Communications in Information and Systems
Volume 18, Number 4, 331–347, 2018

Natural distinct inter-preference between genetic codon and protein secondary structure combinations

Zhuomaji and Xinqi Gong

The central dogma of molecular biology describes the process of genetic information transferred to protein. Many studies have found that genetic codon not only influences the protein amino acid sequence, but also affects protein 3D structure, such as local protein 3D structure may be affected by synonymous codon preferred usage. Here, in addition to the effect of single codons, we furtherly considering the preferences of short codon sequences for protein secondary structures. Also, we studied the preferences of short protein secondary structures for codon sequences. We studied in six cases that how codon combinations with length of N (N-codons) affect protein secondary structure element combinations with the same length (N-secondary structures), where $N = 1, \ldots, 6$. A few distinct codon combination sequences and their corresponding structure sequences were found by calculating Relative Codon Usage (RCU) and Relative Structure Usage (RSU). The preferences of many codon combinations vary for secondary structure combinations when N is different; similar preference patterns were found for protein secondary structure preference for genetic codons. This work is based on the CSandS database. In order to further confirm our conclusion, we selected seven proteins of human that are not in the CSandS to predict its secondary structures from nucleotide sequences using the statistical results. Prediction accuracy can reach 75.72%. It is sufficient to show the imprint of codons on protein structure, and it also indicates that codon usage probably is related to species.

Key words and phrases: codon, short codon sequence, protein secondary structure, short structure sequence, distinct codon combination, predicting secondary structure.

Introduction

The genomic codon is translated into protein by ribosome in living cells, which is a unidirectional flow of explicable information from mRNA sequences to protein sequences. Prediction of the protein structure play an important role in structural bioinformatics [1]. Most methods of predicting the protein structure start with the amino acid sequence. Besides, although numerous articles have linked protein structure to its amino acid sequence, for example, Yau et al. [2] used the natural vector representation to develop a criterion for judging what kinds of amino acid sequences can be natural protein sequences, most of them ignored the effect of codon on structure. In recent years, a great number of evidences demonstrate that synonymous codon affects the protein structure. Comprehensive analyses have also been devoted for connection between protein structure and synonymous codon usage [3, 4]. Study of the relationship between genetic codon and protein structure has led to some databases [3, 5].

In accordance with the literature, there are no universal distinct codons for protein structure in different species, yet a small subset of distinct codons that have been identified to display structural propensity by some studies. Almost all of these researches consider different kinds of amino acids in the study of the relationship between codons and protein structures. Carrying out the Mantel-Haenszel test (Mantel and Haenszel 1959), an asymptotic test based on Chi-square Distribution for analyzing sets of 2×2 contingency tables from the data stratified by codon family and gene. In CSandS database, Saunders et al. [3] showed that the synonymous codons for leucine (Leu), valine (Val), serine (Ser), glutamic (Glu) have different references for the start of helix (H1). Adzhubei [6] also found that such codons by utilizing the genomes to protein (GTOP) database. However, it is worth mentioning that these studies did not consider the effect of adjacent codons on each codon. The GTOP links genetic sequences to protein sequences deposited in the Protein Data Bank (PDB), which is un-curated and theoretical. It is not suitable for accurate research on codon usage, therefore we selected a relatively accurate database, CSandS [3].

Here we map mRNA sequences to protein secondary structure directly. A lot of codons are found to have different preferences for different protein secondary structures. Also, protein secondary structures are found to show preferences for different codons. Furthermore, we observed a few distinct codon sequences and their corresponding structure sequences. Based on statistics, the prediction of the protein secondary structure starts with the mRNA sequence was performed, accuracy of which is 75.72%.

1. Materials and methods

1.1. Coding sequence and structure database

Our analysis is based on the Coding Sequences and Structure (CSandS) database linking the genomic coding sequences to the solved protein secondary structure sequences (`http://portal.stats.ox.ac.uk/userdata/proteins/csands/`). The database contains 4406 protein structures from many different species: human, E.coli, yeast, mouse and so forth, which was compiled by two main tools JOY and BLASTx [3]. Mapping protein secondary structure to its corresponding protein FASTA sequences by using JOY [7], and BLASTx is used to extract the amino acids and coding sequences [8, 9]. It is conducted by the Martin and coworkers [10] based on a PDB to UniProt [11]. The mapping available is outlined (Figure 1). The CSandS database provides a comprehensive and wide-ranging analysis between codon usage and protein structure, which is currently an accurate and unique non-redundant tool in the linked mRNA to protein structure.

