s_i\}$ of the protein network (S_s), while in the second ensemble (S_{ks}) we fix both strength sequence $\{s_i\}$ and degree sequence $\{k_i\}$. The degree sequence $\{k_i\}$ and the strength sequence $\{s_i\}$ are respectively defined as the number of contacts and the weighted sum of contacts of the nodes in the network.

Network Entropy can be generally defined as

$$S = -\sum_{i<j}\sum_{w=0}^{\infty}\pi_{ij}(w)\log(\pi_{ij}(w)). \tag{3}$$

where weights w are positive and $\pi_{ij}(w)$ is the probability to observe weight w between residue i and j. The constraints previously defined for the calculation of maximum network entropy are written as

$$s_i^{prot} = \sum_j\sum_w w\pi_{ij}(w) \quad \forall\ i \tag{4}$$

$$k_i^{prot} = \sum_j\sum_{w\neq 0}\pi_{ij}(w) \quad \forall\ i, \tag{5}$$

where the average values of strength sequence and degree sequence over the network ensemble are enforced to match the real features of the selected protein, i.e. $\{s_i^{prot}\}$ and $\{k_i^{prot}\}$. The network entropy observable S_R is defined as the ratio between the two entropies with a different number of constraints

$$S_R = S_s/S_{ks} \tag{6}$$

with $S_s \geq S_{ks}$ given the fewer number of constraints. The closer is the value of S_R to 1, the less relevant is the role of the degree sequence constraint $\{k_i\}$, and thus most of the information on possible PCN configurations is enclosed in the strength sequence $\{s_i\}$ only. On the contrary, a large value of S_R implies that the given strength sequence is compatible with a larger number of degree sequences (corresponding to more possible PCNs). Thus, MS proteins could in principle have larger S_R values than TS proteins: having more stable (or metastable) configurations available could be reflected in a larger number of available configurations as measured by S_R.

Laplacian-based observables λ_i. The Laplacian operator L on networks is a positive semi-definite operator that plays a major role in the study of diffusion processes on networks, in node clustering and network visualization[37,38], and it has already been applied to characterize protein features[39]. Given an adjacency matrix A without self loops, we define the Laplacian operator as

$$L = K - A; \quad K_{ij} = k_i \cdot \delta_{ij} \tag{7}$$

Remarkably, in case of a N-lattice network, the eigenvalue problem for the Laplacian operator can be put in analogy with the discretization of an N-dimensional elastic membrane[40]. With this analogy in mind, the eigenvalues of the Laplacian matrix can be associated with the oscillating frequencies (harmonics) of the vibrating modes on the membrane, with the largest eigenvalues corresponding to the highest frequencies.

Since we suppose that the relevant information on folding kinetics can be contained in the long-range contacts of the native folded state[41], we decided to partially remove the backbone contacts from the original PCN. In more detail, for each protein we evaluated b, the number of d–diagonals (the set of links between nodes at a distance of d residues along the backbone, see Eq. 8) needed to break the protein network in more than one component. Then, $b - 1$ diagonals were removed from the PCN. We verified that with the full PCN our classification performance was significantly reduced by 2–3% on average, justifying our choice to emphasize the role of long-range connections with respect to the protein backbone. We remark that this procedure is specific for each protein, i.e. the parameter b depends on the PCN of each protein, and moreover, once removed $b - 1$ diagonals, the PCN is still connected, thus generating a unique eigenvalue spectrum for L. Also other authors considered a reduced PCN[11,14], but they used a unique threshold to define long-range interactions, taking only inter-residue distances $d > 12$ independently from protein size and structure (see Fig. 1). Once the laplacian spectrum of this modified version of PCN was computed for each protein, we considered as observables $\{\lambda_i\}$, the largest eigenvalues of L rescaled by the number of residues N_C. This rescaling was chosen because we observed a dependence of the largest eigenvalues on N_C, so our observables could in principle be linearly dependent on the number of residues. According to the vibrational interpretation of the Laplacian, these eigenvalues represent the highest vibrational frequencies associated with the long-range structure of the protein.

Inter-residue link density R_0. Each PCN is an adjacency matrix, in which the d-diagonals

$$PCN_{ij}, \forall\ i, j: |i - j| = d \tag{8}$$

contain all the links between residues with a sequence separation equal to d (ranging from 1 to $N_C - 1$) with respect to the protein backbone. The observable R_0 is defined as the ratio between the number of d-diagonals without links and the number of residues N_C of the protein, i.e.

$$R_0 = \frac{1}{N_C}\sum_d \delta\left(\sum_{i<j,|i-j|=d} PCN_{ij},\ 0\right) \tag{9}$$

where δ is the Kronecker delta. For a given protein, a low value of R_0 implies that the residues interact at many different levels of sequence separation (different values of d). On the contrary, a high value of R_0 indicates that, in such protein structure, the residue interactions are more localized and show less cooperativity.

