Exploring the correlation between the folding rates of proteins and the entanglement of their native states

Marco Baiesi

Enzo Orlandini

Flavio Seno

Antonio Trovato

Dipartimento di Fisica e Astronomia "Galileo Galilei", Università di Padova, Via Marzolo 8, 35131, Padova, Italy
INFN, Sezione di Padova, Via Marzolo 8, 35131, Padova, Italy

Abstract.

The folding of a protein towards its native state is a rather complicated process. However there are empirical evidences that the folding time correlates with the contact order, a simple measure of the spatial organisation of the native state of the protein. Contact order is related to the average length of the main chain loops formed by amino acids which are in contact. Here we argue that folding kinetics can be influenced also by the entanglement that loops may undergo within the overall three dimensional protein structure. In order to explore such possibility, we introduce a novel descriptor, which we call "maximum intrachain contact entanglement". Specifically, we measure the maximum Gaussian entanglement between any looped portion of a protein and any other non-overlapping subchain of the same protein, which is easily computed by discretized line integrals on the coordinates of the C_α atoms. By analyzing experimental data sets of two-state and multistate folders, we show that also the new index is a good predictor of the folding rate. Moreover, being only partially correlated with previous methods, it can be integrated with them to yield more accurate predictions.

PACS numbers:

Submitted to: *J. Phys. A: Math. Gen.*
Keywords: Protein native structure, folding rates, topology, linking number

1. Introduction

Simple paradigms very often play an invaluable role to help understanding complex systems. A well known example is given by protein folding. Protein folding is the physical process by which a protein chain aquires its final three dimensional structure, the native state, that is usually biologically functional, in a reproducible manner. The characteristic time of this process is named folding time. Protein folding is complex because of the sheer size of protein molecules, the twenty types of constituent amino acids with distinct side chains, and the essential role played by the environment. Nevertheless, it is by now widely accepted that several aspects of the process driving a sequence of amino-acids to the corresponding native structure can be inferred by simple descriptors of the native-state geometry [1, 2]. For instance, it has been shown that the folding nucleus of a protein, including the residues whose interactions are essential for the folding to the native state, can be predicted through simulations of homopolymer models based on the mere knowledge of the native contact map, that is of the whole list of residue pairs in contact with each other [3, 4, 5, 6]. Other evidences of this simplicity are the ability of effective energy scores, derived by a statistical analysis of folded protein structures, to discriminate real native states among set of competing decoys [7, 8, 9, 10, 11] and the finding that the universe of possible proteins folds can be derived by simple coarse-grained models of polymers, which capture few universal properties typical of all amino-acids [12, 13, 14].

A further evidence of this emerging simplicity is the empirical result of Plaxco and coworkers [15, 16], who found a significant correlation between experimental folding rates of proteins, e.g. the inverse folding times, and a simple descriptor of the native state organisation, such as the *contact order*, that is, the average chemical length (in terms of the number of amino-acids) of the loop formed between residues which are in contact. This results is somewhat surprising because folding inevitably involves states other than the native one and these conformations might affect the kinetic process. Despite some evidence that this correlation is weak for proteins belonging to the all-β structural class [17], later studies confirmed correlations between folding rates and other descriptors of the native state organisation. These descriptors are *long range order* [18], the *number of native contacts* [19, 20], the *total contact distance* [21], the *cliquishness* [22], the *local secondary structure content* [23] and the *chain crosslinks contact order* [24].

The contact order and all its possible variants are descriptors that focus on the network of pairs of residues that are nearby in space regardless of the full spatial arrangement of the protein conformation. More realistically, one can however think of non-local descriptors that capture the degree of self-entanglement of the whole protein backbone, seen as a curve in a three dimensional space. An example is the *writhing number*, a measure of how a curve winds around itself in space [25]. In protein physics, the writhing number was first used in [26] to quantify the amount of self-threading in native state and it was later extended to perform a systematic classification of existing protein folds [27, 28].

After the seminal observation that the backbone of a protein may self entangle into physical knots [29, 30, 31, 32, 33, 34, 35, 36], a growing attention has been devoted to find new, topologically inspired, descriptors for quantifying accurately the winding of a protein with itself or with other molecules. Specific descriptors have been proposed to measure the amount and location of mutual entanglement between protein complexes [37, 40, 39, 38] or to detect specific topological knots, links and lassos within a single chain [40, 41, 42].

The aim of this study is to explore the correlation between protein folding rates and a novel topological descriptor of proteins three dimensional entanglement, which we name *maximum intrachain contact entanglement*. This indicator is the maximum value of the mutual entanglement measured between any looped portion of a protein and any other non-overlapping subchain extracted from the same protein. As a measure of the mutual entanglement, we consider the Gaussian double integral of two oriented curves. For closed curves this measure reduces to the Gauss linking number, a topological invariant that quantifies how pairs of loops are (homologically) linked [43]. Being quite easy to compute, the Gauss linking number has been extensively used in the past to characterise the mutual entanglement of diluted and concentrated solutions of linear polymers [44, 45, 46, 47], to estimate the linking probability and link complexity of pair of loops under geometrical constraints [48, 49, 50, 51] as well as to identify threadings in dense solutions of unlinked loops diffusing in a gel [52].

By exploring a data set of 48 proteins [22, 24, 53] for which the folding time is known experimentally, we compute the linear correlation coefficient between the values of our descriptor and the experimental folding rates. We show that the maximum intrachain contact entanglement captures aspects that are different from those highlighted by the contact order, and we describe how the two descriptors can be combined to improve the predictions of folding rates.

2. Methods

2.1. A topologically inspired descriptor

It is well known that the Gauss double integral

$$G \equiv \frac{1}{4\pi} \oint_{\gamma_1} \oint_{\gamma_2} \frac{\boldsymbol{r}^{(1)} - \boldsymbol{r}^{(2)}}{\left|\boldsymbol{r}^{(1)} - \boldsymbol{r}^{(2)}\right|^3} \cdot (d\boldsymbol{r}^{(1)} \times d\boldsymbol{r}^{(2)}) \tag{1}$$

between two closed curves γ_1 and γ_2 in $\mathbb{R}^3$ gives an integer number, known as the linking number, whose value is a topological invariant. A nice feature of this measure, however, is that it provides a meaningful assessment of the mutual entanglement also if either one or both curves are open [44, 47, 37].

Our strategy here is to consider as γ_1 and γ_2 any pair (γ_i, γ_j) of non-overlapping subchains extracted from the same protein backbone and to compute their Gaussian entanglement G_{ij}. Since the backbone of a native protein structure with N residues can

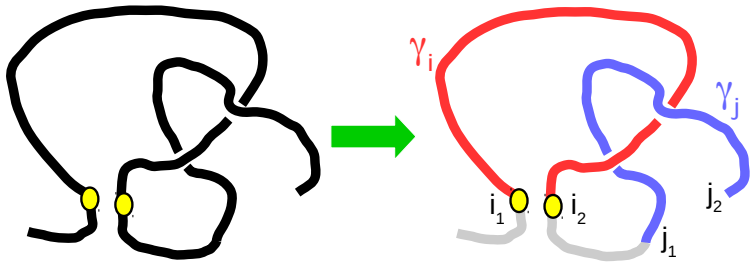

Figure 1. Sketch of the procedure we use for computing G'_{ij}. First, a looped portion γ_i ($i_1 \to i_2$, red) is identified when C_α coordinates of amino acids i_1 and i_2 (yellow ovals) are closer than $d = 9\mathring{A}$. Then, the double sum (4) is computed for any other portion γ_j ($j_1 \to j_2$, blue) preceding or following γ_i.

