A novel method of analyzing proline synonymous codons in *E. coli* ☆

Ming-Lei Wang[a,b,*], Jiang-Ning Song[a,b], Wen-Bo Xu[c], Wei-Jiang Li[a,b]

[a]*The Key Laboratory of Industrial Biotechnology, Ministry of Education, Southern Yangtze University, Wuxi 214036, Jiangsu, China*
[b]*School of Biotechnology, Southern Yangtze University, Wuxi 214036, Jiangsu, China*
[c]*School of Information Technology, Southern Yangtze University, Wuxi 214036, Jiangsu, China*

Accepted 13 September 2004

Available online 25 September 2004

Edited by Robert B. Russell

Abstract Proline is a special imino acid in protein and the isomerization of the prolyl peptide bond has notable biological significance and influences the final structure of protein greatly, so the correlation between proline synonymous codon usage and local amino acid, the correlation between proline synonymous codon usage and the isomerization of the prolyl peptide bond were both investigated in the *Escherichia coli* genome by using a novel method based on information theory. The results show that in peptide chain, the residue at the first position C-terminal influences the usage of proline synonymous codon greatly and proline synonymous codons contain some factors influencing the isomerization of the prolyl peptide bond.

Keywords: Proline synonymous codon; *E. coli*; Information theory

1. Introduction

The planar peptide bond occurs predominantly in the *trans* isomerization due to partial double bond character of the planarity [1]. Planarity is maintained by a rotational barrier of about 20 kcal/mol (1 cal = 4.184 J). The difference in energy is approximately 2.5–2.6 kcal/mol between the *trans* and the *cis* isomerization. But prolyl residue is a unique one among the amino acids for it is an imino acid lacking the amide hydrogen atom rather than an amino acid. For an imide bond in Pro-containing peptides, the difference in energy is only 0.5 kcal/mol between the *trans* and the *cis* form, and the energy barrier is also less, 13 kcal/mol. So, only 0.03–0.05% Xaa–non-Pro peptide bonds occur in the *cis* form, while about 4–5% Xaa–Pro peptide bonds occur in the *cis* form [1–3].

Peptidyl-prolyl *cis*/*trans* isomerization is of considerable biological significance [4]. Peptidyl-prolyl *cis*/*trans* isomerization has been frequently found as a rate limiting step in the folding of proteins [5]. Prolyl residue plays an important role in the final structure of protein [6], e.g., the *cis*/*trans* isomerization of prolyl peptide bonds has been suggested to dominate the folding of the alpha subunit of tryptophan synthase from *Escherichia coli* (alphaTS) [6]; a conserved *cis* peptide bond is necessary for the activity of Bowman–Birk inhibitor protein [7]. In fact, isomer-specificity has been observed directly for protein–ligand and enzyme–substrate interactions and for protein folding, and indirectly through the catalytic effects of peptidyl-prolyl *cis*/*trans* isomerases [8–13]. Some theoretical prediction researches on prolyl isomerization have also been carried out and these predictions are based on the amino acid sequences [14,15].

It is commonly considered that the spatial structure of protein is determined by the primary structure, i.e., all the information needed by protein when folding into the natural structure are contained by the amino acid sequence. The sequence of nucleic acids contains more information than the amino acid sequence because of the degeneracy of the codons. Has the information ever been used by the life process and does it correlate with the protein folding? Because of the character of codons and the difference of the cellular environment, the processes of translation of synonymous codons are distinctly different. The difference will influence the dynamic process of folding of peptide chains. It is indicated by investigations emerging in the recent years that there is a correlation between the synonymous codons and the protein structure [16–19].

Many researches on synonymous codons were based on hypothesis-test method [17–19]. This method can surely find strong correlation signals, but its shortcoming is also obvious. First, hypothesis-test method can only provide the result of ‘‘yes’’ or ‘‘no’’. When correlation signals are lying in strong back-round noise, the result is often uncertain. Unfortunately, we always face this complexion in statistic research to bio-macromolecular sequences. Second, when hypothesis-test method says ‘‘no’’, it does not mean that the correlation does not exist but means the signal cannot be distinguished from the noise. Third, hypothesis-test method cannot tell us the degree of the correlation.

At present, most researches on prolyl peptide bond are based on the local or global amino acid sequences and the researches on the synonymous codons are rather less.

In this work, we tried to apply a novel method based on the information theory to reveal the hidden correlation between the prolyl synonymous codons usage and the local residues near the prolyl peptide bond, as well as the prolyl synonymous codons usage and peptidyl-prolyl *cis*/*trans* isomerization, in *E. coli*. So, we may acquire some useful information to predict prolyl isomerization.

☆ Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2004.09.034.