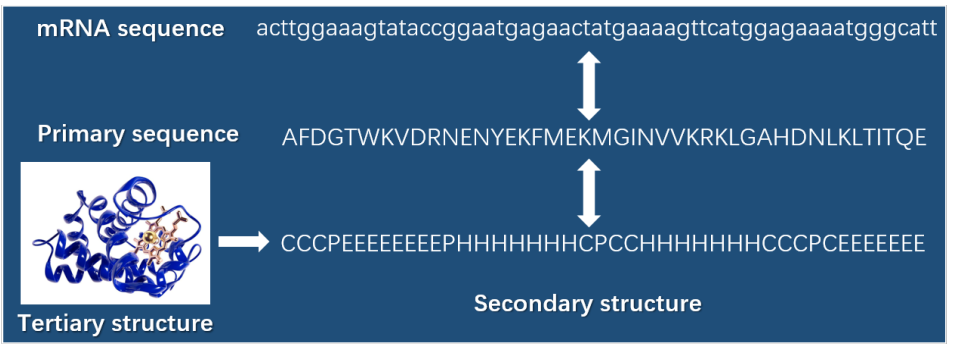

Figure 1: **Data of coding sequence and protein structure from the CSandS database.** mRNA codons are mapped to protein secondary structures through amino acid sequences.

1.2. Relating codons to protein structures

There are four major protein secondary structure classes in the CSandS database, which are strand (E), helix (H), coli (C) and paperclip (P). Saunders et al. [3] defined nine secondary structure classes, and only considered the situation that one codon (referred to as 1-codon) was mapped to one

particular secondary structure (1-structure) (here we denote it as 1-1 situation) through CSandS database. They ignored the influences of adjacent codons on each codon.

For this reason, we assigned N-codons to N-Secondary Structures (written as N-SS below) (N-N situation), where $N = 1, 2, \ldots, 6$, and then to count their frequencies. Here we do not consider different species. We considered the effect of codon combinations to protein secondary structure combinations for the length from 1 to 6. For example, for the length of 1, since three nucleotides encoding one amino acid, the codon combinations from the four single codons, A, C, G, T, will contain $4^3 = 64$ kind of codons; the corresponding secondary structure combinations have 4 kinds, such as, E, H, C and P. For the length of 2, the codon combinations from single codons, A, C, G, T, will contain $4^3 \times 4^3 = 4096$ kinds, such as, AAA, ACT, AAC, ATT, and etc.; the corresponding secondary structure combinations have $4 \times 4 = 16$ kinds, such as, EE, EH, EC, EP, HE, HH, HC, HP, CE, CH, CC, CP, PE, PH, PC, PP.

One of the purposes of this work is to find distinct pair of codon combination (N-codons) and secondary structure combination (N-SS). If $N > 6$ then there are no distinct pairs. we choose the largest $N = 6$ so as to achieve our goal.

1.3. Measures calculations

Subsequently, we counted relative frequency that each codon combination used to encode a structure element combination, and then generated six $n \times m$ contingency tables for N-N situations ($N = 1, \ldots, 6$) respectively. The measure Relative Codon Usage (RCU) relates codon usage to protein structure at the amino acid level, which can measure the preference of the codon for the structure based on these six contingency tables. Another similar measure Relative Structure Usage (RSU) is used to describe the propensity of the protein secondary structure to the codon.

For all 'N-N', the RCU of the jth codon for the ith Secondary Structure (SS) is defined as follows:

$$\mathrm{RCU}_{ij} = \frac{x_{ij}}{\frac{1}{n}\sum_{j=1}^{n} x_{ij}}. \tag{1}$$

Here x_{ij} stands for the number of times of the jth N-codon for the ith N-SS ($i = 1, \ldots, m$, m is the number of arbitrary secondary structure combinations), and n is given by the number of codon combinations (N-codon) for the N-SS. The RCU with a value of 1 means there is no codon usage

preference. A RCU> 1 indicates that the codon usage frequency higher than other codons, and vice versa. We also calculate the Relative Structure Usage (RSU) of each codon combination, it can be described as follows:

$$\text{RSU}_{ij} = \frac{x_{ij}}{\frac{1}{m}\sum_{i=1}^{m} x_{ij}}. \tag{2}$$

In this equation, all parameters are same in Equation (1). RSU= 1 indicates that there is no structure usage preference, A RSU> 1 means that the propensity of the secondary structure is higher than other structures, and vice versa.

Furthermore, if only consider RCU and RSU, leading to extreme situations not suitable for fully explaining the preference. For example, in 5-5 situation, the RCU of the codon GCTCTGTGGGGCGCG is 90394 (largest) to its structure PHEEE, but the corresponding counts is 1, thus we can't say its preference is strong. In order to take into account RSU (and RCU) and the CounT (CT) synthetically, we then calculate their product (referred to as RSU*CT and RCU*CT).

So as to determine whether codon and structure (two variables) are related, we need to test for independence of these two variables. In addition, it should be noted that structure and codon are categorical variables rather than quantitative variables. Accordingly, the chi-squared test, which is a statistical hypothesis test, is undertaken for 1-1 situation. The null hypothesis (H_0) is that codon usage preference is independent of the protein secondary structure encoded by these codons. Under the null hypothesis, we calculate the test statistic in the following way:

$$\chi^2 = \sum \frac{(x_{ij} - E(x_{ij}))^2}{E(x_{ij})}, \tag{3}$$

where x_{ij} is shown in Equation (1), $E(N_{ij})$ denote the expected number of observations that can be described by $\frac{\sum x_{ij}}{n \times m}$. The chi-square statistic follows approximately a chi-squared distribution of $(n-1) \times (m-1)$ degrees of freedom. For this test, we take a significance level of 5% that means if p-value is less than or equal to 0.05, then one rejects the null hypothesis (Figure 2).

