SEARCH OF POSSIBLE READING FRAME SHIFTS IN GENES

E. KOROTKOV
Centre of Bioengineering RAS
Moscow, RUSSIA
e-mail: genekorotkov@gmail.com

Abstract

The definition of a phase shift of triplet periodicity (TP) is introduced. The mathematical algorithm for detection of a phase shift of TP in nucleotide sequences has been developed. The gene sequences from the Kegg-46 data bank were have been analyzed to search for genes with TP phase shifts. The presence of a phase shift of TP for 318329 genes (~10 % from the total number of genes in the Kegg-46 data bank) has been demonstrated. We show that TP phase shifts are determined by the shifts of a reading frame (RF) in genes. The connection between phase shifts of TP and RF shifts in genes is discussed.

1 Introduction

Mutations in gene sequences arise as substitutions, deletions and insertions of DNA bases and as deletions, insertions and inversions of whole DNA fragments [1-6]. Substitutions of DNA bases can induce substitutions of amino acids in proteins and a substitution of one base can change only one amino acid in amino acid sequences. These amino acid alterations often have a strong influence on a protein structure and a protein ability to accomplish the biological function. However deletions or insertions of DNA bases can change a long region of the amino acid sequence because of a shift of the RF if a size of the deletion or the insertion is not a number divisible by 3. In this case all amino acid sequences below the point of the RF shift are changed. Therefore deletions and insertions can be considered as more important evolutionary events than base substitutions. The better understanding of the RF shift influence on the protein structure and functions is possible if we could develop a mathematical method for more complete detection of RF shifts in the known gene sequences.

The present work is solving the next problems. It is attempt is to find all genes in the Kegg data base with TP phase shifts. This work uses a new mathematical approach for revealing TP phase shifts in DNA. The Kegg-46 data bank was analyzed by the developed mathematical method and we have found 318329 genes with statistically important TP phase shifts which can show the presence of the RF shifts in these genes.

2 Algorithm of searching for the TP phase shift

Let x shows the position of a base $s(x)$ in a sequence S and let x be chosen as $L_1 + 3n$, where $n = 0, 1, \ldots, (L - L_1)/3$, where L_1 is divisible by 3 and lies in the interval from

60 to 600 bases. Let us consider the subsequence $S(x-L_1+1,x)$. For this subsequence we calculate the matrix of TP $M_1(x-L_1+1,x)$ for RF of T_1 in a sequence S. The subsequences $S(x+1,x+L_1$, $S(x+2,x+L_1+1)$ and $S(x+3,x+L_1+2)$ are also considered, and for these subsequences we calculate the TP matrices $M_1(x+1,x+L_1)$, $M_2(x+2,x+L_1+1)$ and $M_3(x+3,x+L_1+2)$ for RF of T_1, T_2 and T_3, respectively, in a sequence S. If a shift of RF on 1 or 2 bases occurs just after the position x in a sequence S, then the matrix $M_1(x-L_1+1,x)$ is more similar to $M_2(x+2,x+L_1+1)$ or to $M_3(x+3,x+L_1+2)$ matrix. If a shift of RF after position x is absent, then the matrix $M_1(x-L_1+1,x)$ is more similar to the matrix $M_1(x+1,x+L_1)$. Then for each of the four TP matrices another matrix was calculated which elements were the arguments of the normal distribution. Each element of such matrix was calculated using the following formula:

$$n(i,j)=\frac{m(i,j)-Lp(i,j)}{\sqrt{Lp(i,j)\,(1-p(i,j))}},\ p(i,j)=\frac{x(i)y(j)}{L^2},$$

where $m(i,j)$ is the element of a matrix M_1, M_2 or M_3, $n(i,j)$ - normally distributed value. As a result, we obtain for each of the matrices $M_1(x-L_1+1,x)$, $M_1(x+1,x+L_1)$, $M_2(x+2,x+L_1+1)$ and $M_3(x+3,x+L_1+2)$ the matrices V_1, W_1, W_2 and W_3. The differences between the matrix V_1 and each of the matrices W_1, W_2 and W_3 were calculated as:

$$D(1,k)=\sum_{i=1}^{4}\sum_{j=1}^{3}\left(\frac{v_1(i,j)-w_k(i,j)}{\sqrt{2}}\right)^2$$

for $k=1$, 2 and 3. In the capacity of a function U which allows to make the conclusion regarding the differences of two matrices of TP, we selected function $D(1,k)$. The value $D(1,k)$ is distributed as χ^2 with 6 degrees of freedom if the matrices calculated for random sequences are being compared. Then we calculated three probabilities of the fact that the random value distributed as χ^2 with 6 degrees of freedom will be greater than or equal to $D(1,k)$, $k=1,2,3$. Let us denote these probabilities P_{11}, P_{12}, P_{13}. If two matrices are similar, then the value of D equals zero and the value of P equals 1.0. In a case when two matrices being compared are different, the value of D will be greater than zero and the value of P will be less than 1.0.

