Biomath Communications

www.biomathforum.org/biomath/index.php/conference

Statistical Analysis of Codon Pairs Usage in Prokaryotic Genomes

Ivan Ivanov, Kiril Kirilov
Institute of Molecular Biology, Bulgarian Academy of Sciences,
“Acad. G. Bonchev” Str., Bld. 21, 1113 Sofia, Bulgaria

Keywords: codon pairs, codon usage, codon pairs usage, genetic code

1 Introduction

Genetic information is coded by the collinear arrangement of four nitrogen bases - adenine (A), guanine (G), cytosine (C) and thymine/uracil (T/U) along with the polynucleotide chains of DNA and RNA. They are combined in triplets called *codons*. Each of the 64 codons, except for UAA, UAG and UGA, codes for one (out of twenty) amino acids [1]. The genetic information is translated/decoded by the help of transfer RNAs (tRNAs). The latter bear complementary triplets, called anticodons and also a single covalently bound amino acid. The place of decoding is the *ribosome*. It contains two sites (A and P) for binding of tRNAs [2–4]. Therefore the accommodation of two tRNAs, each carrying one amino acid, in these sites is a prerequisite for the sequential formation of dipeptide, tripeptide and *polypeptide* (protein) products. At every moment of translation the two tRNAs in the A and P ribosomal sites are selected on the basis of the two translating codons in mRNA attached to the same ribosome.

2 Background

Genetic code was deciphered in 1961 [5–7]. Then the meaning of all 61 sense and the three stop codons was determined and the first genetic code dictionary was created. It is proven now that the genetic code is degenerated, i.e.

except for two amino acids (methionine and tryptophan) all the rest (18 amino acids) are coded by more than a single codon. New thoroughfares for investigation of the genetic code were paved after the year 2000 when new powerful and low-cost methods for DNA sequencing were launched for full genome sequencing [8,9]. They allowed the genomes of thousands of prokaryotic (bacterial) and eukaryotic (nuclear) organisms to be sequenced during the last 1–2 decades and enormous amount of sequencing data to be accumulated in the world DNA databases. The latter allowed investigation of the genetic code at a new level determination of the codon usage pattern in different organisms. These studies led to the discovery of a great difference in codon preference between the organisms belonging to different taxonomic groups.

3 Scientific idea

Universal genetic code includes 61 sense and 3 stop/termination codons. In all organisms the 61 sense codons are decoded by 44 to 47 different tRNAs [10]. Bearing in mind that the number of alpha-amino acids in the proteins is 20, this means that 44–47 of all sense codons are decoded by specific/unique tRNAs whereas the rest (14–17 codons) are read by non specific tRNAs, i.e. tRNAs recognizing more than one codon. As already mentioned the formation of a peptide bond between two amino acids in the proteins requires two neighbor codons (*a codon pair*) to be decoded simultaneously, i.e. two tRNAs must occupy the two (A and P) ribosomal sites at one at the same time. We presume that for steric reasons related with the different size and spatial structure of the tRNAs not all combinations of tRNAs by two are equally compatible. Due to this some codon pairs should be *preferable* and others *avoided.* Since the preferable and avoided codon pairs are selected and fixed in the genome during evolution, we expect that the frequency of occurrence of the preferable codon pairs is higher in comparison with the rare codon pairs. Moreover, we postulate also that the frequency of occurrence (i.e. the preference) of the codon pairs might have biological functions. For instance, the codon pairs usage could serves as a modulating factor of translation, i.e. the codon pairs could speed up or delay the translation of mRNA depending on their frequency of occurrence/usage. This hypothesis can be verified by determining the frequency of occurrence of all combinations of codon pairs in the genome. Comparing the most frequently used and avoided codon pairs in organisms belonging to different taxonomic groups we could shed more light on the evolution of genetic code and might reveal the biological function of the genetic code degeneration in the recent organisms. The biological function of codon pairs usage can be easily checked by insertion of synthetic frequently used and rare codon pairs in appropriate expression plasmids to study their effect on gene expression.