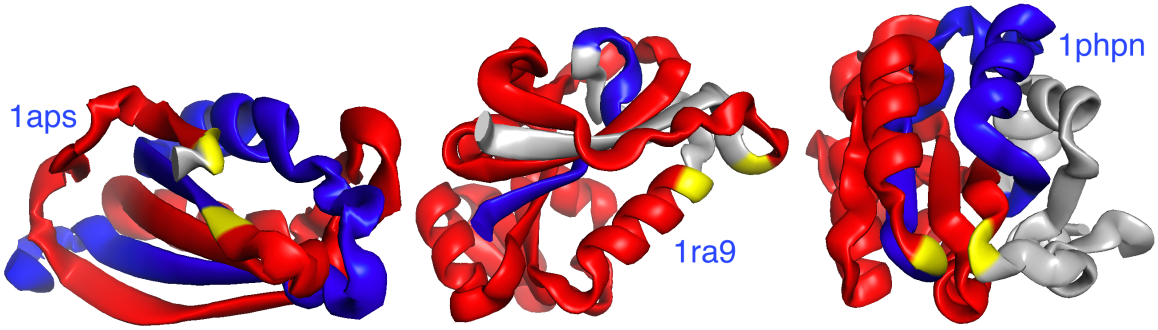

Figure 2. Following the color code of figure 1, examples of protein native structures in which we identified the subchains γ_i ($i_1 \to i_2$, yellow-red-yellow) and γ_j ($j_1 \to j_2$, blue) yielding the maximum intrachain contact entanglement.

be described as a discrete chain of monomers ($i = 1, \ldots, N$) placed at the positions $\boldsymbol{r}_i$ of the C_α atoms, it is natural to define the average positions

$$\boldsymbol{R}_i \equiv \frac{1}{2}(\boldsymbol{r}_i + \boldsymbol{r}_{i+1})\,, \tag{2}$$

and the bond vectors

$$\boldsymbol{dR}_i = \boldsymbol{r}_{i+1} - \boldsymbol{r}_i. \tag{3}$$

Hence, for a Given subchain γ_i with monomers from index i_1 to i_2, and a subchain γ_j with monomers from index j_1 to j_2 such that $\gamma_j \cap \gamma_i = \emptyset$, their Gaussian entanglement is given by

$$G'_{ij} \equiv \frac{1}{4\pi} \sum_{i=i_1}^{i_2-1} \sum_{j=j_1}^{j_2-1} \frac{\boldsymbol{R}_i - \boldsymbol{R}_j}{|\boldsymbol{R}_i - \boldsymbol{R}_j|^3} \cdot (\boldsymbol{dR_i} \times \boldsymbol{dR_j}) \tag{4}$$

where the prime in G' highlights the fact that the measure is for open chains. Note that, unlike in our previous study [37] where the entanglement was estimated between two different protein backbones, here the pairs of subchains (γ_i, γ_j) are extracted from the same protein backbone.

The definition (4) is rather generic and can be applied to any pair (γ_i, γ_j) of not overlapping portions of the protein backbone. Here we specialize the analysis to the

subset of (γ_i, γ_j) where the subchain γ_i has its first (i_1) and last (i_2) residues forming a *contact* $(i_1 \div i_2)$, i.e. when $|\boldsymbol{r}_{i_1} - \boldsymbol{r}_{i_2}| < d$, with $d = 9\mathring{A}$. With this restriction γ_i is essentially a loop. The same restriction is not applied to γ_j, which can either precede or follow γ_i along the protein backbone.

This way of detecting entangled configurations is sketched in figure 1. It is similar to that leading to the definition of "lassos" [42], although here we do not restrict the contacts to chemically strong bonds as in cystein pairs.

A preliminary analysis of our data set showed that such entangled configurations are not rare. Given their topological complexity, it is reasonable to think that these native states host proteins which might fold slowly, especially when the mutual entanglement between γ_i and γ_j is considerable.

Motivated by the considerations above, we perform a statistical analysis to test the existence of a negative correlation between folding rates and a quantitative measure of the intrachain entanglement present in a protein native structure. We define this measure to be the largest absolute value of the mutual entanglement found for all possible pairs (γ_i, γ_j) having the lasso structure discussed above:

$$|G'|_c = \max_{[i_1,i_2],[j_1,j_2]} |G'_{ij}|. \tag{5}$$

The index "c" indicates that this maximum intrachain *contact* entanglement is subject to the loop constraint of having the ends of γ_i in contact with each other, $i_1 \div i_2$.

With the same definition of contact between non consecutive residues we can introduce the absolute contact order (ACO). This is the average chemical distance $|j - i|$ between monomers in contact. Supposing that there are n_c of these contacts in the native state of a protein, we have

$$\text{ACO} \equiv \frac{1}{n_c} \sum_{i \div j} |j - i| \tag{6}$$

The relative contact order (RCO) is simply the ACO divided by the chain length N, which is the average of normalized chemical distances $|j-i|/N$ of residues in contact [15].

2.2. Data sets

We use two separate data sets. A first data set for two-state folders includes single-domain, non-disulfide-bonded proteins that have been reported to fold via two-state kinetics under at least some conditions [54]. We use folding rates as reported previously [24, 54], see table 1.

The second data set, for multistate folders, is summarised in table 2 and includes proteins that exhibit one or more folding intermediates in water, the entries 34-57 in table 1 from Ref. [53]. We use folding rates from that table with two exceptions. For 1ra9, we use the folding rate reported instead in Refs. [55, 56]. We then removed 1cbi and 1ifc from the data set, as they are both homologous to 1opa, thus sharing essentially

Table 1. Data set for two-state folders. N is the number of C_α atoms with available coordinates used in the computation of $|G'|_c$, ACO, RCO. 1bnza refer to chain A in the 1bnz protein-DNA complex. 1div.n refer to the N-terminal domain of protein 1div. 1hz6a refer to chain A in the 1hz6 protein complex. 1lmb3 refer to chain 3 in the 1lmb protein-DNA complex. 1urna refer to chain A in the 1urn protein-RNA complex.

