35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Role for Gene Sequence, Codon Bias and mRNA Folding Energy in Modulating Structural Symmetry of Proteins

Xiaojuan Shen, Shixiong Chen, and Guanglin Li, *Senior Member, IEEE*

***Abstract*—Structural symmetry in proteins is commonly observed in the majority of fundamental protein folds. Meanwhile, nascent polypeptide chains of proteins have the potential to start the co-translational folding process and this process can have drastic effects on protein structure. Thus we are interested in understanding mechanisms that gene adopts in specifying structural symmetry in proteins. In the present paper, we reveal that for two representative symmetric proteins from $(\alpha\beta)_8$-barrel fold and beta-trefoil fold, intragenic symmetry is detected in the corresponding gene sequences. Codon bias and mRNA folding energy might be involved in mediating translation speed for the formation of structural symmetry: at least one major decrease in both codon bias and mRNA folding energy can be observed in the connecting region of the symmetric substructures along the codon sequence. Results suggest that gene duplication and fusion is responsible for structural symmetry in these proteins, and the usage of rare codons or higher order of secondary structure near the boundaries of symmetric substructures might be selected in order to slow down translation speed for effectively co-translational folding process of symmetric proteins.**

I. Introduction

Internal symmetry of tertiary structures is often a feature of natural proteins[1]. A number of investigators have been interested in symmetric protein structures[2], their role in protein function and evolution[3, 4], and their utility in protein engineering and design[5]. It is reported that among ten fundamental protein superfolds proposed by Thornton, six have internal structural symmetry[6]. Gene duplication and subsequent fusion of the duplicated genes is hypothesized to be the origin of structural symmetry in proteins[7]. The $(\alpha\beta)_8$-barrel is a highly symmetric protein fold consisting of eight repeats of a β-strand/α-helix structure. The crystal structure of two $(\alpha\beta)_8$-barrel proteins HisA and HisF that are members of the histidine biosynthetic pathway show 2-fold internal symmetry in structures[8]. Experimental data also certify that the two proteins evolved from $(\alpha\beta)_4$ half-barrel [9]. The beta-trefoil is another common protein fold that exhibits 3-fold structural symmetry[4]. Each of the repeating domains contains 40-50 amino acids in length and is composed of four β-strands of two anti-parallel β-hairpins. The beta-trefoil fold is also widely held to have evolved from gene duplication and fusion processes. One single gene duplication and fusion event yield structural 2-fold symmetry, and higher order of symmetry requires subsequent duplication and fusion.

On the other hand, substantial evidence support that nascent proteins largely acquires spatial structures while still attached to the ribosome[10, 11], and variations of translation process may have drastic effects on the folding efficiency of newly synthesized proteins[12]. The translation speed along mRNAs is nonuniform and changes in translation rates may result from the presence of rare or slowly translated codons or local stable mRNA structure in the translated mRNA[10, 13, 14]. For example, replacement of rare codons by frequently used ones in the genes from E.coli or S.cerevisiae leads to faster translation but with reduced activity of the encoded proteins[15]. A silent mutation in the human gene MDR1 results in the encoded P-glycoprotein fold differently, suggesting that the conformation of the product is altered by the speed of protein synthesis[16]. It is now recognized that the degeneracy of the genetic code allows additional layer for mRNA to carry structural information relating with the encoded proteins [17, 18].

Given that internal symmetry is a common feature among protein structures, and protein can fold co-translationally on the ribosome, there must be some conserved mechanisms that gene adopts to modulate structural symmetry in proteins. In this paper, we investigate this issue for two representative proteins from $(\alpha\beta)_8$-barrel fold and beta-trefoil fold. As described above, proteins from the two protein folds contain 2-fold and 3-fold symmetry in structure, respectively. We first study the protein structure based intragenic symmetry using modified recurrence plot method [19-23]. Our results reveal that the same degree of symmetry in gene sequences of the symmetric proteins can be detected which provide clear evidence for the gene duplication and fusion hypothesis of the two proteins. Meanwhile, local codon bias and mRNA folding energy distribution along the codon sequences is also analyzed to investigate their correlation with protein symmetry. Results reveal that at least one major decrease of codon bias and mRNA folding energy in the linker peptide connecting symmetric substructures of the two representative proteins can be observed, which suggest that slow translated regions near the boundaries of symmetric substructures might be under selection pressure and provide as regulation to cooperate the co-translational protein folding.

*This work was supported in part by the National Science Foundation of China under Grant (#60971076), the Shenzhen Public Platform for Patient-specific Orthopedic Technology and Manufacturing Service, and the Guangdong Innovation Research Team Fund for Low-cost Healthcare Technologies.

