Vetriselvan Divya et al /J. Pharm. Sci. & Res. Vol. 8(2), 2016, 118-120

ISSN:0975-1459
Journal of Pharmaceutical Sciences and Research
www.jpsr.pharmainfo.in

Computational Models for Predicting Folding Rates of Proteins

Vetriselvan Divya and Thirunavukkarasu Sivaraman*
Structural Biology Lab, Department of Bioinformatics, School of Chemical and Biotechnology, SASTRA University, Thanjavur – 613401, Tamil Nadu, India.

Abstract

Understanding the chief forces governing the folding rates of proteins is essential to facilitate *de novo* protein designing and as well to overcome protein misfolding problems. The folding rates and folding pathways of proteins can be characterized at high resolution using variety of kinetic experiments and as well by computational methods. In these contexts, this article exemplify exclusively various 'topological descriptors' reported for predicting folding rates of proteins in the literature from 1998 to till date. Moreover, principles, unique features and limitations of each model of the computational approaches have been systematically evaluated.

Keywords: Contact order, Folding rates, Statistical models and Topology.

1. INTRODUCTION

Proteins are the most important structural and functional biomolecules, which play vital roles in carrying out many biological activities such as immune system reactions, signal transduction, gene expression, storage, translocation and many more. In general, proteins are synthesized in ribosomes of eukaryotic organisms as linear polypeptide chains, which then acquire unique biologically active three-dimensional (3D) structures [1-3]. Though much has been learnt on 'structure-function' relationships of proteins, the mechanisms by which these functional proteins are folded from their unstructured conformations to biologically active conformations are still puzzling [4-7]. Aberrations in the protein folding processes often impair the functions of proteins leading to devastating consequences [8-10]. An array of debilitating diseases may be caused either due to accumulation of misfolded proteins forming toxic aggregates (Alzheimer's disease, Creutzfeldt-Jacob disease, Familial amyloidosis) or absence of correctly folded proteins from the site of action (cystic fibrosis, phenylketonuria, Medium-chain acyl-CoA dehydrogenase deficiency). In addition, protein aggregation will also adversely affect the *in vitro* laboratory experiments and industrial protein production [11, 12]. Therefore, unravelling the protein folding puzzle will help on understanding the mechanisms of the protein misfolding/aggregation and also on stimulating drug designing processes [13, 14].

Notwithstanding the advantages of sophisticated biophysical techniques that are being used to study refolding of proteins at high resolution and as well in sub-millisecond time scales, the experiments are highly challenging from technical standpoints and also require sound knowledge on the structural architectures and conformational stabilities of proteins. In these contexts, computational tools will be an excellent alternative to predict folding rates of proteins solely on the basis of 3D structures of the proteins. This article reviews various pertinent statistical models documented in the literature from 1998 to till date for predicting the folding rates of proteins.

2. COMPUTATIONAL METHODS FOR PREDICTING FOLDING RATES OF PROTEINS

Various topological descriptors such as contact order, long range order, total contact distance, absolute contact order, chain topology parameter, fraction local contacts, long range contact order and NN contacts order demonstrated to date in the literature for calculating folding rates of proteins have been systematically discussed in the following section.

2.1 Contact order (CO)

Contact order is the first topological descriptor proposed to correlate the relationships between folding rates of two-state folding proteins and the parameter. The contact order is defined as shown in the following equation [15].

$$CO = \frac{1}{n_c n_r} \sum_{\substack{k=1 \\ |i-j| > l_{cut}}}^{n_c} |i-j|$$

wherein, n_r is protein length represented as number of amino acids. However, amino acids constituting disordered regions of a protein are generally excluded from the calculation; n_c is total number of contacting residue-residue pairs present in the distance cut-off (R_{cut}) of ≤ 6 Å for each residue of a protein; |i – j| is the sequence separation between contacting residues 'i' and 'j'; l_{cut} is generally set to 2 for enumerating contacting residue - residue pairs.