[*] Corresponding author. Fax: +86-510-5869645.
E-mail address: wml_yh@yahoo.com.cn (M.-L. Wang).

0014-5793/$22.00 © 2004 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2004.09.034

2. Materials and methods

2.1. Materials

EcoGene. A secondary database of *E. coli* from GenBank. 4304 gene sequences, about 4Mb, are included in the database [20].

EcoPDB. A high quality dataset of *E. coli* gene sequences and corresponding protein structures. This dataset includes 190 *E. coli* gene and corresponding Protein Data Bank structures determined by X-ray diffraction method with resolutions better than 2.5 Å [21]. Sequence identity between each pair of the sequences in EcoPDB was below 30%, except 1RPJA, 1VEWA, 1A99A, 1KAS, and 2LBP, which were not used in our research. All the entries in EcoPDB can be found in supplementary materials.

2.2. Methods

Correlation between the prolyl synonymous codons usage and the local residues near the prolyl peptide bonds. pro denoted proline and *r* denoted codons. The position of amino acid *a* was called the relative position *m* of the *pro* if the *m*th amino acid is *a* when counted from a certain *pro* to C-terminal. If $m<0$, amino acid *a* located the N-terminal side of *pro*. We defined:

$$I_m(pro|a) = \sum_{r \in pro} p_m^{pro}(r|a) \ln \frac{p_m^{pro}(r|a)}{p^{pro}(r)}$$

where $p_m^{pro}(r|a)$ stood for the relative usage frequency of codon *r* of *pro*, when amino acid *a* located the relative position *m* of *pro*; $p^{pro}(r)$ stood for the unconditional relative usage frequency of codon *r*. $\sum_{r \in pro}$ stood for the summations running over all codons of *pro* in the dataset.

According to the definition

$$\sum_{r \in pro} p_m^{pro}(r|a) = \sum_{r \in pro} p^{pro}(r) = 1$$

$I_m(pro|a) \geqslant 0$ also can be proved.

Only when, $p_m^{pro}(r|a) = p^{pro}(r)$, $I_m(pro|a)$ was equal to 0, i.e., only when the amino acid *a* locating the relative position *m* had no influence on the relative usage frequency of codons of *pro*, $I_m(pro|a)$ was equal to zero. Otherwise, the greater the influence was, the greater the difference between the conditional usage frequency and the unconditional usage frequency, i.e., the greater the value of $I_m(pro|a)$ was. So, $I_m(pro|a)$ could be used to weigh the degree of the influence of amino acid *a* on the relative usage frequency of codons of *pro*.

In the whole EcoGene, $I_m(pro|a)$ of all the 20 types of amino acids flanking the prolyl residue (10 residues of N-terminal and 10 residues of C-terminal, respectively, i.e., $m = -10 \sim +10$) were computed.

All the conformations of prolyl peptide bond in the EcoPDB were computed. According to the definition of PDB, those conformations with ω dihedral angles between −30° and +30° were *cis*.

Dinucleotide correlation between the third nucleotide in a prolyl synonymous codon and the first nucleotide in the subsequent codon. Similar to the method described above, dinucleotide correlation was also computed.

pro denoted proline and *r* denoted codons. *n* denoted the first nucleotide in the subsequent codon following the prolyl codon. We defined:

$$I(pro|n) = \sum_{r \in pro} p^{pro}(r|n) \ln \frac{p^{pro}(r|n)}{p^{pro}(r)}$$

where $p^{pro}(r|n)$ stood for the relative usage frequency of codon *r* of *pro*, when the first nucleotide in the subsequent codon was *n*, obviously $n \in (A, G, C, U)$; $p^{pro}(r)$ stood for the unconditional relative usage frequency of codon *r*. $\sum_{r \in pro}$ stood for the summations running over all codons of *pro* in the dataset.

Similar to the cause described above, the greater the influence of the first nucleotide in the subsequent codon was, the greater the value of $I(pro|n)$ was.

Correlation between the prolyl synonymous codons usage and the conformations of prolyl peptide bonds. The conformation of prolyl peptide bond was *cis* or *trans*. This uncertainty could be described by Shannon entropy. (For convenience, the natural logarithm was used here. The unit of entropy and information was nat, 1 bit = ln2 nat):

$$H(pro) = -\sum_{\Gamma} p(\Gamma|pro) \ln p(\Gamma|pro)$$

where $p(\Gamma|pro)$ stood for the probability of proline adopting the conformation Γ and the summation ran over both the *cis* and *trans* conformations.

The greater the value of $H(pro)$ was, the greater the uncertainty of the conformation of *pro* was; otherwise, the value of $H(pro)$ reached its minimum 0 if the conformation of *pro* had been certain, i.e., only one conformation had been adopted.