X. Shen, S. Chen and G. Li are with the Neural Engineering Center, the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; and Graduate University of Chinese Academy of Sciences, 19A Yuquanlu, Beijing, 100049, China (phone:+86-755-86392219; fax:+86-755-86392299; gl.li@siat.ac.cn).

II. MATERIALS AND METHODS

Modified recurrence plot method

The modified recurrence plot method has been shown to be effective in detecting internal symmetry in both protein sequences and structures [19-23]. Here we will use it to detect protein structure based symmetry in gene sequences and it works as follows: An arbitrary gene sequence is denoted as $S = x_1 x_2 x_3 ... x_N$, where x_i denotes one of the four nucleic acids and *N* is the length of the gene sequence. One constructs a set of all (*N-d+1*) possible segments of *d (d<N)* consecutive symbols:

$X_1 = x_1 x_2 ... x_d$
$X_2 = x_2 x_3 ... x_{d+1}$
.....
$X_i = x_i x_{i+1} ... x_{i+d-1}$
.....
$X_{N-d+1} = x_{N-d+1} x_{N-d+2} ... x_N$

where *i* denotes the location of the first nucleotide of X_i in the sequence. For any given segment X_i, we find how many other segments are similar to it. Hamming distance is used for calculate the similarity between two segments:

$$h(x_i, x_j) = \begin{cases} 1 & x_i = x_j \\ 0 & x_i \neq x_j \end{cases} \qquad (1)$$

and a segment X_j is defined similar to X_i if the percentage of identical nucleotides is larger than a given cutoff of similarity degree *S*. Furthermore, p-value is calculated to assess the statistical significance of the alignment of the two peptide segments. It is defined as the probability of obtaining an alignment with the same similarity by self-alignment of scrambled sequences. The similarity of the two aligned segments is considered statistically significant when p-value is lower than 0.01. The modified recurrence plot is built as follows: the horizontal axis of the modified recurrence plot is the residue index in the sequence and the vertical axis is the segment length *d*. For each segment $X_i(d)$, we can give a value $S_{d,i}$, which is the number of the segments similar to $X_i(d)$. Thus, for all the $(N-d+1)$ segments $(X_1(d), X_2(d), \ldots, X_{N-d+1}(d))$, we obtain a set of the numbers $(S_{d,1}, S_{d,2}, \ldots, S_{d,N-d+1})$. The analysis of similar segments can be done for different segment lengths d.

Local Codon usage bias and mRNA folding energy

The classical *CAI* (Codon Adaption Index) value is used as measurement of codon usage bias[24]. The definition of *CAI* is the geometric mean of the relative synonymous codon usage (*RSCU*) values corresponding to each of the codons used in that sequence, divided by the maximum possible *CAI* for a gene sequence of the same amino acid composition.

$$CAI = CAI_{obs} / CAI_{\max}$$
$$CAI_{obs} = (\prod_{k=1}^{L} RSCU_k)^{1/L} \qquad (2)$$
$$CAI_{\max} = (\prod_{k=1}^{L} RSCU_{k\max})^{1/L}$$

$RSCU_k$ and $RSCU_{max}$ are the *RSCU* value for the kth codon and the maximum *RSCU* value among the synonymous codon group of kth codon, respectively. An *RSCU* value for a codon is the observed frequency of the codon divided by the frequency expected[24]:

$$RSCU_{ij} = \frac{x_{ij}}{\frac{1}{n}\sum_{j=1}^{n_i} x_{ij}} \qquad (3)$$

where x_{ij} is the number of occurrences of the jth codon for ith amino acid, and n_i is the number of synonymous codons for ith amino acid. For computational effectiveness, (2) is computed as:

$$CAI = (\prod_{k=1}^{L} x_{ij} / x_{i\max})^{1/L} = (\prod_{k=1}^{L} w_k)^{1/L} = \exp\frac{1}{L}\sum_{k=1}^{L} \ln w_k \qquad (4)$$

where w_k is the relative adaptiveness of a codon and is denoted as $x_{ij} / x_{i\max}$. Because there is no intrinsic effect of gene length on *CAI*, we use this value to investigate the local codon usage bias within codon sequence. We set a sliding window of 20 consecutive codons and the local *CAI* for all the sliding window of length 20 is calculated and the profile of the distribution is given.