It should be mentioned that CO values are calculated for proteins by taking into considerations of non-hydrogen atoms of amino acids in the proteins. As defined by the above equation, the CO of a protein represents average sequence separation per contact per residue of the protein in the folded state. It has been shown that the correlation coefficient of 0.81 upon correlating the CO values and folding rates of 12 simple two-state folding proteins and the 12 proteins are λ-

Repressor (1LMB3), Equine cyt c (1HRC), Bovine ACBP (2ABD), Ubiquitin (1UBQ), CI-2 (1CIS), ADA2h (1PCA), Protein L (2PTL), HPr (1HDN), Muscle AcP (1APS), CspB (1CSP), TnFN3 (1TEN), FynSH3 (1SHFA). The findings suggest that both shorter-range contacts and long-range contacts established by residues of proteins are essential factors governing the folding rates of the proteins. The CO can be calculated for a protein through the webserver http://www.bakerlab.org/contact_order/

2.2 Long-range order (LRO)

When a protein folds from the unfolded states to the native folded states, residues that are distantly apart in the sequence but involving in close network of structural contacts may play vital roles in kinetic refolding of the protein. This is a prime basis for proposing long- range order to predict folding rates of two-sate folders. The LRO for a given protein can be calculated as illustrates in the following mathematical expression [16].

$$LRO = \frac{\sum n_{ij}}{N}$$

nij = 1 if |i-j| > 12 otherwise nij = 0.

In the above equation, 'N' stands for number of amino acids in a protein; |i – j| is the sequence separation between contacting residues 'i' and 'j'; The l_{cut} and R_{cut} for the LRO calculations were set to 12 residues and 8Å, respectively. Moreover, the LRO calculation considers only Cα atoms of proteins. As defined, LRO is a ratio of total numbers of contacts to length of amino acid sequence. In other words, LRO defines number of long-range contacts per residue.

In order to correlate the LRO and folding rates of two-state folding proteins, the proteins were first grouped into three categories: all-α proteins, all-β proteins and αβ mixed proteins. The correlation coefficients were -0.72 for all-α proteins, -0.92 for all-β proteins and -0.86 for mixed class proteins. When considered proteins belonging to all of the three classes together, the correlation coefficient was found to be -0.78. In these backgrounds, the authors of LRO suggested that statistical models for predicting folding rates for each class of proteins must be framed on the basis of structural classifications of proteins. Moreover, the striking correlations between the LRO and folding rates of proteins implying folding nucleus of proteins may present at an interval of approximately 25 residues.

2.3 Total contact distance (TCD)

As discussed above, both the CO and LRO depict a significant correlation with logarithm of folding rates of a set of proteins suggesting that both local and non-local contacts are playing essential roles in governing the folding rates of the proteins. The TCD incorporates both the CO and LRO together and it is defined as shown below [17].

$$TCD = \frac{1}{n_r^2} \sum_{\substack{k=1 \\ |i-j|>l_{cut}}}^{no} \Delta L_{ij}$$

In the equation, n_r is a number of amino acid residues of a protein and Lij is the sequence separation between contacting residues 'i' and 'j' under defined l_{cut} and R_{cut}. Hence, TCD = CO x LRO, provided both the CO and LRO calculated with same values of l_{cut} and R_{cut}. While the CO and LRO defines sequence separation per contact per residue and long range contacts per residue, respectively, the TCD accounts summation over all the contacts per residue. Interestingly, TCD depicted a correlation coefficient of -0.88 with the folding rates of 28 proteins indicating TCD is a better topological parameter in predicting folding rates of two-state folding proteins comparing that of CO and LRO.

2.4 Absolute contact order (Abs_CO)

All the three topological parameters (CO, LRO & TCD) discussed above have been shown as reliable to predict folding rates of two-state folders only. In other words, the reliability of the parameters for predicting folding rates of multi-state folders was left largely unaddressed. In these connections, the Abs_CO has been proposed to predict folding rates of proteins belonging to both of the two categories and the parameter is defined as represented below herein [18].

$$Abs\_CO = \frac{1}{N} \sum_{k=1}^{N} \Delta L_{ij} = CO \times L$$

In the above equation, the N and L represent the total number of contacts and total number of amino acids in a protein, respectively. The other terms in the equations are as represented in the previous sub-headings. The Abs_CO is the one in which the sum of sequence separation between all pairs of contacting residues are divided by total number of contacts in the protein. Hence, the Abs_CO defines sequence separation per contact of a protein. While CO is independent on chain length of a protein, the Abs_CO accounts the protein size. The authors of the original paper also introduced a parameter 'size-modified contact order' (SMCO). The SMCO is defined as SMCO = CO X L^P and the SMCO becomes CO, when P = 0, and in contrast, the SMCO becomes Abs_CO, when P = 1. It has also been demonstrated that the Abs_CO scales with the chain length as P = 0.70 ± 0.07 and depicted appreciable correlation with folding rates of peptides and as well both types of the two-state and multi-state proteins.

