FEBS Letters 589 (2015) 2179–2185

ELSEVIER

FEBS Letters

journal homepage: www.FEBSLetters.org

Co-evolutionary constraints of globular proteins correlate with their folding rates

Saurav Mallik, Sudip Kundu *

Department of Biophysics, Molecular Biology and Bioinformatics, University of Calcutta, India
Center of Excellence in Systems Biology and Biomedical Engineering (TEQIP Phase-II), University of Calcutta, India

ARTICLE INFO

Article history:
Received 18 March 2015
Revised 9 June 2015
Accepted 24 June 2015
Available online 8 July 2015

Edited by A. Valencia

Keywords:
Protein folding
Co-evolution
Relative co-evolution order
Contact order

ABSTRACT

Folding rates ($\ln k_f$) of globular proteins correlate with their biophysical properties, but relationship between $\ln k_f$ and patterns of sequence evolution remains elusive. We introduce 'relative co-evolution order' (*rCEO*) as length-normalized average primary chain separation of co-evolving pairs (CEPs), which negatively correlates with $\ln k_f$. In addition to pairs in native 3D contact, indirectly connected and structurally remote CEPs probably also play critical roles in protein folding. Correlation between *rCEO* and $\ln k_f$ is stronger in multi-state proteins than two-state proteins, contrasting the case of contact order (*co*), where stronger correlation is found in two-state proteins. Finally, *rCEO*, *co* and $\ln k_f$ are fitted into a 3D linear correlation.

1. Introduction

A number of studies are performed in recent years to correlate folding rates ($\ln k_f$) of globular proteins with their biophysical properties; these include length [1], sequence composition [2], secondary structural makeup [3], 3D topology [4,5] etc. Small proteins generally fold faster than large ones, which results a negative correlation (−0.64) between proteins' length and $\ln k_f$ [1]. Folding rate also depends on the amino acid composition, resulting 96% correlation between the two parameters [2]. The secondary structural makeup, that is to be generated during folding, also negatively correlates (−0.82) with $\ln k_f$. Further, $\ln k_f$ depends on the 3D topology of the native structure. Contact order (*co*), a measure of protein 'topology' in 3D space, is defined as the average primary chain separation of the native atomic contacts, and it negatively correlates (−0.74) with $\ln k_f$ [5].

Research interests have recently been diversified to understand the association between protein folding and evolution. Analyzing homologous sequences of proteins with known folding kinetics, Plaxco et al. [6] reported a significant correlation between the contributions of individual sequence positions (not individual amino acids) to the transition state structure. This indicated that a protein evolves by conserving the structure of its folding transition state ensemble, rather than conserving specific interactions among amino acids [6]. As a consequence, strong sequence conservation does not necessarily indicate participation in transition state ensemble [7,8]. In recent years, the effects of point mutation on the folding mechanism are also being investigated, in which point mutations are induced in small globular proteins (both conserved and non-conserved sites) to investigate consequent changes in their folding free energy as well as folding rate [9]. Parallel to experimental studies, several theoretical works predict the effect of point mutations on folding landscape [10,11]. These studies show that both conserved and non-conserved positions can alter the folding rate while mutated and the rate can vary in wide spectrum.

Mutations are random and unavoidable in the course of evolution. But the fixation of mutations is not random, but it depends on many factors, including the maintenance of folding landscape and structural integrity [12–14]. For example, if two sites are under some biophysical constraint(s), then mutation occurring at one site alters the selection pressure on the other, inducing a complementary change [15]. This evolutionary phenomenon is termed as 'co-evolution' and it is associated with a wide spectrum of biophysical constraints, including tertiary and quaternary atomic contacts as well as long-distance functional constraints [15].

Author contributions: S.M. and S.K. designed research; S.M. implemented computational methodologies, performed research and analyzed data; S.M. and S.K. wrote the paper.

* Corresponding author at: University College of Science, Technology and Engineering, University of Calcutta, 92, Acharya Prafulla Chandra Road, Kolkata 700009, India.

E-mail address: skbmbg@caluniv.ac.in (S. Kundu).

http://dx.doi.org/10.1016/j.febslet.2015.06.032
0014-5793/© 2015 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

Such coordinated reciprocal mutations during biological evolution are, therefore, fundamentally different from experimentally induced mutations. Hence, a systematic investigation is required to test whether the coordinated fashion of biological mutations has some association with folding rate.

Here we identify the intra-molecular co-evolving residue pairs (CEPs) of globular proteins by several available methods to find whether the co-evolutionary patterns correlate with their experimentally derived folding rates. We introduce a parameter: relative co-evolution order (*rCEO*), defined as the length-normalized average primary chain separation of the co-evolving pairs and identify a significant negative correlation between *rCEO* and $\ln k_f$. Our results indicate that not only CEPs in native 3D contact, but structurally remote and indirectly contacting CEPs play critical roles in protein folding as well. Finally, *rCEO* and *co* are integrated into a 3D linear correlation with $\ln k_f$. These results might be an important step in understanding the association between the folding constraints of biomolecules and their evolution.