We use Matlab rnafold function to calculate the mRNA folding free energy that predicts the folding energy of a secondary structure associated with minimum free energy of the RNA sequence. We set a sliding window of 40 nucleotides long and shift by one nucleotide along the gene sequence and the profile of the distribution is given. The length of 40 nucleotides is approximately the footprint length of ribosome on mRNA.

III. RESULTS AND DISCUSSION

Protein 1QO2 from $(\alpha\beta)_8$-barrel fold and protein 1KNM from beta-trefoil fold are selected as representatives. Gene sequences of the two proteins are downloaded from ExPASy and the genome data are from genbank FTP. The codon usage frequency for a species is calculated based on the whole genome data. Protein 1QO2 is the gene products of HisA from hyperthermophile Thermotoga maritime. It is an enzyme that converts N9- [(59-phosphoribosyl)-formimino] 5- aminoimidazol-4- carboxamid ribonucleotide into the 59-phosphoribulosyl isomer[8]. The crystal structure of this protein reveals 2-fold symmetry with each substructure contains four repeats of β-strand/α-helix model. The cartoon structure of 1QO2 is given in Figure 1 (upper). The modified recurrence plot of the nucleotide sequences of this protein reveals 2-fold symmetry in the nucleotide sequence (Figure 1 lower), with the first segment contains nucleotides 1-342 and the second segment contains nucleotides 343-726. The similarity degree for this protein is set as 35%. The profile of codon usage bias reveals a major decrease in the middle of the

codon sequence (Figure 2 upper). A decrease of *CAI* denotes that there are an increasing number of rare codons in the corresponding segment. Results suggest that rare or less frequently used codons are more likely to be used near the boundary of the two symmetric substructures. In the profile of folding energy distribution, decreases in the middle region of the nucleotide sequence can also be observed. It indicates that codons that facilitate the formation of higher order of secondary structure in local mRNA sequence are more favorable near the middle regions. Protein 1KNM is Hydrolase from streptomyces lividans and its crystal structure reveals 3-fold symmetry. The cartoon structure of 1KNM is given in Figure 3 (upper). The modified recurrence plot of the nucleotide sequence of 1KNM reveals 3-fold symmetry (Figure 3 lower), with each segment length containing 120 nucleotides. The similarity degree for this protein is set as 50%. Two major decreases of codon usage bias within the codon sequence can be observed (Figure 4 upper), suggesting that rare codons are selected for during these regions. In the distribution of local folding free energy (Figure 4 lower), two regions with consecutive decreases in the folding free energy can also be detected, indicating that higher order of structure in mRNA is likely to occur for these regions. Results reveal that for proteins from $(\alpha\beta)_8$-barrel fold and beta-trefoil fold, gene duplication and fusion evidence can be found in the corresponding gene sequences. At the same time, codon usage bias and mRNA folding free energy might be mechanisms that involved in modulating the co-translational folding process to coordinate the effective folding of symmetric substructures.

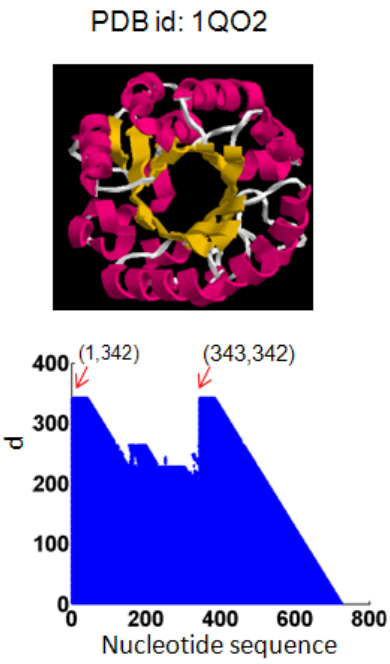

Figure 1 Tertiary structure (upper) of protein 1QO2 and modified recurrence plot (lower) for its nucleotide sequence. The recurrence plot exhibit a clear 2-fold symmetry in the nucleotide sequence, with the first subsegment from 1 to 342 nucleotide and the second subsegment from 342 to the end.