PDB code	ln(rate)	N	$\|G'\|_c$	ACO	RCO
1afi	0.6	72	0.77	22.99	0.32
1aps	-1.47	98	1.62	34.1	0.35
1aye	6.63	80	0.27	13.83	0.17
1bnza	6.95	64	0.27	16.39	0.26
1bzp	11.12	153	0.47	24.28	0.16
1csp	6.54	67	0.4	19.65	0.29
1div.n	6.61	56	0.84	13.	0.23
1fkb	1.38	107	0.96	32.4	0.3
1hrc	8.75	104	0.56	23.56	0.23
1hz6a	4.1	62	0.54	17.36	0.28
1imq	7.28	86	0.5	18.24	0.21
1lmb3	11.01	80	0.3	15.14	0.19
1pgb	5.66	56	0.39	17.2	0.31
1poh	2.69	85	0.49	27.04	0.32
1psf	1.17	69	0.47	20.86	0.3
1shf	4.54	59	0.71	18.12	0.31
1ten	1.06	89	0.67	28.84	0.32
1tit	3.48	89	0.61	30.98	0.35
1ubq	7.35	76	0.47	21.82	0.29
1urna	2.5	96	1.15	28.34	0.3
1wit	0.41	93	0.72	33.35	0.36
256b	12.2	106	0.33	15.53	0.15
2abd	6.56	86	0.6	21.56	0.25
2ci2	4.03	64	0.68	20.23	0.32
2pdd	9.67	41	0.3	9.04	0.22
2vik	7.48	126	0.86	27.02	0.21

the same native structure. We kept 1opa because it has the intermediate rate among the three.

Note that two proteins, 1tit and 1ubq, belong to both data sets, since they are multistate folders in water, while switching to two-state kinetics upon different conditions.

3. Results

For each protein structure in the data sets described in section 2, we consider four different descriptors: the chain length N, the ACO, the RCO and the maximum intrachain contact entanglement ($|G'|_c$).

The values of $|G'|_c$ should be compared to the reference value of 1 found for two closed curves that form a standard Hopf link (the same as for two flat linked rings). We

Table 2. Data set for multistate folders. N is the number of C_α atoms with available coordinates used in the computation of $|G'|_c$, ACO, RCO. 1phpn and 1phpc refer to the N-terminal and, respectively, C-terminal domains of 1php. 1qopa and 1qopb refer to the chains A and, respectively, B of the 1qopa protein complex.

PDB code	ln(rate)	N	$\|G'\|_c$	ACO	RCO
1a6n	1.1	151	0.48	25.71	0.17
1aon	0.8	155	1.35	39.1	0.25
1bni	2.6	108	0.6	19.32	0.18
1brs	3.4	87	0.43	19.8	0.23
1cei	5.8	85	0.38	15.18	0.18
1eal	1.3	127	0.29	25.58	0.2
1fnf	5.5	94	0.7	28.27	0.3
1hng	1.8	97	0.78	31.04	0.32
1opa	1.4	133	0.34	29.92	0.22
1ra9	-2.46	159	1.65	40.71	0.26
1sce	4.2	101	0.46	23.41	0.23
1tit	3.6	89	0.61	30.98	0.35
1ubq	5.9	76	0.47	21.82	0.29
2a5e	3.5	156	0.56	15.83	0.1
2cro	3.7	65	0.25	13.39	0.21
2lzm	4.1	164	0.36	16.07	0.1
2rn2	0.1	155	1.	39.29	0.25
3chy	1.	128	0.98	18.96	0.15
1phpc	-3.5	219	1.23	32.51	0.15
1phpn	2.3	175	1.3	36.25	0.21
1qopa	-2.5	268	1.43	41.16	0.15
1qopb	-6.9	392	1.43	55.09	0.14

find $|G'|_c \geq 1$ for 9 out of the overall 46 proteins analyzed in this work (see tables 1,2). The largest value in our study is $|G'|_c = 1.65$ for the multistate protein 1ra9.

3.1. Two-state proteins

The linear correlations of the descriptors we consider with the natural logarithm of the experimentally measured folding rate are shown in figure 3 for two-state folder together with the corresponding Pearson correlation coefficient r. As already known [24], chain length is essentially not correlated with the folding rate ($r = 0.13$) for two-state folders (see figure 3(a)), whereas the best performance (see figure 3(d)) is achieved by RCO ($r = -0.87$), with negative correlation implying that slow folders have native structures with contacting residues that are on average well separated along the sequence. Although with a lower quality with respect to RCO, the correlation of ACO with the folding rate of two-state folders is still significant ($r = -0.65$) and with the proper (negative) slope (see figure 3(c)).

For the novel topological descriptor that we introduce in this work, the maximum intrachain contact entanglement $|G'|_c$, the correlation is essentially as good ($r = -0.64$)