2. Results

There are a total of 6373354 secondary structure elements in the database that contains 4406 protein chains, leaving 6300614 by removing some useless

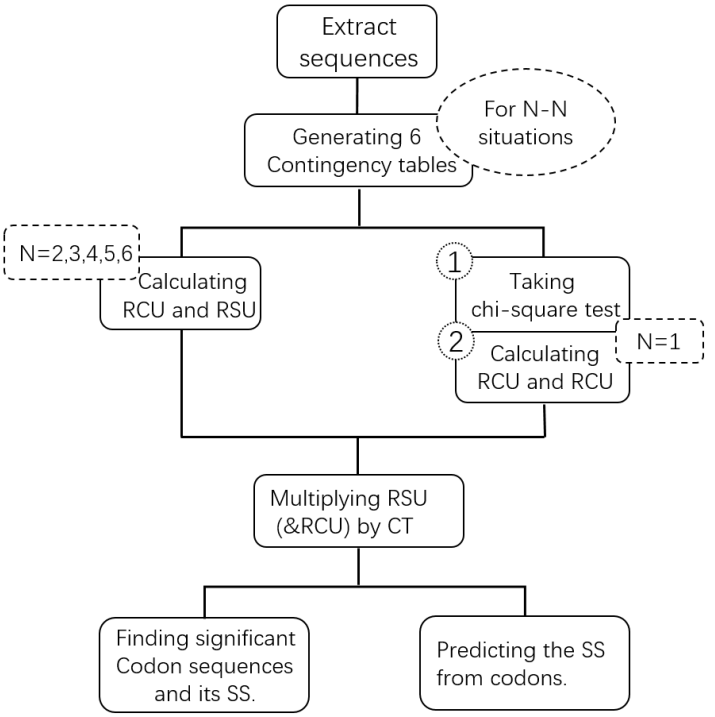

Figure 2: **The flow chart of our study.** From obtaining original mRNA sequence and protein secondary structure data to give out last result. After extracting sequence data from the database, we calculated the counts of the N-codons to the N-Secondary Structures(N-SS) for N-N situations and then generated 6 $n \times m$ contingency tables. Next the chi-square test was taken in 1-1 situation. In order to find the 'distinct' codon sequences and its secondary structures, we calculated each (RCU*CT, RSU*CT). Based on statistics, the protein structure sequences were predicted from mRNA sequences using RSU*CT value. SS is the abbreviation of Secondary Structure here.

records. In the CSandS database, T rather than U is used in the mRNA sequence.

2.1. 1-1 situation

We carry out chi-square test with null hypothesis of independence for the 1-1 situation, results are shown in the following Table 1. It's clear that the

p-value is less than 0.05, so we should reject the null hypothesis. That is, the counts of codons in secondary structure is related to the secondary structure they encoded. In order to examine the strength of association, we then applied a Pearson's Contingency Coefficient (denote as PC) measure. Table 2 shows that the relationship of codon and secondary structure is close (with PC= 0.528). What kind of relationship between the protein structure and codon in 1-1 situation is there? therefore we calculate the RSU*CT and the RCU*CT. By doing this, two distinct codons GGC and GGT have high propensity at paperclips respectively (P, RCU*CT= 1190773 (highest), RCU*CT= 483173.8), a codon CTG have high propensity at helix (H) with RCU*CT= 361007.1, were found within their codon family. Conversely, these two structures P and H also prefers to be encoded by these three codons with high RSU*CT that are 160218.8, 100536.9 and 192821.6 respectively (see table 3). While TAG (stop codon) is under represented at coil (C) (RCU*CT= 0.00224 (lowest), RSU*CT= 1.4382), which means that the codon TAG encoded into structure coil (C) rarely in many species. Especially, the codon ATT and the structure coil (C) has approximately no usage and structure preference with RCU is 1.0425 and RSU is 1.1063 respectively.

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1914659.95[a]	189	.000
Likelihood Ratio	1213539.124	189	.000
N of Valid Cases	4954324		
a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 11.4			

Table 1: **The result of Chi-Square Test for 1-1 situation.** The chi-square test statistic χ^2 is 6627.030, and the p-value is less than 0.05.

On the whole, however, the codons that encoding coil (C) and helix (H) are distinctly more than the other two secondary structures (P and E) (Figure 3) that is different from the preference of several codons mentioned above. Moreover, it is a remarkable fact that the proportion of codon usage of four codons GGC, GGT, GGA, GGG for paperclips (P) is obviously higher than other codons, which also accounts for the largest proportion of all four structures (Figure 3).