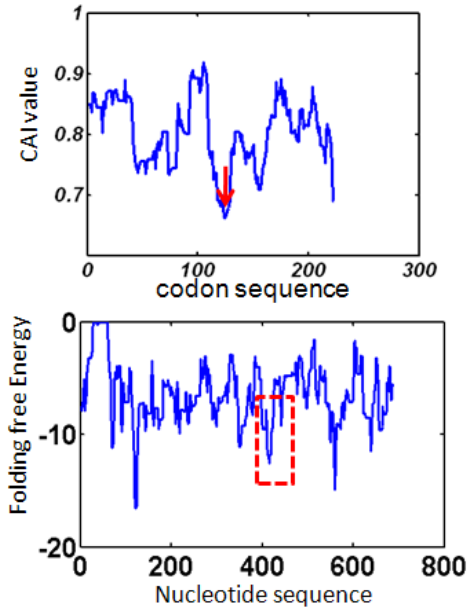

Figure 2 Profile of local codon usage bias distribution (upper) and folding free energy distribution (lower) for protein 1QO2. In the distribution of local *CAI* value, a decrease of CAI can be observed in the middle of the codon sequence (pointed out by red arrow). In the distribution of local folding free energy, a decrease in the middle region of the nucleotide sequence can also be observed and the region is shown with dashed square lines.

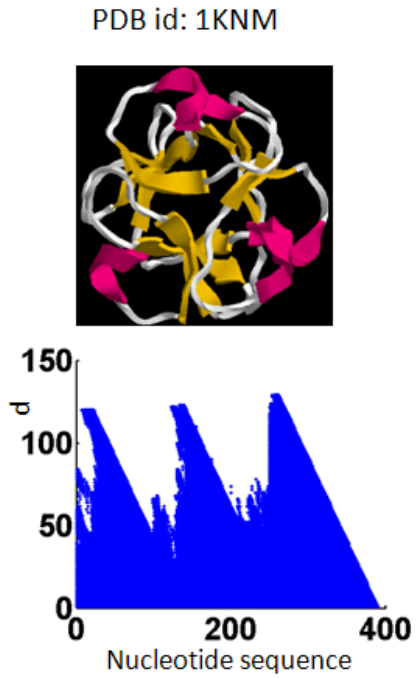

Figure 3 Tertiary structure (upper) of protein 1KNM and modified recurrence plot (lower) for its nucleotide sequence. The modified recurrence plot exhibits a clear three-fold symmetry in the nucleotide sequence, with each segment containing 120 nucleotides.

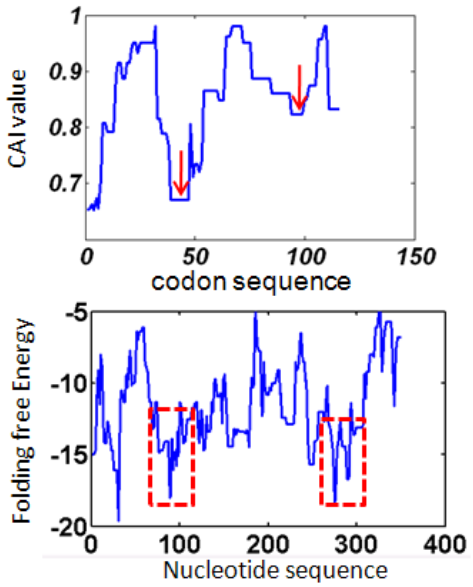

Figure 4 Profile of local codon usage bias distribution (upper) and folding free energy distribution (lower) for protein 1KNM. In the distribution of local *CAI* value, two major decreases of *CAI* can be observed within the codon sequence (pointed out by red arrows). In the distribution of local folding free energy, regions with consecutive decreases of folding energy within the nucleotide sequence can also be observed and are shown with dashed square lines.

IV. Conclusion

In this paper, we studied the relationship between the structural symmetry of proteins and the nucleotide sequence, codon bias, and mRNA folding energy in two representative proteins from $(\alpha\beta)_8$ -barrel fold and beta-trefoil fold. Our results reveal that the same degree of intragenic symmetry is detected in both the two selected proteins which provide clear evidence for the gene duplication and fusion hypothesis. Results also show that major decreases or consecutive decreases in local codon usage bias or mRNA folding energy near the boundaries of the symmetric substructures can be detected, suggesting that codon usage bias and higher order of mRNA structure might be factors that modulate translation process for the formation of structure symmetry. Since only two representative proteins were analyzed in the preliminary study, more proteins from different folds and from different species would be carried out in the future to further testify the universality before we draw a general conclusion on the relationships and mechanisms proposed in this paper.

Acknowledgment

The authors would like to thank Professor Yi Xiao from Huazhong University of Science and Technology for his valuable suggestion and discussion for this work.

References

[1] C. A. Orengo, D. T. Jones, and J. M. Thornton, "Protein superfamilies and domain superfolds," *Nature,* vol. 372, pp. 631-634, 1994.

[2] C. Kim, J. Basner, and B. Lee, "Detecting internally symmetric protein structures," *BMC bioinformatics,* vol. 11, p. 303, 2010.