2. Materials and methods

2.1. Protein dataset

An initial dataset of 94 proteins with experimentally determined folding rates is collected. This dataset is then filtered based on three criteria: (i) proteins for which at least 1000 homologous sequences are available (ii) the protein family must be present within at least one complete phylum, (iii) the 3D structure of at least one homolog must be experimentally determined. The final dataset of 37 bacterial proteins (25 two-state and 13 multi-state) are provided in Supplementary Table S1. In addition, we have analyzed the bacterial 30S ribosomal complex (Supplementary extended methods).

2.2. Co-evolution analysis

Homologous sequences of each protein (the PDB sequence is used as the query) are collected using protein–protein BLAST [16]; highly similar sequences (95% similarity cutoff) are removed to maintain diversity required for co-evolution analysis. We have employed a number of currently available co-evolution analysis methods [17] to estimate *rCEO* and have compared their results. Those include basic Mutual Information [18–20], DCA [21] and GREMLIN [22]. In Mutual Information (*MI*) method, the *MI* score between two positions in an alignment is given by:

$$MI(i,j) = \sum_{a,b} P(a_i, b_j) \times \log\left(\frac{P(a_i, b_j)}{P(a_i) \times P(b_j)}\right) \tag{1}$$

where $P(a_i,b_j)$ is the joint probability distribution of residues '*a*' and '*b*', located at *i*-th and *j*-th position of the MSA respectively. $P(a_i)$ and $P(b_j)$ are marginal probability distributions of residues '*a*' and '*b*'. In *MI* approach, there are several potential sources of background errors, such as small alignment size, phylogenetic effects, positions of high entropy and invariable sites [19,23]. Supplementary extended methods includes a detailed discussion on minimizing background errors. The *rcwMI* filtering approach is employed in filtering step. Each site pair score is weighted against the average score of its constituting sites [19], and the Row–Column-Weighted score *rcwMI* is defined as:

$$rcwMI(i,j) = \frac{M_{ij}}{(MI_{i.} + MI_{.j} - 2MI_{ij})/(n-1)} \tag{2}$$

where $MI_{i.}$ and $MI_{.j}$ are the summation of the *MI* values of residues *i* and *j* respectively, to all other residues in the MSA. M_{ij} is the *MI* between residues *i* and *j*. A probability density spectrum of *rcwMI* scores is generated and top hits are chosen from the subset of the entire spectrum above the one-tailed 99.9% confidence interval. The residue pairs associated with these top 0.01% *rcwMI* scores are considered as co-evolving.

In addition, two advanced methods DCA and GREMLIN are employed in our analysis. *MI* calculates the correlation of each residue pair (*i,j*) independently. In DCA method, the coupling of the pair *i* and *j* is computed taking into account the effect of other positions in the alignment. A detailed description and implementation of this method can be found in Ref. [21]. GREMLIN integrates sequence co-evolution and structural context information using a pseudo-likelihood approach, allowing accurate contact predictions from fewer homologous sequences. A detailed description of GREMLIN approach can be found in Ref. [22].

2.3. Estimating contact order

The absolute contact order (*co*) of a protein structure is defined as [5]:

$$co = \frac{1}{n_c} \sum_{i>j} \Delta(i,j)|s_i - s_j| \tag{3}$$

where n_c is the total number of contacts, s_i and s_j are the sequence positions of residues *i* and *j*, and $\Delta(i,j)$ is the selection criteria that includes *i* and *j* into analysis only if they are in contact and if $|i-j| > 4$. This $|i-j| > 4$ criterion ensures that the contacts included in *co* estimation are directly associated with 3D topology of the proteins, rather than secondary structures. If any two atoms from two different amino acids (*i* and *j*) are within a cutoff distance (5 Å), the amino acids are considered to be connected.

2.4. Estimating relative co-evolution order

We introduce a parameter, termed as the relative co-evolution order (*rCEO*) defined as:

$$rCEO = \frac{1}{L \times n_{CEP}} \sum_{i>j} \Delta(i,j)|s_i - s_j| \tag{4}$$

where *L* is length of the amino acid chain, n_{CEP} is the number of CEPs, s_i and s_j are the sequence positions of residues *i* and *j* and $\Delta(i,j)$ is the selection criteria that includes *i* and *j* into analysis if they are co-evolving and if $|i-j| > 4$.

2.5. Classifying CEPs according to 3D contacts

Co-evolution analysis reveals two types of CEPs, based on their 3D contacts. If any two atoms from two different amino acids are within a cutoff distance (5 Å), the amino acids are considered to be in direct physical contact; otherwise they are not in direct contact. The second group is further classified into two sub-groups: (i) structurally remote CEPs and (ii) CEPs in indirect physical contact (if A contacts with both B and C, then B and C are in indirect contact). The *rCEO* estimated from these four classes are denoted as, $rCEO\langle\text{dc}\rangle$, $rCEO\langle\text{nc}\rangle$, $rCEO\langle\text{sr}\rangle$ and $rCEO\langle\text{ic}\rangle$, respectively. In addition, the method used for co-evolution analysis is also mentioned, whenever relevant (e.g., for *rCEO* estimated in MI method, using directly contacting CEPs, we use) $rCEO\langle\text{dc/MI}\rangle$.