4 Genomic databases

For *E. coli* the full sets of 4290 open reading frames (ORFs) and the subset of 2658 protein coding sequences were obtained from the Kyoto Encyclopedia of Genes and Genomes [15]. For the other prokaryotes genomes DNA sequence data were obtained from the DDBJ [16], EMBL [17] and GenBank [18].

5 Methodology

The full set of ORFs and also the subset of experimentally proven protein coding sequences for each organism were analyzed by our own computer program written in Perl (`http://www.bio21.bas.bg/codonpairs/`) [11]. This programme divides the string of codon pairs into two frames, thus mimicking simultaneous codon-anticodon interactions of two adjacent tRNAs on the translating ribosome. The string of codon pairs for each coding sequence begins with an initiator codon (A1) and the second codon (A2), continues with the second and the third (A2, A3), the third and the fourth (A3, A4), etc. and finishes with the combination An:ASTOP (where n is the penultimate codon of the coding sequence, preceding the termination codon. Thus the theoretical number of codon pairs is 3904 of which 3721 are combinations of sense codons (sense:sense) and the rest (183) are combinations of sense and stop (sense:stop) codons. For each codon pair, our programme estimates the following parameters: *Observed Number of Occurrence (NOBS), Expected Number of Occurrence (NEXP), Expected Random Deviation (DEXP), Normalized offset value (r)* (for definitions see below). The routine software applied for these calculations utilizes Fasta format files that carry information about the protein coding regions of the genes only but not for the areas preceding the start and following the stop codon. Taking into considerations the significance of these two gene areas and the fact that their statistical analysis requires new software allowing using other file formats such as the GenBank (gbk), we developed a new programme named Gene Triplet Analysis (GTA) working with gbk files [12]. The GTA programme is written in Java^{TM} SE, uses NetBeans IDE 6.1, and works in BioJava environment with good inter-platform compatibility.

6 Results and discussion

The results in this study are based on the investigation of codon pairs usage in more than 260 organisms having completely sequenced genomes [11]. They belong to the two prokaryotic kingdoms *Archaea* and *Bacteria* and are represented by 2 families for *Archaea* and 13 families for *Bacteria*. In order to reveal the link between codon usage and gene expression, four local DNA databases were created for each organism: a) full set of all protein coding sequences; b) subset of highly expressed genes; c) subset of ribosomal proteins genes (as representatives of extremely highly expressed genes in all living organisms) and d) subset of weakly expressed genes. These databases were explored as described above (see Methodology) and the data were used to compare the codon pairs preferences in: a) different bacterial genomes; b) bacterial strains; c) subsets of genes in individual genomes and also to find correlation between codon pairs usage and prokaryotic taxonomy.

Our study started with the analysis of codon pairs usage in *E. coli* (recognized as a "golden standard" in prokaryotic genetics) and continued with the rest of the bacterial genomes. We found that the frequency of occurrence of the 3904 codon pairs (comprising both sense:sense and sense:stop pairs) in *E. coli* varies between zero and 4913 times per genome. For most of the pairs a significant difference between the real (NOBS) and statistically predicted (NEXP) frequency of occurrence has been observed. We found that codon pairs usage in the two subsets of 334 highly expressed and 303 poorly expressed *E.coli* genes is different [11]. Based on the criterion ΔREG (see Definitions) we classified the codon pairs as *"hypothetically attenuating"* (CUG:GAG, UUU:UUC, CAG:GAG, UUA:CUG, GGU:GUA, GCU:GGU) and *"hypothetically enhancing"* (CAA:GAG, AUG:UGU, GAC:-GUA, GGU:CUG, CCA:AGC) translation. Our results revealed also a great deviation in codon usage pattern between bacterial species belonging to different taxons. This deviation is not randomly spread over the different groups of synonymous codons [13]. Comparing different subset of genes we showed that the codon usage pattern in the low expressed genes is similar to that of the full set of genes, whereas this of the highly expressed and particularly of the ribosomal protein genes is different. Besides some very rare codon pairs, we have identified also 19 missing pairs all of which represented combination of sense and stop codons. Surprisingly, the type of stop codon in these pairs was biased. Except for one pair only (ACU:UGA), where the stop codon was UGA, in all other missing codon pairs the stop codon was UAG. Our analysis revealed also that the sense codons in the missing pairs belonged to the group of rare codons.