Symmetric Measures			
		Value	Approx. sig.
Nominal by Nominal	Contingency Coefficient	.528	.000
N of Valid Cases		4954324	

Table 2: **Pearson's Contingency Coefficient.** Pearson's Contingency Coefficient (PC). PC $= \sqrt{\frac{\chi^2}{N+\chi^2}}$. χ^2 is the chi-square test statistic given before, where $N = n \times m$, n, m is given in Equation (1). PC $\in [0, 1)$, 0 indicating no association, and 1 meaning the opposite.

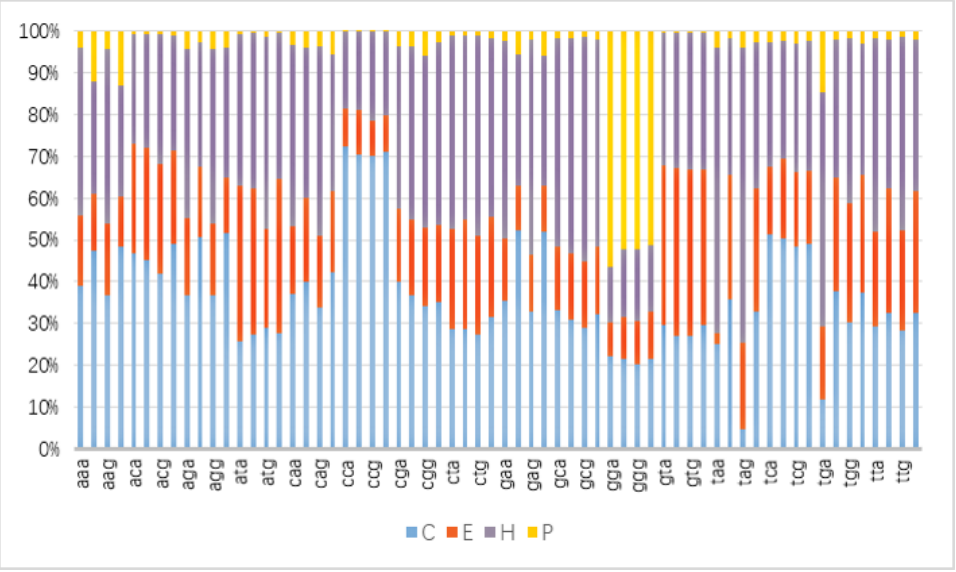

Figure 3: **The proportion of the counts of 61 codons for encoding 4 different protein secondary structures.** It shows the relative proportion of codon usage for the four different kind of secondary structures. Four structures are helix (H), strand (E), coil (C), paperclips (P) respectively.

2.2. N-N situations ($N = 2, \ldots, 6$)

N neighboring codons (N-codons) are mapped to corresponding N protein secondary structure elements (N-SS). Similarly, the RCU*CT and the RSU*CT are also used for describing codon usage and structure usage respectively. There are 12 $n \times m$ contingency tables displaying RCU*CT and

RSU*CT distributions; where n is the number of codon sequences (N-codons) and m indicates the number of secondary structure element sequences (N-SS). We found that these situations differ significantly from the 1-1 situation.

codon	SS	RCU*CT	RSU*CT
ggc	P	1190773	160218.8
ggt	P	483173.8	100536.9
ctg	H	361007.1	192821.6
gatggc	CP	181788.7	28823.1
gtggtg	EE	98628.9	45646.26
agtaacttc	HPP	2948065	14579.400000
aactatcgggga	PCPP	27245131	50106.1
gagagtaacttc	HHPP	17148840	48040.1
tgggacaagaacttc	HPEEC	70808582	110391.428571
aactctggctaccac	ECPCE	53783875	133568.281250
catactgtggataaaaag	CEEPPC	112560188.9	320524.000000
tacggaatcctacagatc	EPCCPE	85718030.58	399716.477816
aacaccaaatatggggat	EHHHPC	104436816.3	309327.524229

Table 3: **The distinct codon combinations (N-codons) and corresponding protein secondary structure combinations (N-SS)**. There are 3, 2, 1, 2, 2, 3 distinct codons and structures respectively in the 1-1, 2-2, 3-3, 4-4, 5-5, 6-6 situations. This result based on Figure 4 (marked with blue wireframe). We choose the N-codons and its N-SS as the 'distinct' that means their value of RSU*CT and RCU*CT are relatively large.

As shown in Figure 4, we know the codon sequence in a propensity for structure change within different N. We found two distinct 2-codons and its corresponding 2-SS TGGC-CP, GTGGTG-EE of 2-2 situation by using the value of RCU*CT and RSU*CT (see Table 3 and Figure 4). For 4-4 and 6-6 situations, there are 2 and 3 such distinct of codon combinations and its structure combinations with both high RCU*CT and RSU*CT (Figure 4 and Table 3). As can be seen from the Figure 4, one distinct of codon combinations and its structure combinations found in both 3-3 and 5-5 situations with extremely high RCU*CT but not RSU*CT (Table 3). where 'distinct' means that the preference of codon for its corresponding protein structure and the preference of secondary structure for its codon are both relatively high.