The values P_{11}, P_{12} and P_{13} are used for further calculations used in searching for the shifts of TP phase. If the sequences S do not have insertion or deletion of bases (with a length of insertion or deletion not divisible by 3) after position x, then $P_{11}>P_{12}$ and $P_{11}>P_{13}$. If insertion with a length equal to $Q=3i+1$ or deletion with a length equal to $Q=3i+2$, $(i=0,1,\ldots)$ is present, then we say that a transition from RF T_1 to RF T_2 is present after position x. Shift of the TP phase on 1 base can be observed in this case, and then $P_{12}>P_{11}$, $P_{12}>P_{13}$. If insertion with a length equal to $Q=3i+2$ or deletion with a length equal to $Q=3i+1$, $(i=0,1,\ldots)$ is present, then we may say that a transition from RF T_1 to RF T_3 is present. Shift of the TP phase on 2 bases can be observed in this case, and then $P_{13}>P_{11}$, $P_{13}>P_{12}$. It is suitable to use the values $F_1=-log_{10}(P_{11}/P_{12})$, $F_2=-log_{10}(P_{11}/P_{13})$ for searching the shifts of TP phase.

We have varied the L_1 for each position x. The variation was carried out for searching such value of L_1 that gives the maximal values of F_1 or F_2. It allows decreasing the

influence of random noise on F_1 or F_2 because the TP can be different in subsequences of the sequence S. We varied the L_1 within the interval from 60 to 600 bases for each x position, and the step of variation was equal to 3 bases.

After all, we built two graphs for a sequence S on which the maximal values of F_1 or F_2 were shown for each position x. Each maximal value F_1 or F_2 was calculated for some value of L_1. We selected the positions x in which we have found the local maximum for F_1 or F_2. If the value of a local maximum for F_1 or F_2 is greater than some threshold F_0, then we consider that the sequence S has a shift of TP at position x. The threshold F_0 was calculated by using Monte-Carlo method (see section 3).

3 The method of Monte-Carlo for determination of the threshold F_0

We used gene sequences from the Kegg-46 data bank for determination of the thresholds F_0 and F_{00}. When using F_0, the probability that the TP phase shift is caused by the random factors is 18 %, while for F_{00} such a probability equals 6 %. We may say that if F_1 or F_2 is greater than F_0, then the sequence contains the TP shift, and if F_1 or F_2 is greater than F_{00}, then the sequence almost definitely contains such a shift. The total number of genes in this release of Kegg was 3318628. We make the random data bank of gene sequences by the way of mixing up the sequences of each gene. It allows keeping the same distribution gene lengths and same base composition of genes as in Kegg data bank. We divided the gene sequence into three subsequences for keeping the TP in a random gene sequence on the original level. The first subsequence (denoted as C_1) was obtained from a gene sequence by choosing bases which were at the positions equal to $i = 1+3n$. The second and the third subsequences C_2 and C_3 were created by choosing bases which were at the positions $i = 2 + 3n$ and $i = 3 + 3n$, $n = 0, 1, \ldots, L/3 - 1$.

Then we made, by using the random number generator, the sequences of random numbers R_1, R_2 and R_3 which have the length equal to $L/3$. We performed the sorting in ascending order of sequences R_1, R_2, R_3 and recorded the order of permutation made in each sequence. After it we permutated the bases in sequences C_1, C_2 and C_3 as we have done it for the sequences R_1, R_2 and R_3 in the ascending order. We produced the random sequence R which had R_1 sequence at the positions $i = 1 + 3n$, R_2 sequence at the positions $i = 2 + 3n$ and R_3 sequence at the positions $i = 3 + 3n$, $i = 0, 1, \ldots, L/3 - 1$. The length of a sequence R was equal to L and it has the same base composition as a gene sequence. We repeated this procedure for all genes from Kegg-46 data bank.

After producing the random data bank which has the same distribution of gene lengths and TP as Kegg-46 data bank, we selected two levels of F (F_0 and F_{00}) equal to 3.0 and 4.0, respectively, and calculated the number of genes which have one or more local maxima for F_1 or F_2 greater than F_0 and F_{00}, respectively. This number was calculated for Kegg-46 data bank ($N1$) and for random data bank ($N2$). We found that $N2 \approx 0.06N1$ and $N2 \approx 0.18N1$ for the levels F_{00} and F_0, correspondingly.

