Study of the triplet periodicity phase shifts in genes

Eugene V. Korotkov[1] and Maria A. Korotkova[2]

[1] Bioinfomatics laboratory, Centre of Bioengineering, Russian Academy of Sciences, 117312, Prospect 60-tya Oktyabrya, 7/1, email:genekorotkov@gmail.com

[2] Cybernetics department, Moscow Physical Engineering Institute, 115409, Russia, Moscow, Kashirskoe Shosse, 31, email:bioinf@rambler.ru

Abstract

The definition of a phase shift of triplet periodicity (TP) is introduced. The mathematical algorithm for detection of TP phase shift of nucleotide sequences has been developed. Gene sequences from Kegg-46 data bank were analyzed with a purpose of searching genes with a phase shift of TP. The presence of a phase shift of triplet periodicity has been shown for 318329 genes (~10% from the number of genes in Kegg-46). We suppose that shifts of the TP phase may indicate the shifts of reading frame (RF) in genes. A relationship between the phase shifts of TP and the frame 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 the whole DNA fragments [1-2]. Substitutions of DNA bases can induce substitutions of amino acids in protein, and a substitution of one base can change only one amino acid in amino acid sequence. These amino acid alterations very often have a strong influence on protein structure and protein ability to perform its biological function [3]. However, deletions or insertions of DNA bases can change long amino acid sequence due to the shift of a reading frame (RF) in a case when a size of deletion or insertion is not divisible by 3. Amino acid sequences downstream the point of RF shift are changed. In this sense deletions and insertions can be considered as more important evolutionary events than base substitutions. The influence of RF shifts on protein function has been studied relatively poorly due to the difficulty of detection of RF shifts. Though, it is very interesting to study the influence of RF shifts on protein function. If protein does not lose its biological function because of RF shift, it is possible to suggest two hypotheses. Firstly, RF shift may change the unimportant part of protein, and this change of amino acid sequence can not influence the protein function. Secondly, we can suppose that RF shift could create amino acid sequence with similar amino acid function. It is very interesting to find the laws of changes of the amino acids in amino acid sequence that allow to create new amino acid sequences with the same function as initial one has. If upon RF shift the biological function of amino acid sequence has been changed, it is very interesting to reveal the types of amino acid sequence changes which have lead to development of a new biological function of a protein. The results of such studies could be used for designing artificial proteins having the necessary biological functions.

The better understanding of the RF shifts influence on protein structure and function will be possible if we develop the mathematical method for the better detection of RF shifts in the known gene sequences. Currently the main method of searching for RF shifts is based on searching for similarities between amino acid sequences with a help of BLAST program or similar programs [4-8]. To search for a similarity we should find the gene region in which we suppose the RF shift to occur. Then we should translate this gene region according to the new

frame into the hypothetical amino acid sequence. After this we search for the similarities of this hypothetical amino acid sequence in Swiss-prot data bank. We can say that the analyzed gene has a RF shift if we have found the statistically significant similarities in Swiss-prot data bank for a hypothetical amino acid sequence, i.e. if we have found that such hypothetical amino acid sequence does really exist. Some hundreds of genes have been revealed by these methods in which we can suppose the presence of frame shifts with a large probability [4-8].

This method of searching for the frame shifts in genes has some limitations. Firstly, we should choose, using some features, the gene, in which we can assume the presence of a frame shift, and then we should find the gene region in which this frame shift is possible. The total search of the frame shifts for all known genes may require large computer capacities. Secondly, even if we could solve the first problem, the Swiss-prot data bank should contain the amino acid sequences which have the statistically significant similarity with the hypothetical amino acid sequence. But these sequences can be absent because of limited volume of the data bank or since the amino acid sequences have accumulated the high level of differences. As a result, the use of similarity search can reveal only some part of the RF shifts that currently exist in genes.

For more successful revealing of the RF shifts, it is necessary to develop the alternative method for searching of RF shifts to replace the similarity search between hypothetical and real amino acids sequences. The shift of a phase of triplet periodicity (TP), as we will show in the present work, can be a signal of a presence of RF shifts in genes. TP of DNA coding sequences is a common property of all known living organisms [9-19], and it is associated with the gene reading frame [20,21]. The reasons of relation between RF and TP are the genetic code structure (which is almost identical for prokaryotes and eukaryotes), saturation of proteins with certain amino acids, and uneven using of the synonymous codons [22-25]. If a shift of RF takes place in gene, and at the same time a nucleotide sequence of a gene has TP, then this shift can be observed as a shift of the phase of TP (Fig 1). This shift remains for a long time because the TP of DNA sequence is difficult to be changed by a few number of DNA substitutions [26]. The presence of a phase shift of TP in gene sequence may serve as an indicator of possible RF shift in the analyzed gene.

