FEBS Letters 589 (2015) 2179–2185

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/

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:

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

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:

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

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]:

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

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:

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

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*⟨dc⟩, *rCEO*⟨nc⟩, *rCEO*⟨sr⟩ and *rCEO*⟨ic⟩, 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*⟨dc/MI⟩.

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