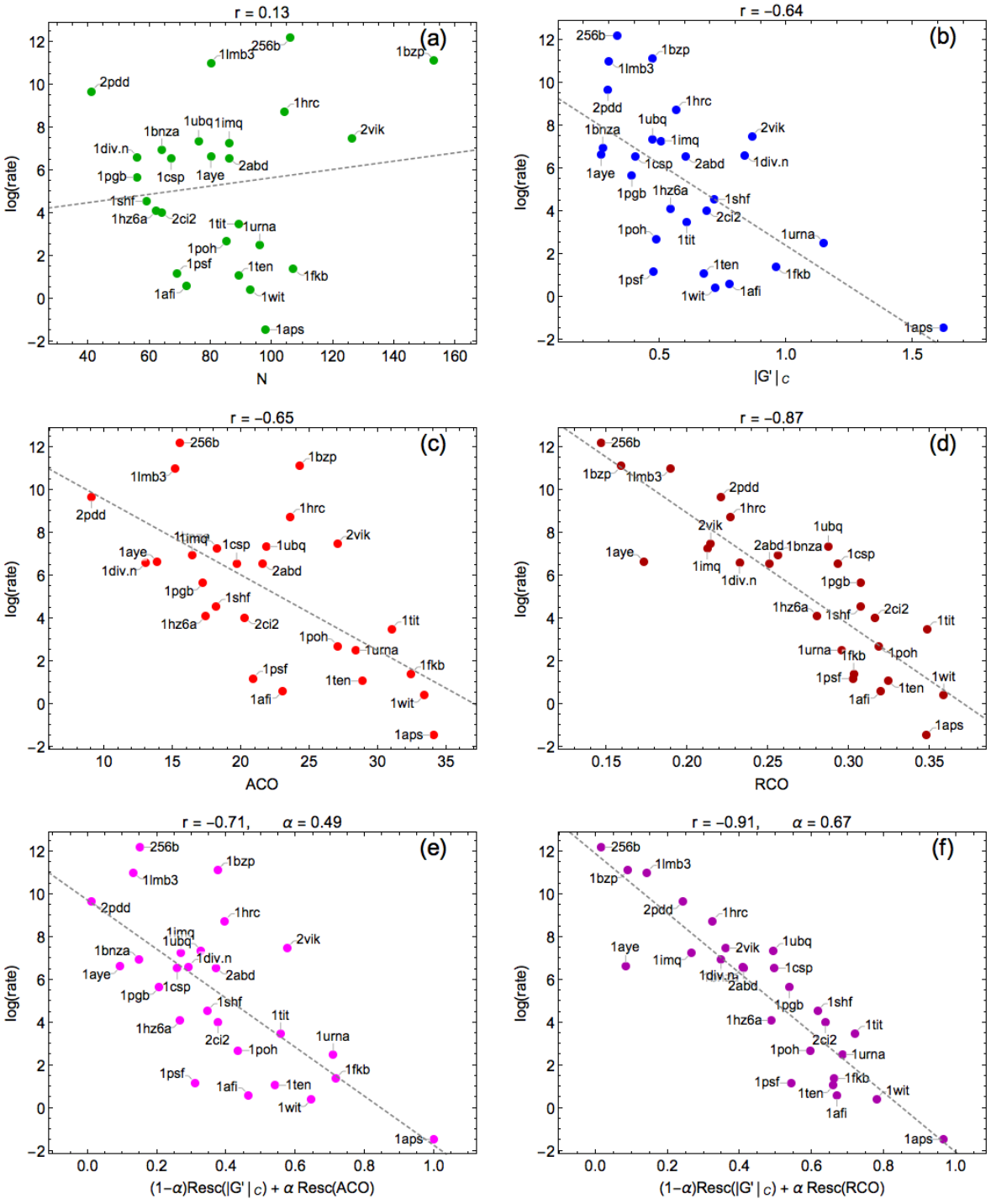

Figure 3. For two-state folders, correlation between the natural logarithm of the folding rate and (a) chain length, (b) the indicator of entanglement $|G'|_c$ proposed in this work, (c) ACO, (d) RCO, (e) a linear combination of rescaled ACO and $|G'|_c$ (see the text), and (f) a similar linear combination of rescaled RCO and $|G'|_c$. The Pearson correlation coefficient r of data is specified in all panels.

as for ACO (see figure 3(b)).

We next consider how one can combine linearly the predicting power of $|G'|_c$ and

the contact order descriptors to achieve correlations with experimental folding rates that are better than the individual cases. To work with homogeneous quantities, acquiring values between 0 and 1 in a data set with $k = 1, \ldots, N_p$ proteins, we rescale linearly any descriptor X_k as

$$\mathrm{Resc}\,(X_k) = \frac{X_k - X_m}{X_M - X_m}\,, \tag{7}$$

where $X_m = \min_k \{X_k\}$ and $X_M = \max_k \{X_k\}$. The Pearson correlation coefficient is then considered for the linear combination

$$(1-\alpha)\,\mathrm{Resc}(|G'|_c) + \alpha\,\mathrm{Resc}(\mathrm{ACO}) \tag{8}$$

as a function of the parameter $\alpha \in [0, 1]$ (and similarly for RCO). Note that $\alpha = 0$ corresponds to consider only $|G'|_c$ while $\alpha = 1$ represents the ACO.

Results for the values of α that yield the higher quality correlations are shown in figure 3(e) for ACO ($r = -0.71$) and in figure 3(f) for RCO ($r = -0.91$). In both cases the performance is increased by combining the contact-order predictor with the novel entanglement-based predictor. The optimal values of α to be used in the mixing, $\alpha = 0.49$ for ACO and $\alpha = 0.67$ for RCO, closer to 0.5 than to 1, show that the structural properties captured by $|G'|_c$ are at least in part complementary to those captured by contact order in the task of predicting folding rates for two-state folders.

Since the increment in the correlation is related to the amount of independent information contained in either descriptors, it important to measure the extent to which the novel descriptor $|G'|_c$ and the other descriptors are mutually correlated. In figures 4(a), 4(b) we show the correlation of $|G'|_c$ with respectively the chain length and ACO. Structural entanglement, as measured by $|G'|_c$, is only slightly correlated with chain length ($r = 0.27$), whereas it exhibits a stronger correlation with the ACO ($r = 0.67$). Some correlation between $|G'|_c$ and ACO should have been expected, since both quantities correlate well with the experimental results (see Discussion).

Finally, we investigate the robustness of the correlation with the folding rates of two-state folders for two of the considered descriptors, $|G'|_c$ and ACO. We perform a leave-one-out analysis by removing, in turn, each single entry from the data sets. The Pearson correlation coefficients computed in all such cases are ranked according to their absolute value in figure 5(a) for $|G'|_c$ and figure 5(b) for ACO. As expected for a not so large number of data, the correlation coefficient can be very sensitive to the removal of single entries from the data set. In particular, the presence of 1aps is found to be crucial for the good performances of both descriptors, much more so for $|G'|_c$, whereas the removal of 1bzp greatly boosts the performance of ACO.

3.2. Multistate proteins

In analogy with figure 3, for multistate folders we plot in figure 6 the correlations between several quantities and the natural logarithm of the experimentally measured folding rate. As already known [57], the correlation of ACO with folding rate is good (Pearson

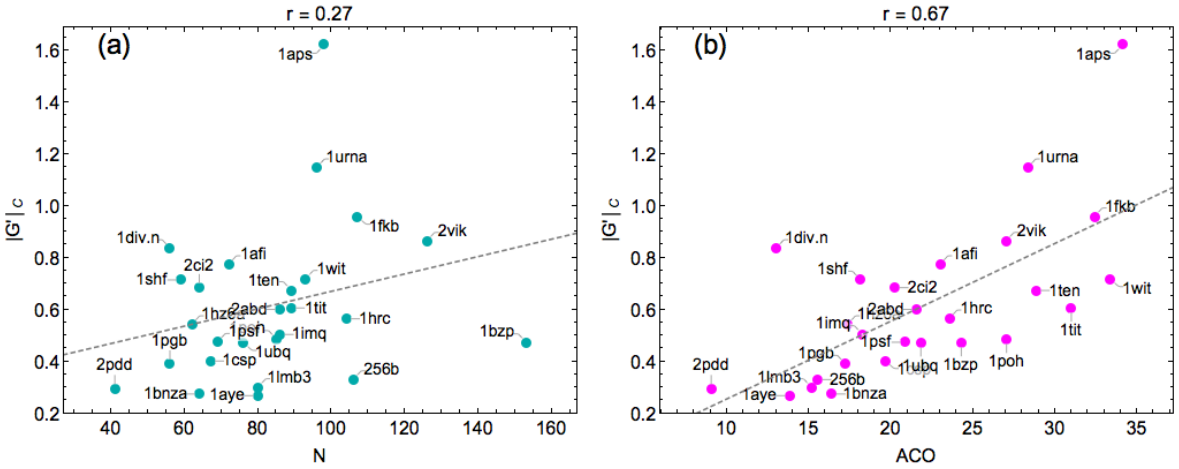

Figure 4. For two-state folders, correlation between $|G'|_c$ and (a) chain length and (b) ACO.