[3] A. Broom, A. C. Doxey, Y. D. Lobsanov, L. G. Berthin, D. R. Rose, P. L. Howell, B. J. McConkey, and E. M. Meiering, "Modular Evolution and the Origins of Symmetry: Reconstruction of a Three-Fold Symmetric Globular Protein," *Structure,* vol. 20, pp. 161-171, 2012.

[4] M. Blaber, J. Lee, and L. Longo, "Emergence of symmetric protein architecture from a simple peptide motif: evolutionary models," *Cellular and Molecular Life Sciences,* pp. 1-8, 2012.

[5] M. Blaber and J. Lee, "Designing proteins from simple motifs: opportunities in Top-Down Symmetric Deconstruction," *Current Opinion in Structural Biology,* 2012.

[6] G. M. Salem, E. G. Hutchinson, C. A. Orengo, and J. M. Thornton, "Correlation of observed fold frequency with the occurrence of local structural motifs," *J Mol Biol,* vol. 287, pp. 969-981, 1999.

[7] A. McLachlan, "Repeating sequences and gene duplication in proteins," *J Mol Biol,* vol. 64, pp. 417-437, 1972.

[8] D. Lang, R. Thoma, M. Henn-Sax, R. Sterner, and M. Wilmanns, "Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion," *Science,* vol. 289, pp. 1546-1550, 2000.

[9] B. Höcker, S. Beismann-Driemeyer, S. Hettwer, A. Lustig, and R. Sterner, "Dissection of a (βα) 8-barrel enzyme into two folded halves," *Nature structural & molecular biology,* vol. 8, pp. 32-36, 2001.

[10] G. Zhang and Z. Ignatova, "Folding at the birth of the nascent chain: coordinating translation with co-translational folding," *Current opinion in structural biology,* vol. 21, pp. 25-31, 2011.

[11] F. U. Hartl and M. Hayer-Hartl, "Converging concepts of protein folding in vitro and in vivo," *Nat Struct Mol Biol,* vol. 16, pp. 574-581, 2009.

[12] C. J. Tsai, Z. E. Sauna, C. Kimchi-Sarfaty, S. V. Ambudkar, M. M. Gottesman, and R. Nussinov, "Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima," *Journal of molecular biology,* vol. 383, pp. 281-291, 2008.

[13] T. Thanaraj and P. Argos, "Ribosome - mediated translational pause and protein domain organization," *Protein Science,* vol. 5, pp. 1594-1612, 2008.

[14] G. Kramer, D. Boehringer, N. Ban, and B. Bukau, "The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins," *Nat Struct Mol Biol,* vol. 16, pp. 589-97, Jun 2009.

[15] A. A. Komar, T. Lesnik, and C. Reiss, "Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation," *Febs Letters,* vol. 462, pp. 387-391, 1999.

[16] C. Kimchi-Sarfaty, J. M. Oh, I. W. Kim, Z. E. Sauna, A. M. Calcagno, S. V. Ambudkar, and M. M. Gottesman, "A" silent" polymorphism in the MDR1 gene changes substrate specificity," *Science,* vol. 315, pp. 525-528, 2007.

[17] A. A. Komar, "A pause for thought along the co-translational folding pathway," *Trends in biochemical sciences,* vol. 34, pp. 16-24, 2009.

[18] M. Jia and L. Luo, "The relation between mRNA folding and protein structure," *Biochem Biophys Res Commun,* vol. 343, pp. 177-82, Apr 28 2006.

[19] R. Xu and Y. Xiao, "A common sequence-associated physicochemical feature for proteins of beta-trefoil family," *Computational biology and chemistry,* vol. 29, pp. 79-82, 2005.

[20] X. Shen, "Conformation and sequence evidence for two-fold symmetry in left-handed beta-helix fold," *Journal of Theoretical Biology,* vol. 285, pp. 77-83, 2011.

[21] Y. Huang and Y. Xiao, "Detection of gene duplication signals of Ig folds from their amino acid sequences," *PROTEINS: Structure, Function, and Bioinformatics,* vol. 68, pp. 267-272, 2007.

[22] M. Li, Y. Huang, and Y. Xiao, "Effects of external interactions on protein sequence - structure relations of beta - trefoil fold," *Proteins: Structure, Function, and Bioinformatics,* vol. 72, pp. 1161-1170, 2008.

[23] J. P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *EPL (Europhysics Letters),* vol. 4, p. 973, 1987.

[24] P. M. Sharp and W. H. Li, "The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications," *Nucleic Acids Res,* vol. 15, pp. 1281-1295, 1987.