4 Results and discussion

We have analyzed 3318628 genes from Kegg data bank, release 46 (www.genome.ad.jp). The total number of genes with F_1 or $F_2 > F_0$ was 318329, while with F_1 or $F_2 > F_{00}$ 174879. Genes with a single shift of TP phase constituted up to 90 % from the total number of genes with a shift of TP phase. Remaining 10 % genes contained more than one case of TP phase shift. In the bank of random sequences (section 3) we have found 58916 and 11063 shifts of TP for F_0 and F_{00}, respectively. This comparison shows that most part of TP phase shifts revealed in Kegg-46 data bank have nonrandom nature for the levels F_0 and F_{00}.

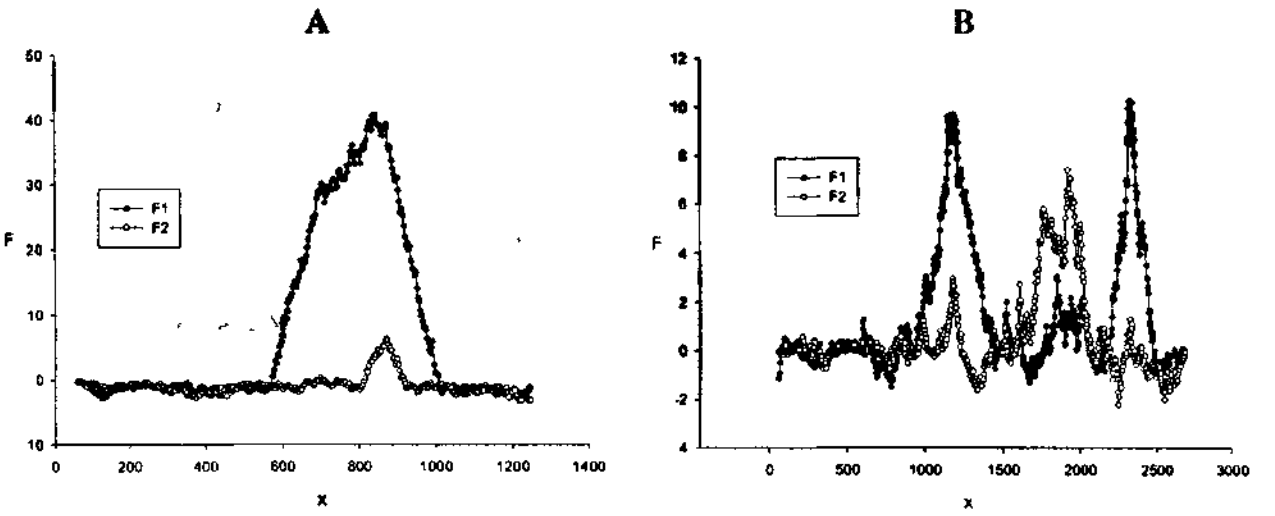

Figure 1: **A**. The functions F_1 and F_2 are shown against position x for the sequence GSU2494 from Kegg data bank. This gene is coding the amino acid sequence of the cytochrome C from *G.sulfurreducens* genome (Swiss-Prot entry is Q74A96_GEOSL). **B**. The functions F_1 and F_2 are shown against position $x$for the sequence F15A2.6 from Kegg data bank. This gene is coding the amino acid sequence of the BR serine/threonine kinase from a genome of *C.elegans*.

An example of a gene with single shift of TP phase is shown in the Fig. 1A. It is a gene of cytochrome from *G.sulfurreducens* (Swiss-Prot entry is Q74A96_GEOSL). As it can be seen in the Fig. 1, the gene has local maximum of F_1 for position 879, whereas the values F_2 are not greater than 1.0 for all positions. It means that TP after 879^{th} base has moved to the beginning of the gene by one base relatively to the TP existing in gene from the 1^{st} to 879^{th} base. It also means that the shift between TP and RF takes place after 879^{th} base. This shift corresponds to the deletion of one base or insertion of two bases after position 879. We can not exclude the possibility of deletion of DNA fragment with a length equal to $Q = 3i + 1$ or insertion of DNA fragment with a length equal to $Q = 3i + 2$, where $i = 1, 2, \ldots$. It could be the reason why the upper part of the pear in Fig. 1A is somewhat smooth. A second example of a gene with shifts of TP phase is shown in the Fig. 1B. The dependencies of F_1 and F_2

from position x are shown in the Fig 1B for gene sequence F15A2.6 from Kegg-46 data bank from genome *C.elegans*. This gene is coding the BR serine/threonine kinase. As it can be seen from the Fig. 1B, this gene contains no less than 4 shifts of TP phase. It is possible to select four positions in gene sequences: 1195, 1816, 1951 and 2326. The deletion or insertion of DNA fragment with a length equal to $Q = 3i+1$ or $Q = 3i+2$ is possible at first, second and fifth positions. The deletion or insertion of DNA fragment with length equal to $Q = 3i + 2$ or $Q = 3i + 1$ is possible at third and fourth positions, where $i = 0, 1, \ldots$. The shifts of a TP phase in this gene are well expressed and values of F_1 and F_2 are much more than 4.0.