3. Results and discussions

3.1. Correlation between rCEO and $\ln k_f$ is exclusive to co

Co-evolution is generally observed between sequence pairs those are biophysically constrained [15]. A high value of the relative co-evolution order (*rCEO*) implies that there are several long-range dependencies between the sequence positions of the corresponding protein. We begin our analysis including only the non-contacting CEPs in *rCEO* estimation ($rCEO\langle \mathrm{nc} \rangle$). This analysis excludes all the CEPs contributing to contact order (*co*) estimation. The Pearson correlation between experimentally determined folding rate ($\ln k_f$) and respective $rCEO\langle \mathrm{nc} \rangle$ is, $r_{\ln k_f | rCEO\langle \mathrm{nc/MI} \rangle} = -0.65$ in MI approach, supported by a strong statistical significance $P < 10^{-6}$ (Fig. 1A). In DCA and GREMLIN, this correlation is slightly lower ($r_{\ln k_f | rCEO\langle \mathrm{nc/DCA} \rangle} = -0.51$, $P < 10^{-3}$; $r_{\ln k_f | rCEO\langle \mathrm{nc/GREMLIN} \rangle} = -0.54$ $P < 10^{-3}$)—although the non-contacting CEPs include a much smaller proportion (8% and 11%, respectively) of all the CEPs predicted in these methods—compared to MI. A strong negative correlation (−0.74) between $\ln k_f$ and *co* is already established in previous studies [4,5]. Given this information, the correlation between $\ln k_f$ and $rCEO\langle \mathrm{nc} \rangle$ indicates that the folding rate of globular proteins has some correlation with the intra-molecular co-evolutionary pattern as well, which is exclusive to contact order.

3.2. Both physically contacting and structurally remote CEPs play critical roles in folding

The physically contacting CEPs ($rCEO\langle \mathrm{dc} \rangle$ group) constitute a significant fraction of the entire CEP set (19% in MI, 79% in GREMLIN and 92% in DCA). An example of co-evolving pairs and native contacts for a globular protein is shown in Supplementary Fig. S2. When we include the $rCEO\langle \mathrm{dc} \rangle$ group CEPs with the $rCEO\langle \mathrm{nc} \rangle$ group CEPs, the correlation with $\ln k_f$ ($r_{\ln k_f | rCEO}$) elevates to −0.78 in MI ($P < 10^{-9}$) (Fig. 1B), −0.91 in DCA ($P < 10^{-15}$) and −0.85 in GREMLIN ($P < 10^{-11}$) (Table 1). This clearly indicates that both the CEPs—physically contacting and non-contacting—play significant roles in protein folding.

Now, to test whether this elevation of correlation is due to some statistical association between *rCEO* and, *co* we investigate the statistical associations of, *rCEO*, $rCEO\langle \mathrm{dc} \rangle$ and $rCEO\langle \mathrm{nc} \rangle$ with *co*. In each case, we generate a scatter plot of the respective parameter with *co* (Supplementary Fig. S3). Considering *a priori* linear correlations, we estimate the lower and upper confidence limits (LCL/UCL) under 95% statistical confidences. The possibility of the respective parameter and *co* being predictors of each other (i.e. correlated to each other) is tested under a null hypothesis of complete association. If a data-point in the scatter plot is located within the confidence band, it is considered as a successful prediction; otherwise, it is a false prediction. We randomly pick up 10 data-points at a time and estimate the percent of them being a successful prediction; this process is repeated 1000 times. For MI method, this analysis rejects the null hypothesis of complete association between $rCEO\langle \mathrm{MI} \rangle$ and *co* with $P < 10^{-7}$ (for DCA and GREMLIN,) $P < 0.05$ and that between $rCEO\langle \mathrm{nc/MI} \rangle$ and *co* with $P < 10^{-10}$ significance (for DCA and GREMLIN,) $P < 0.001$. But for, $rCEO\langle \mathrm{dc} \rangle$ which is in fact a subset of, *co* the null hypothesis persists at $P > 0.1$ for MI and $P > 0.5$ for DCA and GREMLIN.

The predicted CEPs by DCA method, as an example (Fig. 2A), include approximately 70% of the total native contacts. The $rCEO\langle \mathrm{dc/DCA} \rangle$ value computed from these physically contacting CEPs exhibit −0.81 correlation with $\ln k_f$. Addition of the remaining 30% non-coevolving native contacts results a reduced correlation with $\ln k_f$ (−0.81 to −0.74). This suggests that not all pairs in native contacts, i.e. included in *co* analysis, are crucial for protein folding. In fact, pairs those are co-evolving as well as in native contacts probably play critical roles in protein folding. On the other hand, the $rCEO\langle \mathrm{dc/DCA} \rangle$ group CEPs constitutes 92% of all the CEPs and the remaining 8% of CEPs are structurally remote. When we add this 8% CEPs with $rCEO\langle \mathrm{dc/DCA} \rangle$ group, it results a significant elevation of correlation (−0.81 to −0.92). This depicts that structurally remote CEPs probably have some critical role in folding. The essentiality of co-evolving pairs, those are not in direct physical contact, in protein folding and stability is reported in other recent studies as well [21].