Statistical analysis for classification. Fisher Discriminant Analysis, a robust classifier that allows plain interpretations of the classifying parameters due to the simple boundaries separating the classes, was applied to single observables and to their combinations (i.e. couples, triplets and quadruplets of observables). A 10-fold cross-validation with 10000 re-samplings was used to assess the performance of our classifiers, that will be described by the average value over the re-samplings and by the standard deviation as confidence interval. Given the presence of homologous proteins, in each partition of the 10-fold cross-validation all the homologous proteins were kept together, to reduce the risk of overestimating the classifier performance. In order to characterize the homogeneity of the classification performance over both TS and MS classes, we consider the Matthews Correlation Coefficient MCC, defined as

$$MCC = \frac{T_P \cdot T_N - F_P \cdot F_N}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}} \tag{10}$$

where T_P is the number of true positives, T_N the number of true negatives, F_P the number of false positives and F_N the number of false negatives. A coefficient of $+1$ defines a perfect prediction, 0 is nothing more than a random prediction, while -1 reflects total disagreement between prediction and observation.

Classical observables. We investigated the classification power of some classical observables, with the main purpose to have a comparison with the introduced measures. The statistical analysis and the estimation of the classification performances were the same as for the novel observables (see previous Subsection).

We analyzed the performance of the chain length N_C, the average hydrophobicity value $\langle h\rangle$, the contact order CO and the long range contact order LRCO.

The parameter h is chosen since hydrophobic force has always been indicated as one of the major drivers for protein folding[22]. Each amino acid is associated with a hydropathy index h_i, a number representing the hydrophobic or hydrophilic properties of its side-chain, thus each protein can be associated with an average hydrophobicity value $\langle h\rangle$:

$$\langle h\rangle = \frac{1}{N_C}\sum_i h_i^{KD} \tag{11}$$

where h_i^{KD} refers to the hydropathy index of residue i when the Kyte-Doolittle (KD) scale[29] is considered. The average hydrophobicity $\langle h \rangle$ has been often coupled with N_C to classify the protein folding kinetics.

We also considered the classification power of structural topology measures such as contact order[12,13]

$$CO = \frac{1}{N_C L_C} \sum_{ij}^{N_C} PCN_{ij} \cdot |i - j| \tag{12}$$

where L_C is the total number of contacts for the given PCN, and long range contact order[14]

$$LRCO = \frac{1}{N_C^2} \sum_{ij,|i-j|>12}^{N_C} PCN_{ij} \cdot |i - j| \tag{13}$$

which were used before for TS/MS classification purposes.