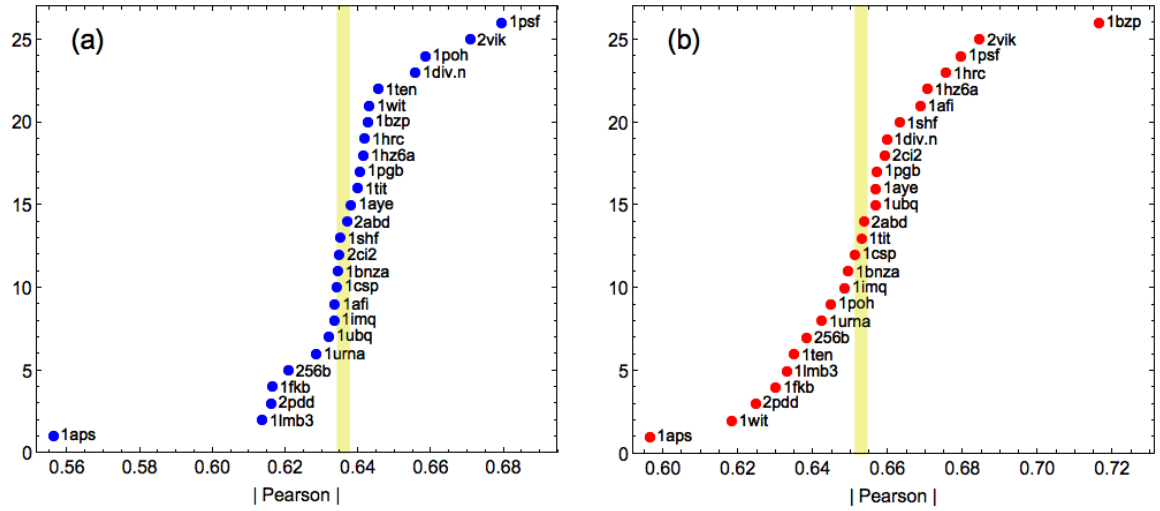

Figure 5. For two-state folders: (a) absolute vale of the Pearson correlation coefficient of log-rate vs. $|G'|_C$ obtained by removing one protein from the database, ranked from the lowest to the highest value. The vertical yellow line indicates the coefficient obtained with the full database. (b) The same for ACO. In both plots, the protein with the lowest value is the most important for obtaining the Pearson coefficient of the full data set, while the protein with the highest value is the one which spoils mostly the global value.

correlation coefficient $r = -0.78$) for multistate folders (see figure 6(c)), whereas the best performance is achieved by chain length ($r = -0.86$, figure 6(a)). Contrary to the case of two-state folders, the correlation of RCO with the folding rate of multistate folders is very poor, even reversing its sign ($r = 0.29$, see figure 6(d)). For $|G'|_c$ the correlation is again almost as good as for ACO ($r = -0.74$, see figure 6(b)).

We next consider how much the linear combination of $|G'|_c$ with either ACO or RCO can improve the correlation with folding rates of the contact-order descriptors. In all cases, we rescale linearly the descriptors X according to (7) and the Pearson

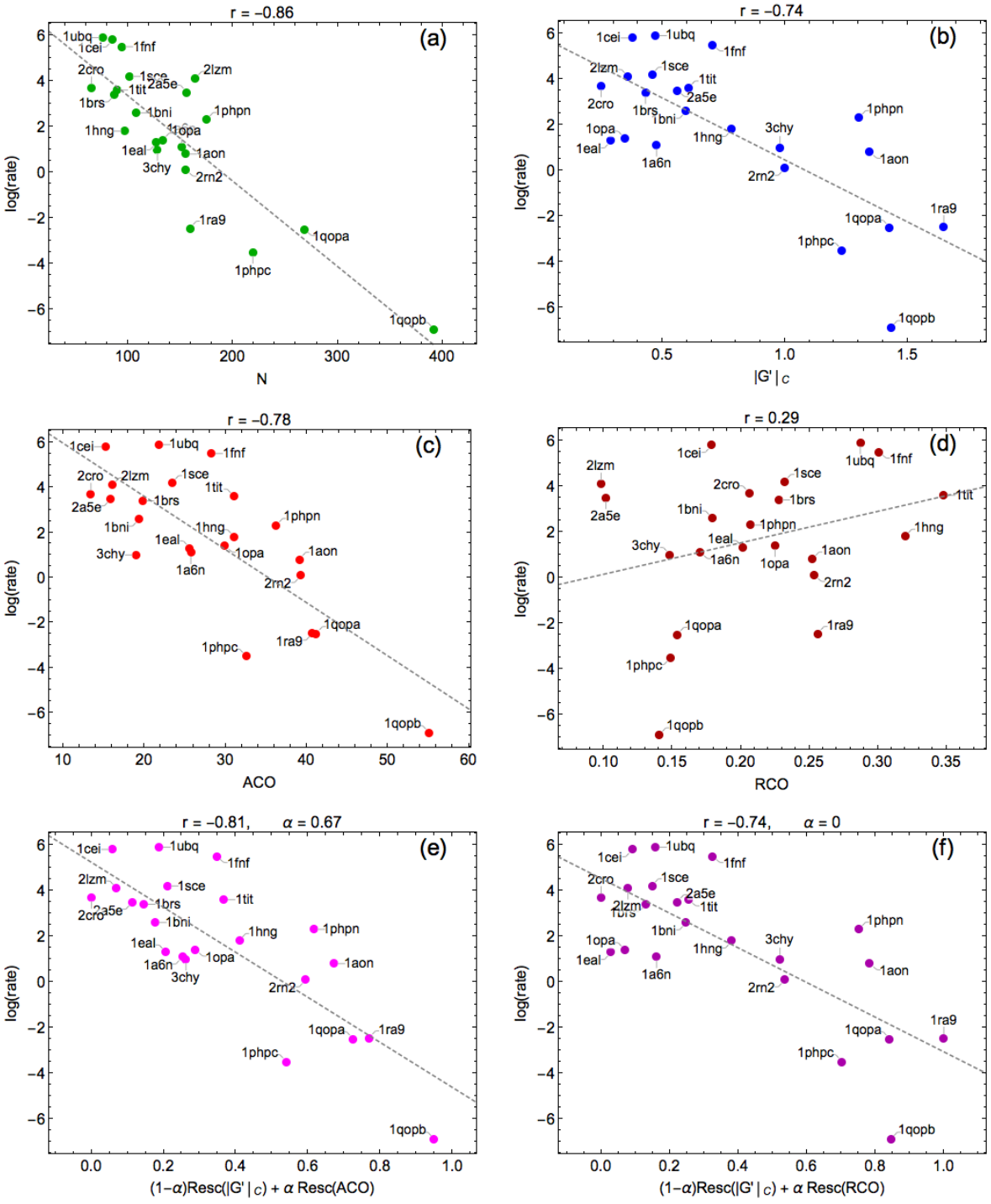

Figure 6. As in figure 3, but for multistate folders.

correlation coefficient is then again considered for the linear combination (8).