There were four types of the codons of proline, denoted by *r*, CCG, CCA, CCC, and CCU. Similar to the above, Shannon entropy of the conformation of the proline coded by *r* could be computed by

$$H(r) = -\sum_{\Gamma} p(\Gamma|r) \ln p(\Gamma|r)$$

where $p(\Gamma|r)$ stood for the frequency of proline, coded by *r*, adopting the conformation Γ.

Obviously, $H(pro) = H(r)$, if proline had only one codon, but it has four types of codons actually. So, the entropy of each codon was different from the entropy of proline. The entropies of four types of codons were averaged according to the usage frequency. The averaged conditional entropy would be obtained.

$$H_r(pro) = \sum_{r \in pro} p(r|pro) H(r)$$

where $p(r|pro)$ stood for the frequency of proline using codon *r* and the summation ran over all four types of codons. The equation indicated the uncertainty of the conformation of proline, when the used codon was known. It could be proved that the averaged conditional entropy was never greater than the unconditional entropy, i.e., the more the known conditions were, the lower the degree of the uncertainty was. The difference between both the entropies could be described as

$$I(pro) = H(pro) - H_r(pro)$$

The equation indicated the reducing of the uncertainty of the proline conformation, i.e., the information provided by the synonymous codons.

$H(pro)$, $H_r(pro)$ and $I(pro)$ of proline in EcoPDB were computed and *t* test was carried out.

3. Results and discussion

3.1. Correlation between the prolyl synonymous codons usage and the local residues near the prolyl peptide bonds

The values of computed $I_m(pro|a)$ could be found in supplementary materials. Fig. 1 was a graphical depiction of $I_m(pro|a)$. Because of the existence of dinucleotide correlation "3–4 correlations" between the third nucleotide in a prolyl synonymous codon and the first nucleotide in the subsequent

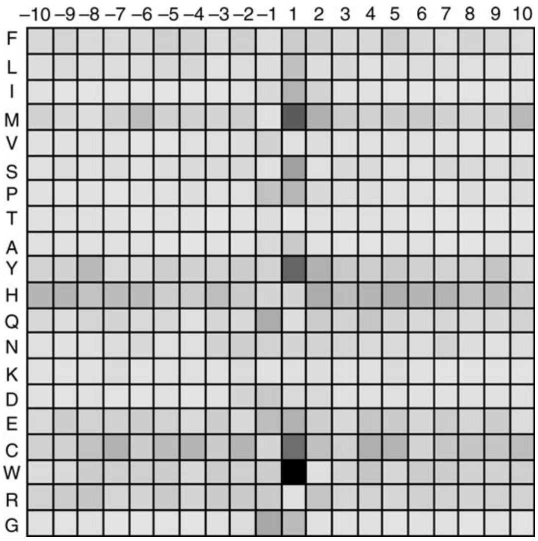

Fig. 1. Influence, on proline synonymous codon usage, of different amino acid residues at different positions. Dark shades indicated high and light shades indicated low influence.

codon [19] that could strengthen the correlation between the prolyl synonymous codons usage and the nearest-neighbor residue of the C-terminal, the influence of this correlation should be removed. So, the value of $I_m(pro|a)$ of the C-terminal nearest-neighbor residue, i.e., the value of $I_m(pro|a)$ when $m = 1$, should be corrected by using $I_m(pro|a) - I(pro|n)$, where n was the first nucleotide in the codon of a. If a was L, S, or R, which has two types of the first nucleotide in the codon, $I(pro|n)$ was averaged according to the proportion of two types of the first nucleotide in the dataset. The values of computed $I(pro|n)$ could be found in supplementary materials. It was depicted in Fig. 1 that the values of $I_m(pro|a)$ of most residues reached their maximum when $m = 1$, i.e., the nearest-neighbor residue of the C-terminal had the greatest influence on the usage of the synonymous codons of proline. And different residues had different degrees of influence. The four residues having the greatest influence in turns were W, M, Y, and C. When $m \neq 1$, the values of $I_m(pro|a)$ of most residues were less than 0.02. But the values of $I_m(pro|a)$ of C, H were always large relatively.

It was presumed that C and H influenced the prolyl synonymous codons usage in global level.

3.2. Correlation between the prolyl synonymous codons usage and the conformations of prolyl peptide bonds

We got 80 *cis* prolyl peptide bonds and 961 *trans* prolyl peptide bonds and divided them into 10 groups randomly. Each group had 100 prolines. $H(pro)$, $H_r(pro)$ and $I(pro)$ of each group were computed. The results were depicted in Table 1.

t test to both groups of $H(pro)$, $H_r(pro)$ was carried out. $t > t_{0.01}$, $P < 0.01$, the difference was significant. So, the prolyl synonymous codons had some factors influencing the conformations of prolyl peptide bonds.