References

1. Tramontano, A. *The Ten Most Wanted Solutions in Protein Bioinformatics*. Chapman & Hall/CRC Mathematical and Computational Biology (CRC Press, 2005).
2. Anfinsen, C. B., Haber, E., Sela, M. & White, F. H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America* **47,** 1309–14 (1961).
3. Englander, S. W. & Mayne, L. The nature of protein folding pathways. *Proceedings of the National Academy of Sciences* **111,** 15873–15880 (2014).
4. Chang, C. C. H., Tey, B. T., Song, J. & Ramanan, R. N. Towards more accurate prediction of protein folding rates: a review of the existing Web-based bioinformatics approaches. *Briefings in bioinformatics* **16,** 314–24 (2015).
5. Huang, J. T., Wang, T., Huang, S. R. & Li, X. Prediction of protein folding rates from simplified secondary structure alphabet. *Journal of theoretical biology* **383,** 1–6 (2015).
6. Nissley, D. A. *et al.* Accurate prediction of cellular co-translational folding indicates proteins can switch from post- to co-translational folding. *Nat. Commun.* **7,** 10341 (2016).
7. Corrales, M. *et al.* Machine Learning: How Much Does It Tell about Protein Folding Rates? *PloS one* **10,** e0143166 (2015).
8. Shen, H.-B. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. *Journal of Biomedical Science and Engineering* **02,** 136–143 (2009).
9. Huang, J. T. & Cheng, J. P. Differentiation between two-state and multi-state folding proteins based on sequence. *Proteins: Structure, Function and Genetics* **72,** 44–49 (2008).
10. Huang, J. T., Xing, D. J. & Huang, W. Relationship between protein folding kinetics and amino acid properties. *Amino Acids* **43,** 567–572 (2012).
11. Song, J., Takemoto, K. & Shen, H. Prediction of protein folding rates from structural topology and complex network properties. *IPSJ Transactions on Bioinformatics* **3,** 40–53 (2010).
12. Plaxco, K. W., Simons, K. T., Ruczinski, I. & Baker, D. Topology, stability, sequence, and length: Defining the determinants of two-state protein folding kinetics. *Biochemistry* **39,** 11177–11183 (2000).
13. Capriotti, E. & Casadio, R. K-Fold: a tool for the prediction of the protein folding kinetic order and rate. *Bioinformatics* (*Oxford, England*) **23,** 385–6 (2007).
14. Gromiha, M. M. & Selvaraj, S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *Journal of molecular biology* **310,** 27–32 (2001).
15. Micheletti, C. Prediction of folding rates and transition-state placement from native-state geometry. *Proteins* **51,** 74–84 (2003).
16. Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M. Small-world view of the amino acids that play a key role in protein folding. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **65,** 1–4 (2002).
17. Di Paola, L., De Ruvo, M., Paci, P., Santoni, D. & Giuliani, A. Protein contact networks: An emerging paradigm in chemistry. *Chemical Reviews* **113,** 1598–1613 (2013).
18. Bartoli, L., Fariselli, P. & Casadio, R. The effect of backbone on the small-world properties of protein contact maps. *Physical biology* **4,** L1–L5 (2007).
19. Bahar, I., Atilgan, A. R. & Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design* **2,** 173–181 (1997).
20. Chang, I., Cieplak, M., Banavar, J. R. & Maritan, A. What can one learn from experiments about the elusive transition state? *Protein science : a publication of the Protein Society* **13,** 2446–57 (2004).
21. Dill, K. A., Fiebig, K. M. & Chan, H. S. Cooperativity in protein-folding kinetics. *Proceedings of the National Academy of Sciences of the United States of America* **90,** 1942–6 (1993).
22. Dill, K. A. & Maccallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **338,** 1042–1047 (2012).
23. Chen, G. & Zhou, J. *Vibration and Damping in Distributed Systems*, Volume 1 (1993).
24. Ivankov, D. N. & Finkelstein, A. V. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 8942–8944 (2004).
25. Aftabuddin, M. & Kundu, S. Weighted and unweighted network of amino acids within protein. *Physica A: Statistical Mechanics and its Applications* **369,** 895–904 (2006).
26. Barah, P. & Sinha, S. Analysis of protein folds using protein contact networks. *Pramana* **71,** 369–378 (2008).
27. Bagler, G. & Sinha, S. Network properties of protein structures. *Physica A: Statistical Mechanics and its Applications* **346,** 27–33 (2005).
28. Brinda, K. V., Surolia, A. & Vishveshwara, S. Insights into the quaternary association of proteins through structure graphs: a case study of lectins. *Biochemical Journal* **391,** 1–15 (2005).
29. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157,** 105–132 (1982).
30. Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18,** 534–552 (1985).
31. Bastolla, U., Farwer, J., Knapp, E. W. & Vendruscolo, M. How to guarantee optimal stability for most representative structures in the Protein Data Bank. *Proteins* **44,** 79–96 (2001).
32. Betancourt, M. R. & Thirumalai, D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein science: a publication of the Protein Society* **8,** 361–9 (1999).
33. Liwo, A. *et al.* A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *Journal of Computational Chemistry* **18,** 849–873 (1997).
34. Park, J. & Newman, M. E. J. Statistical mechanics of networks. *Physical Review E* **70,** 66117 (2004).

35. Anand, K. & Bianconi, G. Gibbs entropy of network ensembles by cavity methods. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **82,** 1–11 (2010).
36. Menichetti, G., Bianconi, G., Castellani, G., Giampieri, E. & Remondini, D. Multiscale characterization of aging and cancer progression by a novel Network Entropy measure. *Mol. BioSyst.* (2015).
37. Chung, F. R. K. Spectral Graph Theory. *Conference Board of the Mathematical Sciences* (1994).
38. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 7821–6 (2002).
39. Burioni, R., Cassi, D., Cecconi, F. & Vulpiani, A. Topological Thermal Instability and Length of Proteins. *Proteins: Structure, Function and Genetics* **55,** 529–535 (2004).
40. Biyikoglu, T., Leydold, J. & Stadler, P. *Laplacian Eigenvectors of Graphs: Perron-Frobenius and Faber-Krahn Type Theorems*. Lecture Notes in Mathematics (Springer Berlin Heidelberg, 2007).
41. Robson, B. & Garnier, J. Protein structure prediction. *Nature* **361,** 506 (1993).

Acknowledgements

D.R. and G.M. acknowledge EU FP7 Mimomics Grant n. 305280 and IT MIUR CNR Flagship Interomics Grant. D.R. acknowledges UniBO Farb Linea 1 Grant FFBO123105.

Author Contributions

D.R. and G.M. conceived the analysis. G.M. performed the analysis. P.F. provided and characterized the protein dataset. P.F., D.R. and G.M. analyzed the results and all authors reviewed the manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Menichetti, G. *et al.* Network measures for protein folding state discrimination. *Sci. Rep.* **6**, 30367; doi: 10.1038/srep30367 (2016).