The results for the values of α that yield the higher quality correlations are shown in figure 6(e) for ACO ($r = -0.81$) and in figure 6(f) for RCO ($r = -0.74$). In both cases the performance is increased by combining the contact-order predictor with the novel entanglement-based predictor. The optimal values of α to be used in the linear

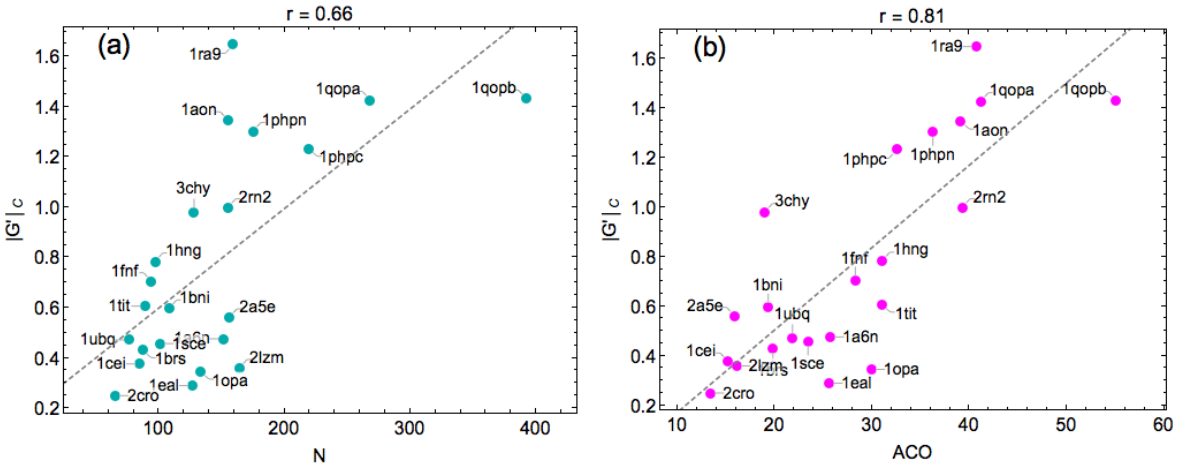

Figure 7. For multistate folders, correlation between $|G'|_c$ and (a) chain length and (b) ACO.

combination is $\alpha = 0.67$ for ACO, confirming that the structural properties captured by $|G'|_c$ are complementary to those captured by contact order in the task of predicting folding rates, also in the case of multistate folders. The linear combination of $|G'|_c$ with RCO is instead illustrative of the case when one of the combined predictors ($|G'|_c$) is much more informative than the other, as evident from the optimal value $\alpha = 0$.

We finally show the correlation of the novel descriptor $|G'|_c$ with other descriptors, such as chain length (figure 7(a)) and ACO (figure 7(b)). Structural entanglement, as measured by $|G'|_c$, is significantly correlated with chain length ($r = 0.66$), whereas it exhibits a good correlation with ACO ($r = 0.81$).

As for the two-state folders, we conclude with a leave-one-out analysis. Figure 8(a) suggests that the performance of $|G'|_c$ is robust, in this case more than that of ACO, which is very sensible to the presence of the protein 1qopb in the data set, as shown in figure 8(b) (the removal of 1qopb from the data set would cause a drop to $r = 0.68$).

4. Discussion

Data sets and performance robustness are sensible issues in the context of folding rate predictions. Different authors typically considered different data sets [57]. Moreover, folding rates may have been measured for the same protein in different conditions. Importantly, given the small number of proteins for which an experimental measure of the folding rate is available, the performance of different predictors can be very sensitive on the presence or absence of even single proteins in the data set.

As a matter of fact, several, yet not all, authors considered separate data sets for two-state and multistate protein folders [57]. The folding of two-state proteins to the native state is a cooperative process characterized by a unique time scale, whose inverse is the folding rate. Multistate proteins exhibit one or more intermediate states in the folding process, resulting in multiple relaxation times being measured. The folding rate

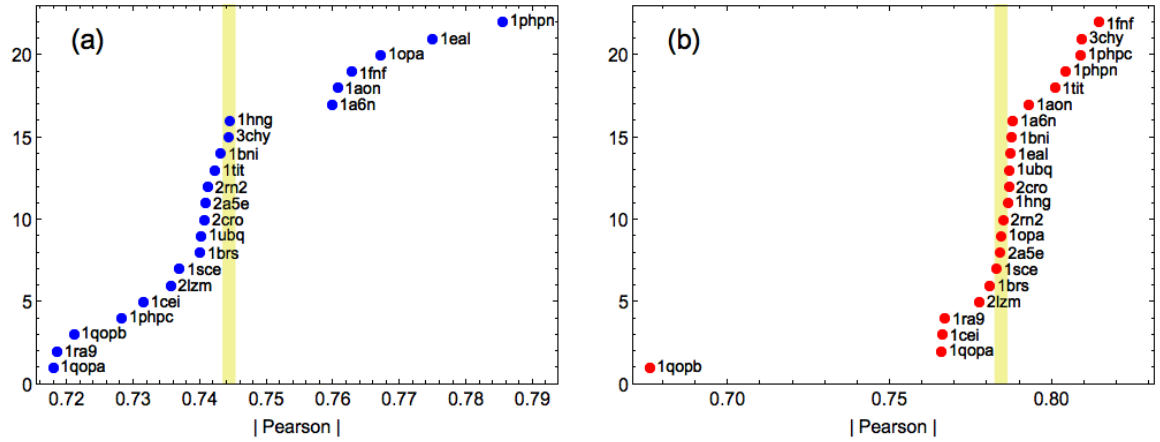

Figure 8. For multistate folders: (a) absolute vale of the Pearson correlation coefficient of log-rate vs. $|G'|_C$ obtained by removing one protein from the database, ranked from the lowest to the highest value. The vertical yellow line indicates the coefficient obtained with the full database. (b) The same for ACO.

of multistate proteins is associated to the final relaxation to the native state [55]. A given protein may switch from two-state to multistate folding behaviour upon changing experimental conditions, so that it can be found in data sets for both categories (see section 2). Other authors [53, 58, 59] considered merged data sets with both two-state and multistate folders, with the goal of testing general theories of protein folding that predict how the folding rate would increase with the chain length, for single-domain proteins.

Our choice here is to keep separate the data sets corresponding to the two different protein classes. The dependence of folding rate on simple descriptors, such as chain length or RCO, is in fact very different in the two classes, see figure 3(a) and (d) vs. figure 6(a) and figure 6(d). We are not aware of any effective explanation of this puzzling behaviour.

More generally, the data shown in figure 3 and figure 6 confirm what found in a previous study [57]. The RCO is a very good predictor for the folding rate of two-state folders but is very poor for multistate ones, to the extent of reversing the sign of the correlation coefficient in the latter case. The converse is true for chain length that is a very good predictor for the folding rate of multistate folders but essentially does not correlate with the folding rate of two-state folders. The ACO is instead a more robust predictor that perform reasonably well for both data sets and it embodies the "surprising simplicity" that characterizes protein folding [1]. The more topologically complex the network of contacts in the native structure, as measured in the case of ACO by the average sequence separation of contacting residue pairs, the longer it takes to fold to that structure.

Several other descriptors were introduced in the past to capture the topological complexity of the network of contacts in the native structures better than ACO. These include long range order, total contact distance, cliquishness, logCO, number of non

local contacts, and number of geometric contacts [18, 19, 21, 22, 23, 24, 56]. All such descriptors are based on the notion of pairwise residue contacts. The performances of the different descriptors in predicting folding rates vary somewhat depending on the considered data sets [57]. It is fair to state that most of the cited predictors, including ACO, exhibit overall similar performances.