From the above results, it may be concluded that the nearest-neighbor residue of the C-terminal had the greatest influence on the usage of the synonymous codons of proline and the usage of the synonymous codons was correlated with the conformation of prolyl peptide bond.

From these, we can also deduce that the nearest-neighbor residue of the C-terminal influences the conformation of prolyl peptide bond. It has been known that the residue preceding the prolyl bond and local sequence flanking the prolyl bond both influence the conformation of prolyl bond [4,15]. The correlation between the prolyl peptide bond conformation and local residue type has also been measured by computing the Z-scores of residues around *cis* prolyl bond [22]. Those results, based on the abundant difference of local residues actually, were not very consistent with ours. It may be resulted from the complicated multi-factors influencing the prolyl peptide bond conformation. Those results could be found in supplementary materials. Our research result shows that the residue at the first position C-terminal influences the isomeric bond by the codon usage. So, the above relation may be a type of regulation to the protein structure on the gene sequence level in *E. coli* genome. We have been investigating this in more genomes. Though we are not very clear to the biochemical mechanism of this regulation now, the result may be helpful for the protein engineering and the theoretical prediction of protein structure.

Table 1
Entropy values of each group

Group	$H(pro)$	$H_r(pro)$	$I(pro)$
1	0.302538	0.283286	0.019252
2	0.325083	0.306427	0.018656
3	0.253639	0.239694	0.013945
4	0.278769	0.256073	0.022697
5	0.226968	0.216534	0.010434
6	0.198515	0.181888	0.016627
7	0.253639	0.232623	0.021016
8	0.253639	0.216138	0.037501
9	0.198515	0.183797	0.014718
10	0.43967	0.427545	0.012125

References

[1] Jabs, A., Weiss, M.S. and Hilgenfeld, R. (1999) J. Mol. Biol. 286, 291–304.

[2] Stewart, D.E., Sarkar, A. and Wampler, J.E. (1990) J. Mol. Biol. 214, 253–260.

[3] Pall, D. and Chakraabarti, P. (1999) J. Mol. Biol. 294, 271–288.

[4] Reimer, U. and Fischer, G. (2002) Biophys. Chem. 96, 203–212.

[5] Reimer, U., Scherer, G., Drewelo, M., Kruber, S., Schutkowski, M. and Fischer, G. (1998) J. Mol. Biol. 279, 449–460.

[6] Wu, Y. and Matthews, C.A. (2002) J. Mol. Biol. 322, 7–13.

[7] Brauer, A.B., Domingo, G.J., Cooke, R.M., Matthews, S.J. and Leatherbarrow, R.J. (2002) Biochemistry 41, 10608–10615.

[8] Lin, L.N. and Brandts, J.F. (1979) Biochemistry 18, 43–47.

[9] Fischer, G., Bang, H., Berger, E. and Schellenberger, A. (1984) Biochim. Biophys. Acta 791, 87–97.

[10] Brandsch, M., Thunecke, F., Kullertz, G., Schutkowski, M., Fischer, G. and Neubert, K. (1998) J. Biol. Chem. 273, 3861–3864.

[11] Ng, K.K.-S. and Weis, W.I. (1998) Biochemistry 37, 17977–17989.

[12] Stoddard, B.L. and Pietrovski, S. (1998) Nat. Struct. Biol. 5, 3–5.

[13] Charbonnier, J.-B., Belin, P., Moutiez, M., Stura, E.A. and Quemeneur, E. (1999) Prot. Sci. 8, 96–105.

[14] Frömmel, C. and Preissner, R. (1990) FEBS Lett. 277, 159–163.

[15] Wang, M.L., Li, W.J. and Xu, W.B. (2004) J. Pept. Res. 63, 23–28.

[16] Komar, A.A., Lesnik, T. and Reiss, C. (1999) FEBS Lett. 462, 387–391.

[17] Adzhubei, A.A., Adzhubei, I.A., Krasheninnikov, I.A. and Neidle, S. (1996) FEBS Lett. 399, 78–82.

[18] Thanaraj, T.A. and Argos, P. (1996) Prot. Sci. 5, 1973–1983.

[19] Oresic, M. and Shalloway, D. (1998) J. Mol. Biol. 281, 31–48.

[20] Rudd, K.E. (2000) Nucleic Acids Res. 28, 60–64.

[21] Li, W.J. and Song, J.N. (2001) J. Wuxi Univ. Light Indu. (Chin.) 20, 340–343.

[22] Li, W.J. and Zhang, Y. (2001) Acta Sci. Nat. Univ. NeiMongol. (Chin.) 32, 12–16.