At a present time the methods for searching of TP have been developed which use the regularities of nucleotide preferences in different positions of triplets of gene sequence. Fourier transformation, hidden Markov models and other statistical methods were used for revealing TP [22-33]. The methods applied were directed on searching for coding regions of genes and separating them from the non-coding regions. The method of information decomposition has been proposed later for searching TP [26-27]. It allows introducing the definition of a class of triplet periodicity as a matrix with dimensions 4x3. The columns represent the positions in triplets, and the rows represent the DNA bases.

Two problems are being solved in the present work. Firstly, we would like to find all genes from Kegg data base having a phase shift of TP. In this paper, mathematical approach for revealing of a phase shift of TP in DNA sequence is developed. Two sequences located one after another, each having the length from 60 to 600 bases, are selected for the search of a shift of TP phase, wherein the length of sequences is divisible by 3. First base of the first and the second sequence always corresponds to the first base of the codon of the RF existing in gene. Then we calculate four matrices of triplet periodicity [26-27]. First matrix corresponds to the RF in the first sequence, while the second one corresponds to the RF in the second sequence. Two matrices left are calculated with a shift on one and two bases, respectively, in the second sequence, and these two matrices correspond to two alternative RFs in the second sequence. It is possible to say that the phase shift of TP is observed between two sequences if the first TP matrix is more similar to the third or fourth TP matrix than the second one. The Kegg data bank was analyzed by the mathematical method developed, and we have found

318329 genes with statistically significant phase shift of TP which may indicate the presence of RF shifts in these genes.

Secondly, we would like to test an assumption that hypothetical amino acid sequences (created with using RF defined by the TP) have the similarities with amino acid sequences from Swiss-Prot data bank. We performed such a testing for regions of genes where the shift of TP phase has been found. We have confirmed the existence of shifts of RF for a part of genes because the statistically significant similarities between some hypothetical amino acid sequences and amino acid sequences from Swiss-prot data bank have been found.

2 Methods and Algorithms

2.1 Conception of the phase of triplet periodicity

Let us consider a coding nucleotide sequence $S = \{s(k), k = 1, 2, ..., L\}$, where each base $s(k)$ is selected from the alphabet $A=\{a,t,c,g\}$, L is the length of a sequence S, and it is divisible by 3. Let us introduce three RF in sequence S and let us designate them as T_1, T_2 and T_3. The base $s(1)$ of a sequence S represents the first, second and third base of codon for RF T_1, T_2 and T_3, correspondingly. The RF T_1 does really exist in a sequence S while RFs T_2 and T_3 are the hypothetical ones. Let us introduce the three matrices of the TP $M_1(i_1, i_2)$, $M_2(i_1, i_2)$ and $M_3(i_1, i_2)$, which are the matrices of TP calculated for RF T_1, T_2 and T_3 for the region of a sequence S from the base i_1 to the base i_2. Let us denote this subsequence as $S(i_1, i_2)$. The elements of matrices $m_1(i,j)$, $m_2(i,j)$ and $m_3(i,j)$ show the number of the bases of a type i in a sequence S (i=1 for a, i=2 for t, i=3 for c, i=4 for g,) which are in j position of a coding triplet (j equals 1,2 or 3) for RF T_1, T_2 and T_3, respectively (Korotkov et.al., 2003; Korotkov et.al., 2003a). Let us suppose that the initial phase of TP determined by matrices M_1, M_2, M_3 equals to the coordinate k of that base from $s(1)$, $s(2)$ and $s(3)$, which is a member of the first position of the first coding triplet of RF T_1, T_2 and T_3, respectively. According to this, the initial phase of TP determined by matrices M_1, M_2, M_3 equals to 1, 2 and 3, respectively.