Furthermore, a majority of the structurally remote CEPs are in indirect native contact (Supplementary Fig. S4), which cannot be captured in *co* analysis. This result demonstrates the importance of indirectly connected pairs in protein folding, which we shall further discuss in the next sections. In Supplementary Fig. S5, we have presented linear regression plots of contact order and co-evolution parameters fitted with folding rates.

3.3. Comparison between two-state and multi-state proteins

We have tested whether the molecular nature of folding process has some effect on the correlation between $\ln k_f$ and *rCEO*. In all three methods, the correlation between $\ln k_f$ and *rCEO* is stronger in proteins exhibiting multi-state folding (e.g.,

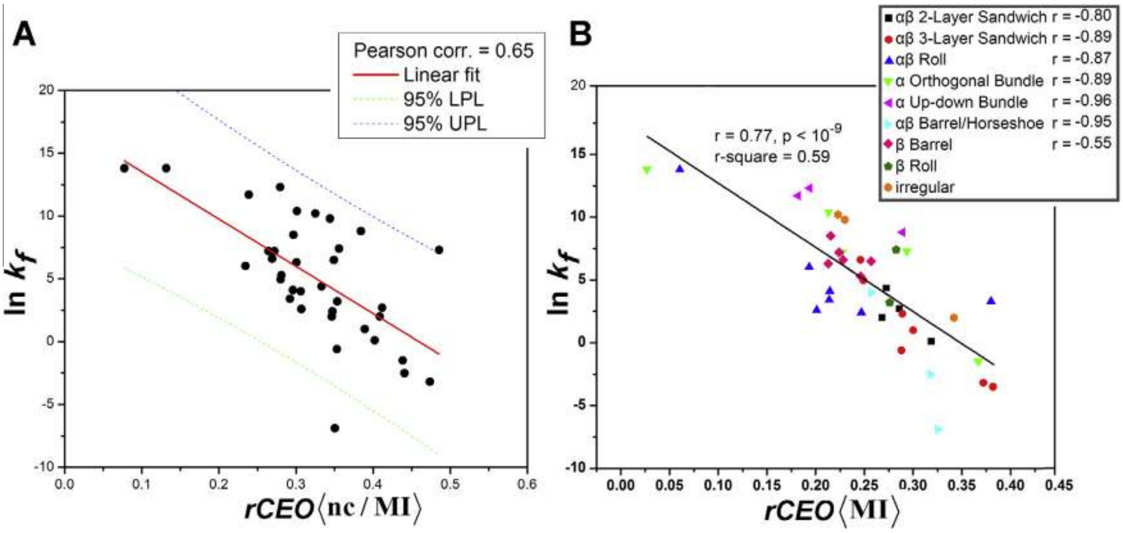

Fig. 1. (A) The linear regression fitting of $rCEO\langle \mathrm{nc/MI} \rangle$ and folding rate (red line) is shown here; the lower and upper prediction limits for this correlation under 95% statistical confidence are also highlighted. (B) The linear regression fitting of $rCEO\langle \mathrm{MI} \rangle$ and folding rate; Pearson correlations for proteins with different architectures are mentioned. Proteins having β-roll and irregular architecture were too few to infer any statistical correlation.

Table 1
Correlation coefficients of co-evolution and structure-derived parameters with protein folding rates (*rCEO* = relative co-evolution order, *co* = contact order, $\ln k_f$ = folding rate) estimated in various co-evolution analysis methods. 3D correlations do not include negative sign by definition. The *P*-values and the ρ-values in the parentheses represent the statistical significance and the respective 95% confidence intervals of the respective correlations.