In this work, in fact, our main focus was not to establish which is the best predictor of folding rates nor to build such an optimal algorithm. We instead introduced a novel descriptor, the maximum intrachain contact entanglement $|G'|_c$, not directly related to the contact order. It is rather based on the concept of the mutual entanglement between two portions of a protein chain that is inherently associated to contact formation. We then showed that $|G'|_c$ can be used to predict folding rates with a performance comparable to the one achieved by ACO, for both data sets of two-state and multistate folders, as shown in figure 3(b),(c) and in figure 6(b),(c). Figure 5 further shows that ACO performance is slightly more robust for two-state folders, whereas figure 8 shows that $|G'|_c$ performance is instead more robust for multistate folders.

Our main message is related to the complementary nature of the $|G'|_c$ and ACO descriptors in capturing the topological complexity of protein native structures at two different levels. Not only the separation along the sequence between pairs of contacting residues is important, but also the possible entanglement of other chain portions with the loop connecting two contacting residues (see figure 1) plays a relevant role. Note that the former feature refers to the topological complexity of the network of native contacts, whereas the latter relates to the topology of the protein chain as a curve in the three-dimensional space. The explicit consideration of the three-dimensional topological properties of the native structures represents one of the main novelties of the present paper.

The two descriptors $|G'|_c$ and ACO are indeed correlated for both data sets of two-state and multistate folders, as shown in figure 4(b) and in figure 7(b). However, a linear combination of $|G'|_c$ and ACO, after proper rescaling of the two quantities achieves a better performance than ACO alone, or $|G'|_c$ alone, for both data sets, as shown in figure 4(e) and in figure 7(e). Similarly, a linear combination of $|G'|_c$ and RCO, after proper rescaling of the two quantities achieves a better performance than RCO alone for two-state folders, as shown in figure 4(f).

The definition we chose for the entanglement descriptor requires that one of the two subchains is looped, and hence that the structure identified by a high $|G'|_c$ resembles a lasso. This definition is in line with similar analyses in the literature. Nevertheless, as Gauss double integrals do not require the looping condition for any of the two subchains, more flexible descriptors may be put forward to assess the degree of entanglement. General Gauss double integrals thus constitute a method for future characterisations of the topological complexity of single proteins, which may find applications also in contexts different from the prediction of folding rates considered in this work.

As an illustrative example, we conclude our discussion by reporting that $|G'|_c = 1.21$ for the Human single-domain protein K-Ras (167 residues as a single chain in the PDB

ID 3GFT). K-Ras fluctuations in its native ensemble were recently shown by atomistic simulations to exhibit anomalous non-ergodic kinetics over several decades [60]. The same behaviour was reported for two larger multi-domain proteins. The observed kinetics was quantitatively well described by a continuous time random walk with a heavy-tailed waiting time distributions [61].

The maximum intrachain contact entanglement found for K-Ras is higher than the one expected for single-domain proteins with similar length based on a linear interpolation for the data sets considered in this paper (see figures 4(a),7(a)). It may be then appealing to speculate whether the locking of the protein chain into conformations that are entangled, according to the $|G'|_c$ descriptor, or to similar ones, could play some role in shaping the non-ergodic kinetics described above. Highly entangled conformations could in fact explain the presence of deep traps in the energy landscape where the protein chain would remain stuck for extended periods of time. Clearly, such hypothesis needs to be thoroughly validated by further studies. The generalization of the maximum intrachain contact entanglement indicator to the case of multi-domain proteins should also be considered in this respect.

Acknowledgments

We thank Stu Whittington for having introduced us into the exciting subject of topology of random polymers and for having been an inspiring teacher and friend. We also thank an anonymous referee for pointing out reference [60].

[1] Baker D 2000 A surprising simplicity to protein folding *Nature* **405** 309
[2] Dokholyan NV, Lewyin L, Ding F and Shakhnovich EI 2002 Topological determinants of protein folding *Proc. Natl. Aca Sci. USA* **99** 8637
[3] Micheletti C, Banavar JR, Maritan A and Seno F 1999 Protein structures and optimal folding from a geometrical variational principle *Phys. Rev. Lett.* **82** 3372
[4] Munoz V and Eaton WA 1999 A simple model for calculating the kinetics of protein folding from three-dimensional structures *Proc. Natl. Aca Sci. USA* **96** 11311
[5] Galitzkaya OV and Finkelstein AV (1999) A theoretical search for folding/unfolding nuclei in three-dimensional protein structures *Proc. Natl. Aca Sci. USA* **96** 11299
[6] Alm E and Baker D (1999) Prediction of protein-folding mechanisms from free-energy landscapes derived from native structure *Proc. Natl. Aca Sci. USA* **96** 11305
[7] Park B and Levitt M 1996 Energy functions that discriminta X-ray and near native folds from well constructed decoys *J. Mol. Biol.* **258** 367
[8] Samudrala R and Levitt M 1998 Decoy 'R' Us: a database of incorrect protein conformations to improve protein structure prediction *Protein Sci.* **9** 11305
[9] Buchete NV, Straub JE and Thirumali D 2006 Developments of novel statistical potentials for protein folding recognition *Curr. opin. Struct. Biol.* **14** 225
[10] Seno F, Trovato A, Banavar JR and Maritan A 2008 Maximum entropy approach for deducing amino acid interactions in proteins *Phys. Rev. Lett.***100** 078102
[11] Cossio P, Granata D, Laio A, Seno F and Trovato A 2012 A simple and efifcient potential for scoring ensembles of protein structures *Sci. Rep.* **2** 351
[12] Hoang TX, Marsella L, Trovato A, Seno F, Banavar JR and Maritan A 2006 Common attributes

of of native-state structures of proteins, disordered proteins and amyloid *Proc. Natl. Aca Sci. USA* **103** 6883

[13] Zang Y, Hubner I , Arakaki A, Shahnovich E and Skolnick J 2006 On the origin and highly likely completeness of single domain protein structures *Proc. Natl. Aca Sci. USA* **103** 2605

[14] Cossio P, Trovato A, Pietrucci F, Seno F, Maritan A and Laio A 2010 Exploring the universe of protein structures beyond the Protein data Bank *PLoS Comput Biol* **6**, e100957

[15] Plaxco KW, Simons KT and Baker D 1998 Contact order, transition state placement and the refolding rates of single domain proteins *J. Mol. Biol.* **277** 985

[16] Plaxco KW, Simons KT, Ruczinski, I and Baker D 2000 Topology, stability, sequence and length: defining the determinants of two-state protein folding kinetics *Biochemistry* **39** 11177

[17] Cieplak M and Hoang TX 2003 Universality classes in folding times of proteins *Biophysical Journal* **84** 475

[18] Gromiha MM and Selvaraj S 2001 Comparison between long-range interaction and contact order in determining the folding rate of two state proteins: application of long range order to folding rate prediction *J. Mol. Biol.* **310** 27