Then we should define the conditions which show us the existence of TP phase shift after the base $s(x)$ in a sequence S. Firstly, TP should exist in the sequence S. Conditions of TP presence and quantitative measure for revealing TP in a sequence S, or in any its subsequence, are shown below in chapter 2.2. Secondly, we should introduce the quantitative measure of a difference between TP matrices. Let us introduce some function U and let us assume that two TP matrices are similar to each other if $U \le U_0$. Otherwise we consider two TP matrices as different matrices. The detailed description of the function U is given in a chapter 2.3. We assume that a sequence S after base $s(x)$ has a phase shift of TP on 1 base if next conditions are implemented simultaneously:

$$\left\{\begin{array}{l} U\{M_1(1,\ x),\ M_2(x+1,\ L)\} \le U_0 \\ U\{M_1(1,\ x),\ M_1(x+1,\ L)\} > U_0 \\ U\{M_1(1,\ x),\ M_3(x+1,\ L)\} > U_0 \end{array}\right\} \quad (1)$$

We assume that a sequence S after base $s(x)$ has shift of TP phase on 2 bases if the following conditions are valid simultaneously:

$$\left\{\begin{array}{l} U\{M_1(1,\ x),\ M_3(x+1,\ L)\} \le U_0 \\ U\{M_1(1,\ x),\ M_1(x+1,\ L)\} > U_0 \\ U\{M_1(1,\ x),\ M_2(x+1,\ L)\} > U_0 \end{array}\right\} \quad (2)$$

If the following conditions are valid:

$$\left\{\begin{array}{l} U\{M_1(1,\ x),\ M_1(x+1,\ L)\} \le U_0 \\ U\{M_1(1,\ x),\ M_2(x+1,\ L)\} > U_0 \\ U\{M_1(1,\ x),\ M_3(x+1,\ L)\} > U_0 \end{array}\right\} \quad (3)$$

than we assume that phase of TP after base *s(x)* remains without a change , i.e. shift of TP phase equals zero in this case. Shift of a phase is calculated as difference of initial phases of TP matrices for subsequences $S(x+1,L)$ and $S(1,x)$ which are the members of first formula of the conditions (1-3). Shift of TP phase on 1 or 2 bases corresponds to insertion of $1+3n$ or $2+3n$ bases (or deletion of $2+3n$ or $1+3n$ bases*)* after the bases s(x)*, n=0,1,2,3…*

2.2 Triplet periodicity of a sequence *S*

TP matrices can be considered as cross-tabulated data [34]. Let us consider the M_1 matrix. For the matrices M_2 and M_3 all conclusions will be analogous. The rows of a matrix M_1 represent the bases of a sequence *S*, while the columns represent the base positions in coding triplets of the RF T_1. We conclude that the sequence *S* has a TP if there is a correlation between bases and base positions. We may suggest that this correlation exists if mutual information I_3 between bases of a sequence *S* and base positions in the coding triplets is greater than some value I_0 [21]. Mutual information is calculated with a help of formula [34]:

$$I_3 = \sum_{i=1}^{4}\sum_{j=1}^{3} m_1(i,\ j)\ln m1(i,\ j) - \sum_{i=1}^{4} x(i)\ln x(i) - \sum_{j=1}^{3} y(j)\ln y(j) + L_1 \ln L_1 \quad (4)$$

where $x(i) = \sum_{i=1}^{3} m_1(i,\ j)$, $y(j) = \sum_{i=1}^{4} m_1(i,\ j)$. The doubled mutual information ($2I_3$) has a χ^2 distribution with 6 degrees of freedom that allows estimating the statistical significance of TP found. The value of I_3 can be approximately transformed into standard normal distribution as:

$$X_3 = \sqrt{4I_3} - \sqrt{2n-1} \quad (5)$$

The accordance of $2I_3$ to the χ^2 distribution with 6 degrees of freedom and of X_3 to the standard normal distribution is observed for sufficient volume of statistical data, i.e., for sufficiently large length of a sequence *S*. We have tested the accordance of $2I_3$ to χ^2 distribution for different lengths of a sequences *S* to determine the minimum value of *L* for which it is still possible to use χ^2 distribution. The random number generator was used to generate the sets of random DNA sequences for each length from the interval from 30 to 1000 bases. Each set contained 10000 random sequences. Mutual information was calculated for all sequences from each set, and distribution of $2I_3$.was calculated for each set. Then this distribution was compared with the theoretical distribution of χ^2. The distribution of $2I_3$ corresponds to χ^2 distribution for the length $L>60$ bases with a probability greater than 95%. All sequences considered in the present work were longer than 60 bases. This allows using of the χ^2 distribution in the present work for statistical estimations of probability that $2I_3$ is present in the interval from some $2I_0$. to ∞. We searched for the shift of TP phase in a sequence *S*, if for both subsequences $S(1,x)$ and $S(x+1,L)$ the value X_3 was greater than 2.0. It gives the probability less than 0.05 that the TP is random.