Method	Correlations	Two-state proteins	Multi-state proteins	All
MI	$\ln k_f$ and *rCEO*	-0.68 ($P < 10^{-3}$) ($-0.38 < \rho < -0.85$)	-0.89 ($P < 10^{-4}$) ($-0.65 < \rho < -0.97$)	-0.78 ($P < 10^{-9}$) ($-0.61 < \rho < -0.89$)
	$\ln k_f$, *rCEO* and *co*	0.75 ($P < 10^{-4}$) ($0.57 < \rho < 0.86$)	0.93 ($P < 10^{-6}$) ($0.87 < \rho < 0.96$)	0.87 ($P < 10^{-14}$) ($0.76 < \rho < 0.93$)
GREMLIN	$\ln k_f$ and *rCEO*	-0.82 ($P < 10^{-7}$) ($-0.62 < \rho < -0.92$)	-0.87 ($P < 10^{-4}$) ($-0.59 < \rho < -0.96$)	-0.85 ($P < 10^{-11}$) ($-0.72 < \rho < -0.92$)
	$\ln k_f$, *rCEO* and *co*	0.83 ($P < 10^{-9}$) ($0.64 < \rho < 0.92$)	0.92 ($P < 10^{-6}$) ($0.73 < \rho < 0.98$)	0.88 ($P < 10^{-16}$) ($0.78 < \rho < 0.94$)
DCA	$\ln k_f$ and *rCEO*	-0.88 ($P < 10^{-9}$) ($-0.74 < \rho < -0.95$)	-0.94 ($P < 10^{-5}$) ($0.89 < \rho < 0.94$)	-0.92 ($P < 10^{-15}$) ($-0.83 < \rho < -0.95$)
	$\ln k_f$, *rCEO* and *co*	0.90 ($P < 10^{-10}$) ($-0.78 < \rho < -0.96$)	0.96 ($P < 10^{-8}$) ($0.86 < \rho < 0.99$)	0.93 ($P < 10^{-17}$) ($0.87 < \rho < 0.96$)
Contact prediction	$\ln k_f$ and *co*	-0.79 ($P < 10^{-6}$) ($-0.57 < \rho < -0.91$)	-0.64 ($P < 10^{-2}$) ($-0.24 < \rho < -0.89$)	-0.74 ($P < 10^{-6}$) ($-0.55 < \rho < -0.86$)

$r^{\text{multi-state}}_{\ln k_f|rCEO\langle\text{MI}\rangle} = -0.89$, $P < 10^{-4}$) compared to those exhibiting two-state folding ($r^{\text{two-state}}_{\ln k_f|rCEO\langle\text{MI}\rangle} = -0.68$, $P < 10^{-3}$) (Table 1). This scenario is in contrast to the case of *co*, which exhibits stronger correlation with $\ln k_f$ for two-state proteins [24]. In our dataset, we have found $r^{\text{two-state}}_{\ln k_f|co} = -0.79$ ($P < 10^{-6}$) and $r^{\text{multi-state}}_{\ln k_f|co} = -0.64$ ($P < 0.01$) (Table 1). To gain further insight, we have estimated the percentage of physically contacting as well as indirectly contacting CEPs in the two types of proteins. Physically contacting CEPs are present in a higher proportion in two-state proteins (in MI, 24%) than multi-state (15%). On the other hand, all three methods confirm that multi-state proteins include a higher proportion of indirectly contacting (e.g., in MI, 63% in two-state, 69% in multi-state) and structurally remote CEPs (in MI, 13% in two-state, 16% in multi-state).

Both co-evolving pairs and 3D contacts represent biophysical constraints. But co-evolution generally represents critical biophysical constraints, which include, but are not limited to native contacts [15]. Two-state proteins fold into their native 3D structures passing through specific folding intermediates. However, in multi-state proteins, determination of the folding intermediates and development of their secondary structural elements has been proven difficult to resolve due to the inherent complexity of the process [25]. This complexity of multi-state folding might give rise to several critical interactions those no longer persist in native state (therefore, cannot be captured in *co* analysis), but are reflected in the higher proportion of indirectly contacting and structurally remote CEPs (compared to two-state). This depicts the importance of such non-contacting CEPs in protein folding (also discussed in Ref. [21]) and also explains why *co* exhibits

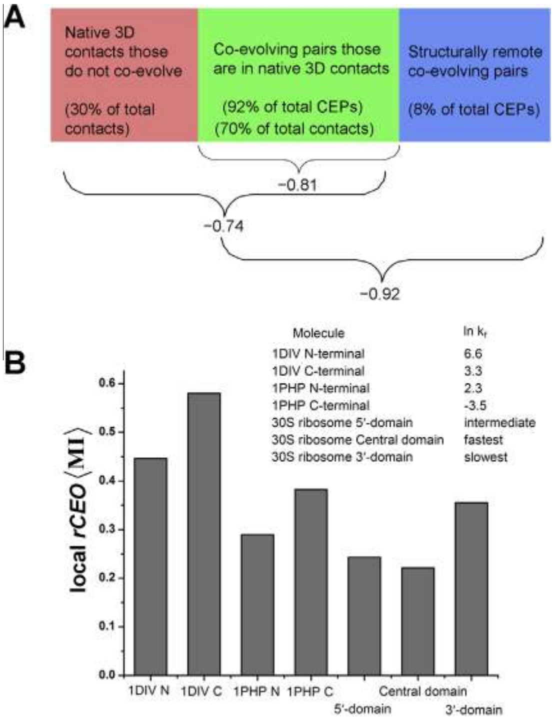

Fig. 2. (A) An illustration of the co-evolving pairs identified in DCA method, the pairs in native 3D contacts, commonality between two groups are shown. The numbers represent the correlations with folding rate of the contact/co-evolution order parameters estimated from the respective groups. (B) Comparison of local *rCEO*⟨MI⟩ at different domains of multi-domain biomolecules with qualitative/quantitative kinetic data is demonstrated (MI method). Rapidly folding domain exhibits lower relative co-evolution order.

weaker correlation with $\ln k_f$ in multi-state proteins, whereas that for *rCEO* is stronger.