2.5 Chain topology parameter (CTP)

The CTP is defined as the sum of the square of the sequence separation between the pair of contacting residues i and j divided by number of inter residue contacts and total number of residues in the protein. The

mathematical expression of the CTP is as follows [19].

$$CTP - \frac{1}{LN}\sum_{k=1}^{N}\Delta S_{ij}^{2}$$

wherein, ΔS_{ij} is the separation in sequence between the contacting residues i and j and other parameters are as represented in the previous sub-headings. The authors of the original paper have shown that the rate of folding of proteins and CTP were in linear relationships within the range of 10^{-1} s^{-1} $\leq k_f \leq 10^{8}$ s^{-1}. Moreover, the CTP is unique for predicting folding rates of isolated helices and β-hairpin small peptides. The comprehensive analyses on the correlation between the CTP and folding rates of over 20 proteins also suggested that short-sequence separations are presumably favourable factors for establishment of stable proteins.

3. Concluding Remarks

In addition to various topological parameters (CO, LRO, TCD, Abs_CO & CTP) as described above for predicting folding rates of peptides and proteins, a few more descriptors such as fraction of local contact, long range contact order and NN contacts have also been documented in the literature. The fraction local contacts are very useful to understand the relative importance of local contacts in protein folding pathways [20, 21]. The long range order parameter clearly differentiates effect of short and long range contacts on folding rates of polypeptides [22]. The NN contacts parameter defines different types of residue-wise structural parameters (short range, medium range, long range and total weighted order parameters) for a protein [http://sblab.sastra.edu/NNCOCalculator.html]. Thus, the NN contacts are very useful parameter to computationally explore folding rates of proteins at residue level. Moreover, in addition to predicting protein folding rates, the structural parameters are also very useful in drug designing. Considering those many facets of applications, we strongly feel that there is a great scope to develop unprecedented computational strategies for addressing structures – folding relationships of proteins in near future [23, 24].

Acknowledgements

We would like to thank all the researchers who have significantly contributed to the topological descriptors for predicting folding rates of proteins belonging to various classes.

References

1. Creighton TE.Biochem J 1990, 270: 1-16.
2. Dobson CM,Karplus M. Current Opinion in Structural Biology 1999, 9: 92-101.
3. Rose GD, Fleming PJ, Banavar JR and Maritan A. 2006, 103: 16623-16633.
4. Chan HS, Dill KA. Physics today 1993, 46: 24-32.
5. Dill KA, MacCallum JL. Science 2012, 338: 1042-1046.
6. Radford SE. Trends Biochem Sci 2000, 25: 611-8.
7. Bystroff C, Simons KT, Han KF and Baker D., Current Opinion in Biotechnology 1996,7: 417-421.
8. Chiti F and Dobson CM. Biochemistry 2006, 75: 333-336.
9. Barral JM, Broadley SA, Schaffar G, Hartl FU. Seminar in Cell and Developmental Biology 2004,15: 17-29.
10. Liu J and Song J. PLoS ONE 2009, 4: doi:10.1371/journal.pone.0007805
11. Engelsman JD, Garidel P, Smulders R, Koll H, Smith B, Bassarab S, Seidl A, Hainzl O, Jiskoot W. PHARMACEUTICAL RESEARCH 2011, 28: 920-993.
12. Finkelstein AV, Galzitskaya OV. Physics of Life Reviews 2004, 1: 23-56.
13. Luheshi LM, Crowther DC, Dobson CM.Current Opinion in Chemical Biology 2008, 12: 25- 31.
14. Uversky VN. Cell Mol Life Sci 2003, 60: 1852-71.
15. Plaxco KW, Simons KT, Baker D. J MolBiol 1998, 277: 985-994.
16. Gromiha MM and Selvaraj S. J MolBiol 2001, 310: 27-32.
17. Zhou H, Zhou Y. Biophysical Journal 2002, 82: 458-463.
18. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV. Protein Science 2003, 12: 2057-62
19. Nolting B, Schalike W, Hampel P, Grundig F, Gantert S , Bandlow W, Qi PX. Journal of Theoretical Biology 2003, 223: 299-307.
20. Mirny L, Shakhnovich E. Annu. Rev. Biophys. Biomol. Struct.2001, 30: 361-96.
21. Kuznetsov IB, Rackovsky S. PROTEINS: Structures, Function, and Bioinformatics 2004, 54: 333-341.
22. Ma BG, Chen LL, Zhang HY. J. Mol. Biol.2007, 370: 439-448.
23. Hao PY, Tsai LB. Opportunities and Challenges for Next-Generation Applied Intelligence 2009, 214: 7-12.
24. Prabhavadhini A, Richa T, Sivaraman T. Journal of Pharmaceutical Science and Research 2015, 7: 159-162.