This study includes also a comparative analysis of the codon pairs pattern of the 260 prokaryotic organisms mentioned above in the light of the formal/classical bacterial taxonomy (based on morphological and biochemical characteristics). Our results showed that in many cases the type of codon pairs usage did not correlate with this taxonomy [11,14].

The data reported here are available at: `www.bio21.bas.bg/`

7 Verification of hypotheses

Based on our results we postulated two hypotheses that can be verified: *1) Different usage of synonymous codon pairs (encoding the same dipeptide) is due to the non equal compatibility between the decoding tRNAs accommodated in the A and P ribosomal sites at the time of translation; 2) The preferential and avoided codon pairs might serves as modulators (enhancers or attenuators) of translations.*

Both hypotheses can be checked and we made first attempts for their verification. To verify the first hypothesis we searched for correlation between the codon pairs usage in *E.coli* and molecular size of the decoding isoacceptor tRNAs (tRNAs decoding different synonymous codons). Our results definitely demonstrated that the usage of tRNAs for decoding of the different types of codon pairs (preferable or avoided) is biased. We observed tendency for preserving a constant molecular volume/size of the tRNA duplexes decoding respective type of codon pair.

To check the second hypothesis, codon pairs with different frequency of occurrence, including missing in the *E. coli* genome codon pairs were chemically synthesized, inserted at appropriate places in two genes (the genes of chloramphenicol acetyltransferase and human calcitonin) and the genes thus modified were expressed in *E. coli.* Our results demonstrated that the missing codon pairs CCU:UAG (Pro:Stop) and CCC:UAG (Pro:Stop) showed a strong inhibiting effect, whereas another missing pairs such as the CCU:AGG (Pro:Arg) had an opposite effect on gene expression [15].

8 Definitions

Observed Number of Occurrence (N_{OBS}) is the real number of occurrence of an individual codon pair in a protein coding sequence. For instance, the full set of 4289 ORFs in the *E.coli* genome contains 1 358 854 codon pairs and the subset of 2656 real protein-coding sequences includes 906 166 codon pairs respectively.

Expected Number of Occurrence (N_{EXP}) for the sense:sense codon pairs is

$$N_{EXP} = (P_{Ax} P_{Ax+1}) \times NTOT$$

where N_{TOT} is the total number of codon pairs and P_{Ax} and P_{Ax+1} are the probabilities of occurrence of the two individual codons Ax and $Ax + 1$ at any position.

Expected Number of Occurrence (N_{EXP}) for the sense:stop codon pairs is

$$N_{EXP} = (P_{An} P_{ASTOP}) \times 4289,$$

where PAn and $PASTOP$ are the probabilities of occurrence of the penultimate and stop codons, respectively.

Expected Random Deviation (DEXP) for the internal (sense:sense) codon pairs is defined as

$$DEXP = [N_{TOT}(P_{Ax}P_{Ax+1}) \times (1 - P_{Ax}P_{Ax+1})]^{1/2}$$

The expected random deviation (DEXP) for the end terminal (sense:stop) codon pairs is

$$DEXP = [4289(PAnPASTOP) \times (1 - PAnPASTOP)]1/2.$$

Normalized offset value (*r*) measures the difference between the observed and randomly expected values. For the sense:sense codon pairs it is defined as:

$$r = (N_{OBS}N_{EXP})/D_{EXP}$$

and for the sense:stop codon pairs as:

$$r = (N_{OBS}N'_{EXP})/D_{EXP}.$$

The normalized offset value r depends on the deviation of the observed versus expected frequency of occurrence, i.e. when $N_{OBS} > N_{EXP}$, r is positive and conversely, when $N_{OBS} < N_{EXP}$, r is negative.