[19] Makarov DE, Keller CA, Plaxco KW and Metiu H 2002 How the folding rate constant of simple, single-domain proteins depends on the number of native contacts *Proc. Natl. Aca Sci. USA* **99** 3535

[20] Makarov DE and Plaxco KW 2003 The topomore search model: a simple, quantitative theory of two-state protein folding kinetics *Protein Sci* **12** 17

[21] Zhou H and Zhou Y 2002 Folding rate prediction using total contact distance *Biophysics J* **82** 458

[22] Micheletti C 2003 Prediction of folding rates and transition state placement from native state geometry *Proteins* **51** 74

[23] Gong HP, Isom DG, Srinivasan R and Rose GD 2003 Local secondary structure content predicts folding rates for simple, two-state proteins *J. Mol. Biol.* **327** 1149

[24] Dixit PD and Weikl TR 2006 A simple measure of native stayte topology and chain connectivity predicts the folding rates of two state proteins with and without crosslinks *Proteins: Structure, Function and Bioinformatics* **64**, 193

[25] Fuller FB 1971 The writhing number of a space curve *Proc. Natl. Acad. Sci. U S A* **68**, 815

[26] Levitt M 1983 Protein folding by restrained energy minimization and molecular dynamics *Journal Molecluar Biology* **170**, 723

[27] Arteca GA and Tapia O 1999 Characterization of Fold Diversity among Proteins with the Same Number of Amino Acid Residues *J. Chem. Inf. Comput. Sci.* **39**, 642

[28] Rogen P and Fain B 2004 Automatic classification of protein structure by using Gauss integrals. *Proc. Natl. Acad. Sci. USA* **100**, 119-124

[29] Taylor WR 2002 A deeply knotted protein structure and how it might fold. *Nature* **406**, 916

[30] Virnau P, Mirny L and Kardar M 2006 Intricate knots in proteins: Functin and Evolution. *PLoS Comput Biol* **2**, 1074

[31] Lua RC and Grosberg AY 2006 Statistics of knots, geometry of conformations, and evolution of proteins. *PLoS Comput. Biol.* **2**, e45

[32] Sulkowska JI, Sulkowski P, Szymczak P and Cieplak, M Stabilizing effect of knots on proteins 2008 *Proc. Natl. Acad. Sci. USA* **105**, 19714

[33] Mallam AL, Morris ER and Jackson SE 2008 Exploring knotting mechanisms in protein folding. *Proc. Natl. Acad. Sci. USA* **105**, 18740

[34] Mallam AL, Rogers JM and Jackson SE 2010 Experimental detection of knotted conformations in denatured proteins. *Proc. Natl. Acad. Sci. USA* **107**, 8189

[35] Jamroz M et al. 2014 KnotProt: a database of proteins with knots and slipknots. *Nucl. Acids Res.* **43(DI)**, D306-D314

[36] Lim NCH and Jackson SE 2015 Molecular knots in biology and chemistry. *J. Phys. Cond. Mat.* **27**, 354101 (

[37] Baiesi M, Orlandini E, Trovato A and Seno F 2016 Linking in domain-swapped protein dimers

Sci. Rep. **6** 33872
[38] Caraglio M, Micheletti C and Orlandini E 2017 Physica Links: defining and detecting inter-chain entanglement *Sci. Rep.* **7** 1156
[39] Zhao Y, Chwastyk M and Cieplak M 2017 Structural entanglements in protein complexes *J. Chem. Phys.* **146** 3Art. N. 225102
[40] Dabrowski-Tumanski P, Jarmolinska AI, Niemyska W, Rawdon EJ, Millett KC and Sulkowska JI 2017 LinkProt: a database collecting information about biological links *Nucleic acids research* **45(D1)** D243
[41] Dabrowski-Tumanski P and Sulkowska JI 2017 Topological knots and links in proteins *Proc. Natl. Acad. Sci. USA* **114**, 3415
[42] Niemyska W, Dabrowski-Tumanski P, Kadlof M, Haglund E, Sulkowski P and Sulkowska JI 2017 Complex lasso: new entangled motifs in proteins *Sci. Rep.* **6** 36895
[43] Rolfsen D 1976 Knots and Links. *Mathematics Lecture Series* **7**, (Publish or Perish, Inc., Houston, Texas).
[44] Doi M and Edwards SF 1986 The Theory of Polymer Dynamics (Oxford University Press, USA)
[45] Orlandini E, Tesi MC and Whittington SG 2000 Polymer entanglement in melts *J. Phys. A: Math. Gen* **33**, L181
[46] Orlandini E and Whittington SG 2004 Entangled polymers in condensed phases *J. Chem. Phys.* **121**, 12094
[47] Panagiotou E, Kroger M and Millett KC 2013 Writhe and mutual entanglement combine to give the entanglement length *Phys. Rev. E* **88**, 062604
[48] Orlandini E, Janse van Rensburg EJ, Tesi MC and Whittington SG 1994 Random linking of lattice polygons *J. Phys. A: Math. Gen.* **27**, 335
[49] Arsuaga J, Blackstone T, Diao Y, Karadayi E and Saito M Linking of uniform random polygons in confined spaces *J. Phys. A: Math. Theo.* **40** 1925
[50] Marko JF 2011 Scaling of Linking and Writhing Numbers for Spherically Confined and Topologically Equilibrated Flexible Polymers *J. Stat. Phys.* **142** 1353
[51] D'Adamo G, Orlandini E and Micheletti C 2017 Linking of Ring Polymers in Slit-Like Confinement *Macromolecule* **50** 1713
[52] Michieletto D, Marenduzzo D, Orlandini E, Alexander GP and Turner MS 2014 Threading Dynamics of Ring Polymers in a Gel *Macro Letters* **3**, 255
[53] Ivankov DN, Garbuzynskiy SO, Alm E, Baker D and Finkelstein AV 2003 Contact order revisited: influence of protein size on the folding rate *Protein Science* **12** 2057
[54] Grantcharova V, Alm EJ, Baker D and Horwich AL 2001 Mechanisms of protein folding *Curr Opin Struct Biol* **11** 70
[55] Kamagata K, Arai M and Kuwajima K 2004 Unification of the Folding Mechanisms of Non-two-state and Two-state Proteins *J Mol Biol* **339** 951
[56] Ouyang Z and Liang J 2008 Predicting protein folding rates from geometric contact and amino acid sequence *Prot Sci* **17** 1256
[57] Weikl TR 2008 Loop-closure principles in protein folding *Arch Biochem Biophys* **469** 67
[58] Li MS, Klimov DK and Thirumalai D 2004 Thermal denaturation and folding rates of single domain proteins: size matters *Polymer* **45** 573
[59] Naganathan AN and Munoz V 2005 Scaling of Folding Times with Protein Size *J Am Chem Soc* **127** 480
[60] Hu X, Hong L, Smith MD, Cheng X and Smith, JC 2016 The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time *Nat Phys* **12** 171
[61] Schulz JHP, Barkai E and Metzler R 2014 Aging renewal theory and application to random walks *Phys Rev X* **4** 011028