3.4. Proteins of different fold classification exhibit their characteristic correlations

Since different protein folds indicate unique 3D architectures, we have tested whether the former is another determinant of the relationship between $\ln k_f$ and *rCEO*. Here, looking into the $r_{\ln k_f|rCEO}$ values among proteins with different fold categories, we observe that proteins with different architecture exhibit their characteristic $r_{\ln k_f|rCEO}$ (Fig. 1B). For example, proteins with α-orthogonal bundle architecture exhibit stronger correlations ($r_{\ln k_f|rCEO\langle \mathrm{MI}\rangle} = -0.89$, $P < 0.05$), than the β-barrel proteins ($r_{\ln k_f|rCEO\langle \mathrm{MI}\rangle} = -0.55$, $P < 0.05$).

3.5. The curious case of multi-domain proteins and 30S ribosome assembly

Two multi-domain proteins are included in our dataset, where different domains fold into their respective native structures at different rates (Supplementary dataset). In each protein, we see that the rapidly folding domain (supported by available kinetic data) exhibits lower *rCEO* (Fig. 2B).

Similar analysis is performed on the bacterial small ribosomal subunit, in which the three domains (5′, Central and 3′ domain) of the 16S ribosomal RNA fold (in this case, the molecular nature of folding is protein-guided RNA folding) at different rates [26]. Rapidly folding Central domain and kinetically trapped 3′ domain exhibit the lowest and the highest co-evolution order respectively (Fig. 2B).

3.6. 3D correlation among $\ln k_f$, rCEO and co

The correlations among, $\ln k_f$ *rCEO* and *co* are represented in a 3D surface plot in Fig. 3A. The three parameters are correlated in a linear relationship, defining a plane surface in the 3D space ($r_{\ln k_f|co|rCEO\langle \mathrm{MI}\rangle} = 0.87$, $R^2_{\ln k_f|co|rCEO\langle \mathrm{MI}\rangle} = 0.74$), supported by a strong statistical significance ($P < 10^{-14}$). This correlation is much stronger in multi-state proteins ($r^{\text{multi-state}}_{\ln k_f|co|rCEO\langle \mathrm{MI}\rangle} = 0.93$), compared to two-state proteins ($r^{\text{two-state}}_{\ln k_f|co|rCEO\langle \mathrm{MI}\rangle} = 0.75$) (Table 1). The exact mathematical relationship between the three parameters for all proteins is mentioned in the following:

$$\ln k_f = 18.43 - 36.87 \times rCEO\langle \mathrm{MI}\rangle - 0.26 \times co \tag{5}$$

The three constants are associated with 1.58, 7.42 and 0.06 standard errors respectively. The linear models for DCA and GREMLIN are included in Table 2. In MI method, $r_{\ln k_f|co|rCEO\langle \mathrm{MI}\rangle}$ is significantly elevated (87%) compared to $r_{\ln k_f|rCEO\langle \mathrm{MI}\rangle}$ (78%), while in DCA (91–93%) and GREMLIN, (85–88%) there are only minor elevations.