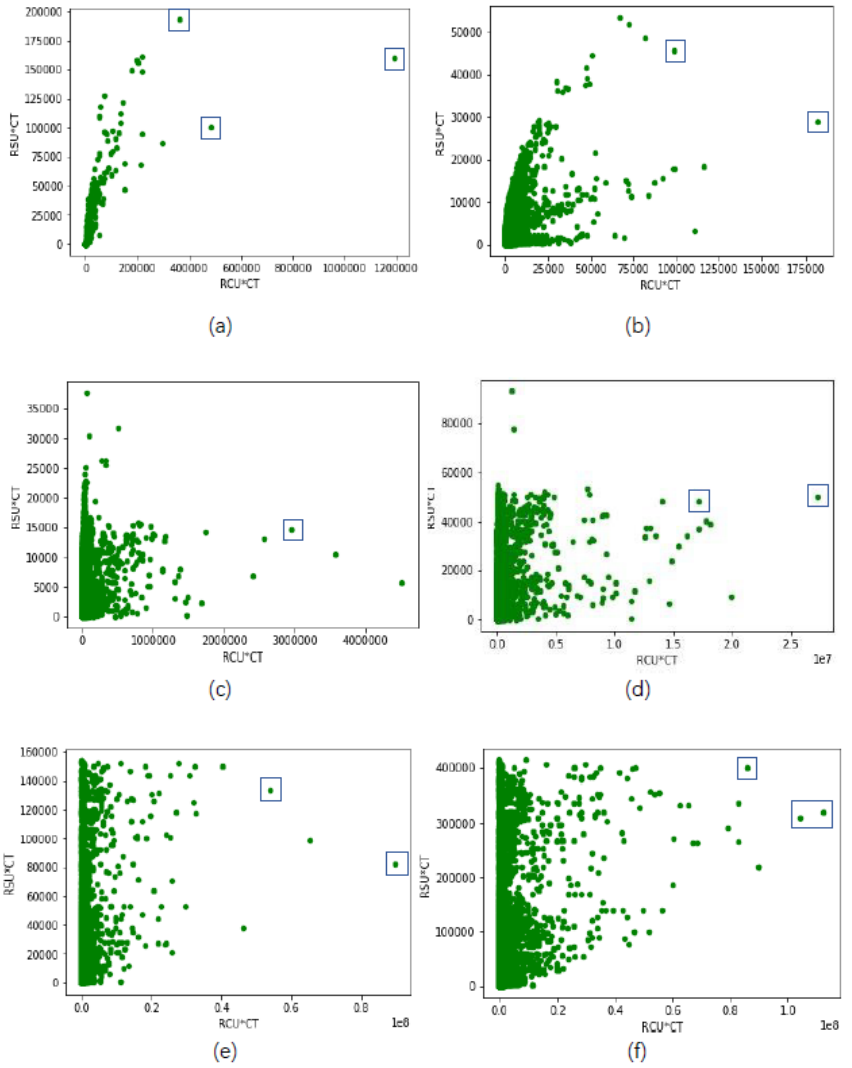

Figure 4: **Result of Multiplying RSU (and RCU) by CT.** (a) 1-1 situation. (b) 2-2 situation. (c) 3-3 situation. (d) 4-4 situation. (e) 5-5 situation. (f) 6-6 situation. There are 3, 2, 1, 2, 2, 3 N-codon and N-SS pairs in from 1-1 to 6-6 situations respectively. It is called 'distinct' means that the short codon sequences and its corresponding short structure sequences have strong preference for each other (listed in Table 3). The distinct N-codons and N-SS marked with blue wireframe.

For example, although there are a total of 18664 6-codons and 60 6-structures of 6-6 situation, only one codon combination AACACCAAATATGGGGAT has the highest probability to encode the secondary structure combination EHHHPC. In contrary, only secondary structure combination EHHHPC most probably encoded by short codon sequence AACACCAAATATGGGGAT. In particular, we select the protein 2HLZ PDB structure to illustrate the example (Figure 5).