We may suppose that in the present work we have found the lowest boundary for the number of genes in which the shift between RF and TP is possible. In reality, this number may be larger. However, even this number (~10 %) is much greater than the fraction of genes with RF shift found earlier (~1 %) [1-3, 5]. This fact indicates that searching for RF by Blast program may be not very effective. Most likely this is related to the fact that the similarity of amino acid sequences may be insignificant due to large number of changes in them or that such amino acid sequences are simply not included in the database. It is the reason why the approach applied for the search of RF shifts, being under its further developing, seems more preferable than application of the similarity search with a help of Blast program or similar programs. Our approach does not require any additional data and is based on the gene sequences only. There always exists a possibility that the similarities are absent in the data bank because the volume of amino acid sequence data bank is limited but the RF shift existing in gene sequence. We think that complete revelation of RF shifts will be possible if we combine our method with similarity search methods. It means that we should study the genes having $F_1 > F_0'$ or $F_2 > F_0'$, where $F_0' < F_0$. The shift of RF for these local maxima can be found if the statistically significant similarities exist for amino acid sequences produced for RF $T2$ and $T3$. Relatively small increase of F_1 or F_2 may indicate the possibility of RF shift in this case, and the fact of RF shift could be proved by the similarity search. On the other hand, the improving of the algorithm applied in present work can be directed on the using of more perfect methods for TP search, like hidden Markov models, for example. We think that the revealing of the RF shifts will be possible in various gene regions even for a large number of insertions and deletions.

It is also important to consider an issue whether the TP phase shift would always indicate the RF shift in a gene. In principle, this may not be claimed with a 100 % probability since there is always some small possibility left that the phase shift of gene's TP is caused by purely random factors (~18 % for F_0) or that the phase shift is caused by the interchange of alpha-helixes and beta-layers in the structures like $\alpha\alpha$, $\beta\beta$, $\alpha\beta$ and $\beta\alpha$ in a protein, where α is an alpha-helix, β is a beta-layer, or by some other secondary or tertiary protein structures. However, if the last statement was true, we would observe the shifts of TP phase in the significantly larger fraction of amino acid sequences since such structures do also occur in proteins in which we do not observe the TP phase shift.

Errors of gene sequencing could produce the shifts between TP and RF also, and it could be in prokaryotic and eukaryotic genes. Accuracy of sequencing is approximately

from 10^{-3} to 10^{-4} in present time. However the situation is much better if the sequencing was done by some times or new gene was annotated and amino acid sequence coding by this gene has an important similarities with another amino acid sequences in Swiss-prot data bank. In this case the probability of mistake become insignificant. It is true for prokaryotic genome firstly where the considerable part of genes are annotated by the search of similarities with known amino acid sequences. It is a reason why we may consider the TP phase shifts in bacterial genes as have no relation with sequencing mistakes. Also, for eukaryotic genes a wrong determination of exon-intron borders is often possible. As a consequence, some fraction of the shifts found between TP and RF for eukaryotic genes could be the result of these errors. But prokaryotic genes do not have introns and analysis of prokaryotic genes permits to exclude the influence of errors in exon-intron border determination. We found that ~4.4 % of prokaryotic genes possess the shifts between TP and RF. It is less than 10 % determined for the genes from the whole Kegg-46 data bank. The reasons for this difference could be the errors in determination of exon-intron borders (or influence of an alternative splicing), greater number of sequencing errors in eukaryotic genes than in prokaryotic ones, and greater speed of shift accumulations in eukaryotic genes.

In the light of these proposals. the TP could be some characteristic for natural testing of gene integrity in the genome [6]. If gene was duplicated in the genome, then such a natural testing of the new copy could be skipped, and this opens the possibilities for evolutionary changing of the gene copy by the way of RF shift and creating the gene with a new biological function as a result.

References

[1] Raes J., Van de Peer Y. (2005) Functional divergence of proteins through frameshift mutations. *Trends Genetics*. Vol. **21**, pp. 428 431.

[2] Okamura K. et al. (2006) Frequent appearance of novel protein-coding sequences by frameshift translation. *Genomics*. Vol. **88**, pp. 690-697.

[3] Claverie J.M. (1993) Detecting frame shifts by amino acid sequence comparison. *J Mol Biol*. Vol. **234**, pp. 1140-1157.

[4] Frenkel F.E., Korotkov E.V. (2008) Classification analysis of triplet periodicity in protein-coding regions of genes. *Gene*. Vol. **421**, pp. 52-60.

[5] Kramer E.M. et al. (2006) A simplified explanation for the frameshift mutation that created a novel C-terminal motif in the APETALA3 gene lineage. *BMC Evolutionary Biology*. Vol. **6**, pp. 30-36.

[6] Trifonov E.N. (1987) Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. *J. Mol. Biol.* Vol. **194**, pp. 643-652.