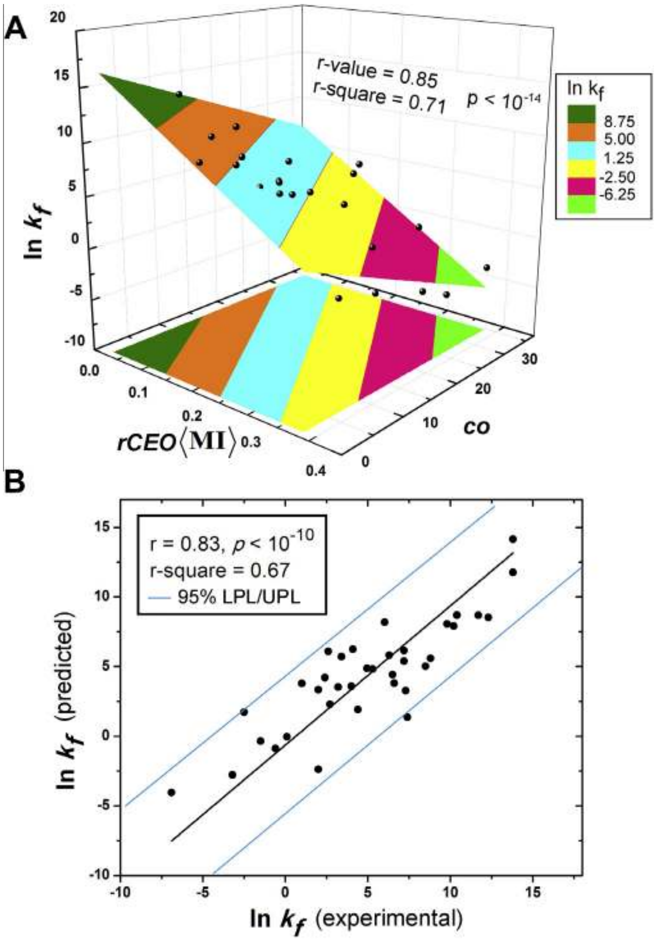

Fig. 3. (A) The 3D plane surface fitting of folding rate ($\ln k_f$), relative co-evolution order ($rCEO\langle \mathrm{MI}\rangle$) and absolute contact order (*co*) is shown as a color map surface for MI method. A projection of this surface on the *XY*-plane is shown as a contour plot. (B) The residual plot between experimentally determined and theoretically predicted folding rates is depicted here for MI method. The lower and upper prediction limits (LPL/UPL) under 95% statistical confidences are highlighted by blue straight lines.

Table 2
Statistics of 3D plane surface fitting (linear correlation) between $\ln k_f$, *rCEO* and *co* in different co-evolution analysis methods.

Method	Statistics of 3D plane surface fitting
MI	Equation: $\ln k_f = 18.43 - 36.85rCEO - 0.26co$ $R = 0.87,\ R^2 = 0.74$
GREMLIN	Equation: $\ln k_f = 19.02 - 51.01rCEO - 0.17co$ $R = 0.88,\ R^2 = 0.77$
DCA	Equation: $\ln k_f = 17.81 - 48.64rCEO - 0.10co$ $R = 0.93,\ R^2 = 0.86$

Our previous results indicate that not all native contacts are critical for folding. MI prediction includes a much higher proportion of structurally remote CEPs and the strong 3D correlation in this method reflects the crucial role of non-contacting CEPs in folding. But MI approach risks the loss of a substantial proportion of pairs those are likely critical for folding. Conversely, DCA and GREMLIN predictions likely include a majority of the connected pairs critical for folding (70% of the native contacts are included in CEPs). Therefore, when we include *co* for a 3D correlation, for DCA and GREMLIN, there are no substantial elevation of correlation (91–93% and 85–88%) compared to that in MI (78–87%). Thus, different methods with different prediction boundaries represent the importance of both physically interacting as well as non-contacting pairs in protein folding.

A residual plot of experimental and predicted $\ln k_f$ (using Eq. (5)) is shown in Fig. 3B. The standard error (*SE*) of prediction is estimated as, $SE = \sigma_{\delta_{error}}/\sqrt{n}$ where $\delta_{error} = \ln k_f|_{predicted} - \ln k_f|_{\exp erimental}$. To correct the effect of small sample size, this estimation is performed 100 times by randomly picking ten δ_{error} values and each time multiplying the *SE* estimate by the finite-population correction factor, $\sqrt{N - n/N - 1}$ where $N = 37$ and $n = 10$. The average of this population, $SE_{corrected} = 0.26$ gives the accurate standard error of the prediction in MI approach. In DCA and GREMLIN, this value is 0.21 and 0.22, respectively. We have further tested the prediction accuracy of our linear model using only a fraction of the proteins to 'train' the model and another smaller fraction to 'test' the prediction. If we randomly choose the 'training' set 1000 times, taking 20 proteins out of 36 at each cycle, the average $SE_{corrected}$ for the randomized 'test' set is 0.31 (MI).

3.7. Probable biophysical basis of the correlation between folding rate and co-evolution order

Theoretical models relate the folding rate with the number of native contacts, under the simple assumption of a native-like transition state [27,28]. However, protein folding includes a series of structural reconstitution processes, for which even the non-contacting residue positions might be constrained to each other as well. In our recent work [29], we have shown that critical non-native structural dependencies drive co-evolutionary phenomenon in 30S ribosomal complex. Here in the case of small globular proteins, results clearly indicate that non-contacting CEPs also make a significant contribution to folding. However, a major fraction of non-contacting CEPs correspond to indirect native contacts, which might be informative to several possibilities, including critical non-native contacts, transition state structure, extended nucleus (described by Fresht [27]) etc., and they might affect the folding rate by contributing to the configurational entropy loss exactly like those in physical contact. In summary, these results might be an important step in understanding the role of folding constraints in protein evolution.

Funding

This work is supported by the Center of Excellence in Systems Biology and Biomedical Engineering (TEQIP Phase II), University of Calcutta, India.

Acknowledgements

The authors acknowledge Tanaya Ray, Harish-Chandra Research Institute, for many conceptual discussions and comments on the data analysis. Authors also acknowledge the anonymous reviewers for their many constructive suggestions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.febslet.2015.06.032.

References

[1] Goldberg, M.E., Semisotnov, G.V., Friguet, B., Kuwajima, K., Ptitsyn, O.B. and Sugai, S. (1990) An early immunoreactive folding intermediate of the tryptophan synthase beta 2 subunit is a 'molten globule'. FEBS Lett. 263, 51–56.