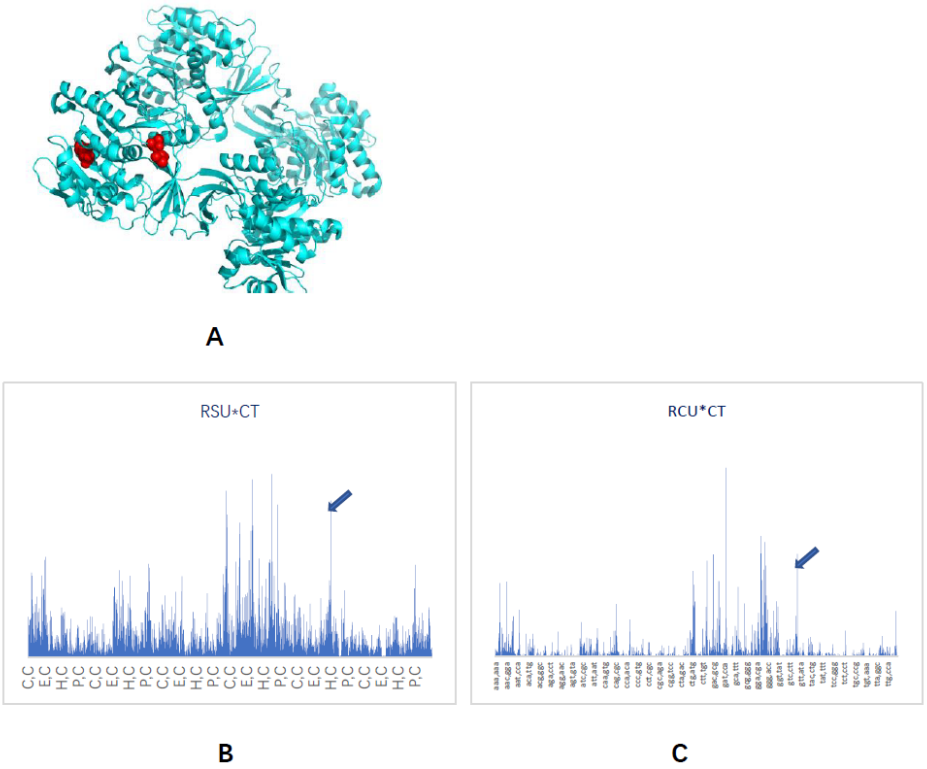

Figure 5: **PDB structure 2HLZ (A) are used to illustrate the example (2-2 situation, the situation of 2-codon aligned with 2-SS).** The positioning of 'distinct' codons GTGGTG and its structures EE of protein 2HLZ are displayed using red spheres representation (A) (RCU*CT= 98628.9, RSU*CT= 45646.26). It is obviously seen from the subgraph B and C, which means that the short codon sequence is over-represented at the short structure sequence EE.

2.3. Predicting

To further illustrate the result that protein secondary structures relate its codons, we conducted a small test: first of all, we take all of the results of RCU*CT and RSU*CT for six situations into a file as a 'dictionary' file. The records of RCU*CT= 0 (RSU*CT= 0 equivalently) are removed from the 'dictionary'. Secondly, we use the dictionary to predict the secondary structure from known mRNA sequence of protein. The process is shown in Figure 6:

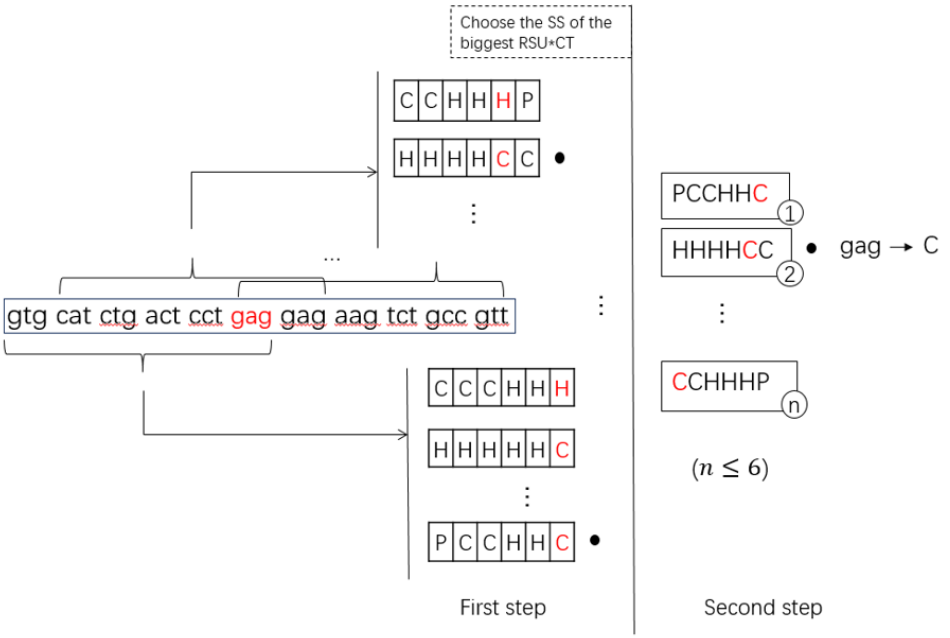

Figure 6: **The process of prediction.** Two steps are taken for this process. For N-N situations, there are N searches performed for a $2N-1$ codons sequence in first step and then some alternative found using the biggest RSU*CT. In the second step, we also use RSU*CT to choose one short structure sequence to predict the Nth codon in this $2N-1$ codons sequence only (marked in red). The process of prediction utilizes from 6- to 1-codon so that all sequences can be predicted. The two steps are repeated until all the sequences have been predicted. All of N-codons was searched against the 'dictionary' file ($N = 1, 2, \ldots, 6$).