Δ_{REG} **value** is defined as: $\Delta_{REG} = r_{high} - r_{low}$. It is significant when the sign of r_{high} is opposite to that of r_{low} and their absolute values are greater than two. This means that if the observed frequency is higher than randomly expected in highly expressed genes ($r_{high} > +2$) and lower in poorly expressed genes ($r_{low} < -2$), then ΔREG takes positive sign and a value greater than four. In contrast, if the frequency of occurrence is higher than the randomly expected in poorly expressed genes ($r_{low} > +2$) and lower than in the highly expressed genes ($r_{high} > +2$), then the sign of ΔREG is negative and its absolute value-is greater than four.

References

[1] Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, et al. (2009) Design parameters to control synthetic gene expression in *Escherichia coli*. PLoS One 4: e7002.

[2] Laurberg M, Asahara H, Korostelev A, Zhu J, Trakhanov S, et al. (2008) Structural basis for translation termination on the 70S ribosome. Nature 454: 852–857.

[3] Korostelev A, Ermolenko DN, Noller HF (2008) Structural dynamics of the ribosome. Curr Opin Chem Biol 12: 674–683.

[4] Leger M, Dulude D, Steinberg SV, Brakier-Gingras L (2007) The three transfer RNAs occupying the A, P and E sites on the ribosome are involved in viral programmed -1 ribosomal frameshift. Nucleic Acids Res 35: 5581–5592.

[5] Nirenberg M (2004) Historical review: Deciphering the genetic code–a personal account. Trends Biochem Sci 29: 46–54.

[6] Yanofsky C (2007) Establishing the triplet nature of the genetic code. Cell 128: 815–818.

[7] Nirenberg MW (1963) The genetic code. II. Sci Am 208: 80–94.

[8] Van Borm S, Belak S, Freimanis G, Fusaro A, Granberg F, et al. (2015) Next-generation sequencing in veterinary medicine: how can the massive amount of information arising from high-throughput technologies improve diagnosis, control, and management of infectious diseases? Methods Mol Biol 1247: 415–436.

[9] McElhoe JA, Holland MM, Makova KD, Su MS, Paul IM, et al. (2014) Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. Forensic Sci Int Genet 13: 20–29.

[10] Anjay A (2012) The Genetic Codes. National Center for Biotechnology Information (NCBI), Bethesda, Maryland, USA.

[11] Bachvarov B, Kirilov K, Ivanov I (2008) Codon usage in prokaryotes. Biotechnology & Biotechnological Equipment 22: 669–682.

[12] Kirilov K, Ivanov I (2012) A programme for determination of codons and codons context frequency of occurrence in sequenced genomes. Biotechnology & Biotechnological Equipment 26: 3310–3314.

[13] Boycheva S, Chkodrov G, Ivanov I (2003) Codon pairs in the genome of *Escherichia coli*. Bioinformatics 19: 987–998.

[14] Kirilov KT, Golshani A, Ivanov IG (2013) Termination codons and stop codon context in bacteria and mammalian mitochondria. Biotechnology & Biotechnological Equipment 27: 4018–4025.

[15] Boycheva SS, Bachvarov BI, Berzal-Heranz A, Ivanov IG (2004) Effect of 3? Terminal Codon Pairs with Different Frequency of Occurrence on the Expression of cat Gene in *Escherichia coli*. Current microbiology 48: 97–101.

[16] `www.tokyo-center.genome.ad.jp/kegg/kegg.html`

[17] `http://www.ddbj.nig.ac.jp/`

[18] `http://www.ebi.ac.uk/embl/`

[19] `http://www.ncbi.nlm.nih.gov/Genbank/index.html`