[2] Gromiha, M.M., Thangakani, A.M. and Selvaraj, S. (2006) FOLD-RATE: prediction of protein folding rates from amino acid sequence. Nucleic Acids Res. 34, W70–W74 (web server issue).

[3] Ivankov, D.N. and Finkelstein, A.V. (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc. Natl. Acad. Sci. U.S.A. 101, 8942–8944.

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

[5] Grantcharova, V., Alm, E.J., Baker, D. and Horwich, A.L. (2001) Mechanisms of protein folding. Curr. Opin. Struct. Biol. 11, 70–82.

[6] Plaxco, K.W., Larson, S., Ruczinski, I., Riddle, D.S., Thayer, E.C., Buchwitz, B., Davidson, A.R. and Baker, D. (2000) Evolutionary conservation in protein folding kinetics. J. Mol. Biol. 298, 303–312.

[7] Martinez, J.C., Pisabarro, M.T. and Serrano, L. (1998) Obligatory steps in protein folding and the conformational diversity of the transition state. Nat. Struct. Biol. 5, 721–729.

[8] Fulton, K.F., Main, E.R., Daggett, V. and Jackson, S.E. (1999) Mapping the interactions present in the transition state for unfolding/folding of FKBP12. J. Mol. Biol. 291, 445–461.

[9] Naganathan, A.N. and Muñoz, V. (2010) Insights into protein folding mechanisms from large scale analysis of mutational effects. Proc. Natl. Acad. Sci. U.S.A. 107, 8611–8616.

[10] Lindorff-Larsen, K., Paci, E., Serrano, L., Dobson, C.M. and Vendruscolo, M. (2003) Calculation of mutational free energy changes in transition states for protein folding. Biophys. J. 85, 1207–1214.

[11] Huang, L.T. and Gromiha, M.M. (2010) First insight into the prediction of protein folding rate change upon point mutation. Bioinformatics 26, 2121–2127.

[12] Weatheritt, R.J. and Babu, M.M. (2013) The hidden codes that shape protein evolution. Science 342, 1325–1326.

[13] Morcos, F., Schafer, N.P., Cheng, R.R., Onuchic, J.N. and Wolynes, P.G. (2014) Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. Proc. Natl. Acad. Sci. U.S.A. 111, 12408–12413.

[14] Makarov, D.E. and Plaxco, K.W. (2003) The topomer search model: a simple, quantitative theory of two-state protein folding kinetics. Protein Sci. 12, 17–26.

[15] Lovell, S.C. and Robertson, D.L. (2010) An integrated view of molecular coevolution in protein–protein interactions. Mol. Biol. Evol. 27, 2567–2575.

[16] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

[17] de Juan, D., Pazos, F. and Valencia, A. (2013) Emerging methods in protein co-evolution. Nat. Rev. Genet. 14, 249–261.

[18] Dunn, S.D., Wahl, L.M. and Gloor, G.B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24, 333–340.

[19] Gouveia-Oliveira, R. and Pedersen, A.G. (2007) Finding coevolving amino acid residues using row & column weighting of mutual information & multi-dimensional amino acid representation. Algorithms Mol. Biol. 2, 12.

[20] Martin, L.C., Gloor, G.B., Dunn, S.D. and Wahl, L.M. (2005) Using information theory to search for co-evolving residues in proteins. Bioinformatics 21, 4116–4124.

[21] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T. and Weigt, M. (2014) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. U.S.A. 108, E1293–E1301.

[22] Kamisetty, H., Ovchinnikov, S. and Baker, D. (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc. Natl. Acad. Sci. U.S.A. 110, 15674–15679.

[23] Wollenberg, K.R. and Atchley, W.R. (2000) Separation of phylogenetic & functional associations in biological sequences by using the parametric bootstrap. Proc. Natl. Acad. Sci. U.S.A. 97, 3288–3291.

[24] Ouyang, Z. and Liang, J. (2008) Predicting protein folding rates from geometric contact and amino acid sequence. Protein Sci. 17, 1256–1263.

[25] Melnik, T.N., Povarnitsyna, T.V., Glukhov, A.S. and Melnik, B.S. (2012) Multi-state proteins: approach allowing experimental determination of the formation order of structure elements in the green fluorescent protein. PLoS ONE 7, e48604.

[26] Bunner, A.E., Beck, A.H. and Williamson, J.R. (2010) Kinetic cooperativity in *Escherichia coli* 30S ribosomal subunit reconstitution reveals additional complexity in the assembly landscape. Proc. Natl. Acad. Sci. U.S.A. 107, 5417–5422.

[27] Fersht, A.R. (2000) Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. Proc. Natl. Acad. Sci. U.S.A. 97, 1525–1529.

[28] Makarov, D.E., Keller, C.A., Plaxco, K.W. and Metiu, H. (2002) How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. Proc. Natl. Acad. Sci. U.S.A. 99, 3535–3539.

[29] Mallik, S., Akashi, H. and Kundu, S. (2015) Assembly constraints drive co-evolution among ribosomal constituents. Nucleic Acids Res. 43, 5352–5363.