The CSandS database contains many different species. Here we randomly choose seven proteins of top 1 represented species human as the target of prediction, which are not in the CSandS database. mRNA sequences of seven proteins were collected from Uniprot database. Their corresponding structure sequences can be downloaded from Pymol database, which is important because for many proteins, the corresponding secondary structure sequences of codon sequences cannot be found in Pymol. The PDB ID of selected seven proteins are 2XNY, 2UZC, 4A9E, 8ICR, 1GNH, 2BX8 and 5XZX. All these proteins with $\leq 30\%$ sequence identity and whose lengths were more than 50 residues. We can take two steps for this process: first step, every nucleotide sequences of the seven proteins was split into 6-codon windows that were also encoded into the same size of the secondary structure windows. The first short codon sequence was searched against the 'dictionary' file. If there are several matched 6-structures found corresponding to this codon sequence, we choose the one with the biggest RSU*CT as an alternative. After that, the second 6-codon from the second codon of a sequence was also searched and used to find another alternative 6-SS. In this way, a total of six sequences were searched against the 'dictionary' file, and a number of 6 or less 6-SS were found, all of which have one common codon; Second step, for these alternative, we also selected the largest RSU*CT one to predict the common codon with corresponding secondary structure (Figure 6); Only one codon can be predicted for eleven codons sequence in two steps of the process at a time. This process is repeated until all the sequences have been predicted. If a six-codon is not in the 'dictionary' file, next we consider 5-5 situation and then 4-4, 3-3, 2-2,1-1 situations will be considered if the previous N-N situation is not found in the 'dictionary' file. That is, the process of prediction utilizes from 6- to 1-codon so that all sequences can be predicted. For all six N-N situations, we use $2N-1$ codons sequence to predict Nth codon at a time in the two steps.

It should be noted that there are only three structure types helix (H), strand (E) and loop (L) in the Pymol database, so we replaced coil (C) and paperclips(P) with loop (L) in the results of prediction. For the seven selected proteins of human that are not in the CSandS database, the prediction accuracy is 75.72%. Although less accurate than other methods such as prediction with amino acid sequence rather than codon sequence, it is enough to illustrate that codons have a significant preference for protein structures.

3. Discussion

Especially, there is a short codon sequence AAAGGCGGC, for example, their RCU*CT is relatively high for structure sequences HEP, but RSU*CT is low (RCU*CT= 28477, RSU*CT= 0.5914). That is, the probability that the 3-SS is encoded by the 3-codon in many three codon combinations is high, but the 3-codon short sequence have a low probability of being encoded into this 3-SS short sequence. In 5-5 situation, two such short codon and structure sequence pairs are also found (GACGTCCAGGCGTGG, PCHPH, RCU*CT= 651082, RSU*CT= 1.809; AAGGGGCATCATGAG, PPCHP, RCU*CT= 651082, RSU*CT= 2.912). This led to a series of problems. Why does this happen? and what kind of biological explanation is there for this phenomenon?

Besides, there is a general situation that the most of codon combinations have no preference for one structure combination, e.g. in the 3-3 situation, 170862 codon combinations can encode structure CCC, only 19.52% of them have preference (RCU$>$ 1) and 80.48% have no preference (RCU$<$ 1) to the structure CCC, which leads a question worth exploring is that what kinds of codon sequences could possibly be protein sequences, so far no one has solved it.

One reason for relatively low accuracy of our predicting method is the data of the CSandS is too small to cover all of the codon combinations. Take 6-6 situation, for example, there are only 834,387 6-codons in the 'dictionary' file, which is only 0.0012% of all the possible 6 codon combinations ($64^6 = 68,719,476,736$). Another point is, the species with the most entries in the CSandS is human, and the prediction accuracy of human protein is higher than other species (Mouse, 60.99%; E.coli, 59.98%; Yeast, 44.47%; BACSU, 43.11%), so we can further speculate that different species probably have different codon preference for protein secondary structure.

4. Conclusion

Here, we extracted the protein secondary structure sequences and its genetic sequences information from CSandS database. Chi-square test demonstrates that there is a relatively high correlation between codons and protein secondary structures for 1-1 situation. We then found some of the 'distinct' short codon sequences and short structure sequences that have distinct propensity to each other by using RCU*CT and RSU*CT value. Our results support codon usage can influence the protein structures. A few codon

combinations have very high preferences for specific protein secondary structure combinations, but many codon combinations have no any preference. Similarly, a few protein secondary structure combinations have very high preferences for specific codon combinations, but many secondary structure combinations have no any preference. In addition, we utilize the value of RSU*CT to predict protein secondary structure sequences from its mRNA sequences for seven proteins of human that are not in the CSandS database. It still further confirms our conclusion and also shows codon usage probably related to different species.

Acknowledgements

This research was supported by National Natural Science Foundation of China (31670725 and 91730301)), and State Key Laboratory of Membrane Biology to Xinqi Gong.

References

[1] *Moult J. Predicting protein three-dimensional structure*, Curr. Opin. Biotechnol. (1999), no. 10, 583–588.

[2] Stephen S.-T. Yau, Wei-Guang Mao, Max Benson, and Rong Lucy He, *Distinguishing proteins from arbitrary amino acid sequences*, Scientific Reports (2015).

[3] Rhodri Saunders and Charlotte M. Deane, *Synonymous codon usage influences the local protein structure observed*, Nucleic Acids Research **38** (2010), no. 19, 6719–6728.

[4] Charlotte M. Deane and Rhodri Saunders, *The imprint of codons on protein structure*, Biotechnol. J. **6** (2011), 641–649.

[5] T. Zhou, M. Weems, CO. Wilke, *Translationally optimal codons associate with structurally sensitive sites in proteins*, Mol. Biol. Evol. **26** (2009), 1571–1580.

[6] A. A. Adzhubei, *Non-random usage of degenerate codons is related to protein three-dimensional structure*, FEBS Lett. **399** (1996), 78–82.

[7] K. Mizuguchi, C. M. Deane, T. L. Blundell, M. S. Johnson, and J. P. Overington, *JOY: protein sequence-structure representation and analysis*, Bioinformatics **14** (1998), 617–623.

[8] L. Duret and D. Mouchiroud, *Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis*, Proc. Natl. Acad. Sci. USA **96** (1999), 4482–4487.

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

[10] A. C. Martin, *Mapping PDB chains to UniProtKB entries*, Bioinformatics **21** (2005), 4297–4301.

[11] UniProt Consortium, *The universal protein resource (uniprot) 2009*, Nucleic Acids Res. **37** (2009), D169–D174.

[12] J. L. Parmley and L. D. Hurst, *Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals*, Mol. Biol. Evol. **24** (2007), 1600–1603.

[13] A. A. Komar, T. Lesnik, and C. Reiss, *Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation*, FEBS Lett. **462** (1999), 387–391.

[14] C. H. Makhoul and E. N. Trifonov, *Distribution of rare triplets along mRNA and their relation to protein folding*, J. Biomol. Struct. Dyn. **20** (2002), 413–420.

[15] W. Gu, X. Sun, and Z. Lu, *Folding type specific secondary structure propensities of synonymous codons folding type specific secondary structure propensities of synonymous codons*, IEEE Transactions Nanobioscience **2** (2003), 150–157.

[16] T. Xie, D. Ding, X. Tao, and D. Dafu, *The relationship between synonymous codon usage and protein structure*, FEBS Lett. **434** (1998), 93–96.

[17] S. K. Gupta, S. Majumdar, T. K. Bhattacharya, and T. C. Ghosh, *Studies on the relationships between the synonymous codon usage and protein secondary structural units*, Biochem. Biophys. Res. Commun. **269** (2000), 692–696.

[18] F. C. Bernstein, T. F. Koetzle, G. J. Williams, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *The Protein Data Bank: a computer-based archival file for macromolecular structures*, J. Mol. Biol. **112** (1977), 535–542.

[19] M. Nakamura, M. Sugiura, *Translation efficiencies of synonymous codons for arginine differ dramatically and are not correlated with codon usage in chloroplasts*, Gene **472** (2011), 50–54.

[20] M. Oresic, M. Dehn, D. Korenblum, and D. Shalloway, *Tracing specific synonymous codon-secondary structure correlations through evolution*, J. Mol. Evol. **56** (2003), 473–484.

[21] M. Jia, L. Luo, and C. Liu, *Statistical correlation between protein secondary structure and messenger RNA stem-loop structure*, Biopolymers **73** (2004), 16–26.

[22] T. Kawabata, S. Fukuchi, K. Homma, M. Ota, J. Araki, T. Ito, N. Ichiyoshi, and K. Nishikawa, *GTOP: a database of protein structures predicted from genome sequences*, Nucleic AcidsRes. **30** (2002), 294–298.

[23] C. L. Wilson, S. J. Hubbard, and A. J. Doig, *A critical assessment of the secondary structure alpha-helices and their termini in proteins*, Protein Eng. **15** (2002), 545–554.

[24] C. L. Wilson, P. E. Boardman, A. J. Doig, and S. J. Hubbard, *Improved prediction for N-termini of alpha-helices using empirical information*, Proteins **57** (2004), 322–330.

[25] M. Oresic, M. Dehn, D. Korenblum, and D. Shalloway, *Tracing specific synonymous codon-secondary structure correlations through evolution*, J. Mol. Evol. **56** (2003), 473–484.

Zhuomaji and Xinqi Gong:
Institute for Mathematical Sciences
Renmin University of China
Beijing 100872, China
E-mail address: xinqigong@ruc.edu.cn