2.3 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 $n=0,1,2,3,\ldots(L-L_1)/3$, 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 $M_1(x\text{-}L_1\text{+}1,\ x)$ for RF of T_1 in a sequence S. The subsequences $S(x\text{+}1,\ x\text{+}L_1)$, $S(x\text{+}2,\ x\text{+}L_1\text{+}1)$ and $S(x\text{+}3,\ x\text{+}L_1\text{+}2)$ are also considered, and for these subsequences we calculate the TP matrices $M_1(x\text{+}1,\ x\text{+}L_1)$, $M_2(x\text{+}2,\ x\text{+}L_1\text{+}1)$ and $M_3(x\text{+}3,\ x\text{+}L_1\text{+}2)$ 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 $M_1(x\text{-}L_1\text{+}1,\ x)$ is more similar to $M_2(x\text{+}2,\ x\text{+}L_1\text{+}1)$ or to $M_3(x\text{+}3,\ x\text{+}L_1\text{+}2)$ matrix. If a shift of RF after position x is not present, then the matrix $M_1(x\text{-}L_1\text{+}1,\ x)$ is more similar to the matrix $M_1(x\text{+}1,\ x\text{+}L_1)$. 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:

$$n(i,\ j) = \frac{m(i,\ j) - Lp(i,\ j)}{\sqrt{Lp(i,\ j)(1 - p(i,\ j))}} \tag{6}$$

$$p(i,\ j) = \frac{x(i)y(j)}{L^2} \tag{7}$$

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 $M_1(x\text{-}L_1\text{+}1,\ x)$, $M_1(x\text{+}1,\ x\text{+}L_1)$, $M_2(x\text{+}2,\ x\text{+}L_1\text{+}1)$ and $M_3(x\text{+}3,\ x\text{+}L_1\text{+}2)$ 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:

$$D(1,k) = \sum_{i=1}^{4}\sum_{j=1}^{3}\left(\frac{v_1(i,\ j) - w_k(i,\ j)}{\sqrt{2}}\right)^2 \tag{8}$$

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 [34]. 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=0,1,2,…) 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=0,1,2,…) 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 $F_1=-\log_{10}(P_{11}/P_{12})$ and $F_2=-\log_{10}(P_{11}/P_{13})$ 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 Chapter 2.4).

Besides of conditions given above, we should be sure that subsequences $S(x\text{-}L_1+1, x)$, $S(x+1, x+L_1)$, $S(x+2, x+L_1+1)$ and $S(x+3, x+L_1+2)$ possess a TP. It is the reason why we take for consideration such subsequences $S(x\text{-}L_1+1, x)$, $S(x+1, x+L_1)$, $S(x+2, x+L_1+1)$ and $S(x+3, x+L_1+2)$ that possess explicit TP (see Chapter 2.2)

2.4 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, $n=0,1,2\ldots, L$/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 and 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, $i=0,1,2,\ldots, L$/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 (as it is considered in the chapter 2.3) 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.

3 Results and Discussion

3.1 The search of genes with TP phase shift in Kegg data base

We have analyzed 3318628 genes from Kegg data bank [35], release 46 (http://www.genome.ad.jp/kegg/). 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 (chapter 2.4) 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}.

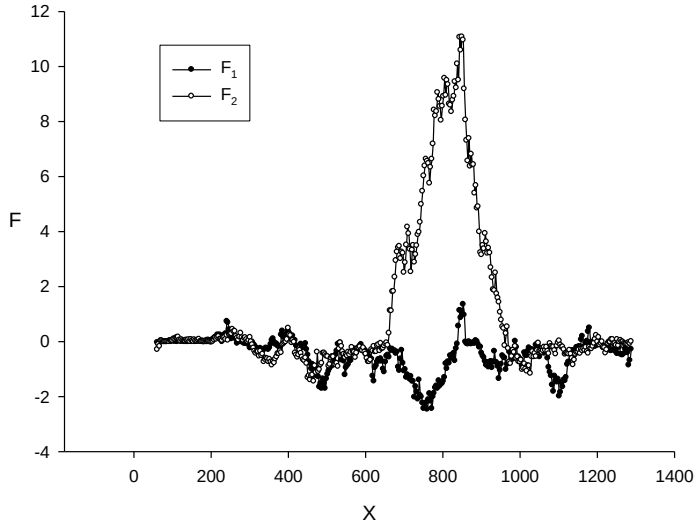

Fig.1. The functions F_1 and F_2 are shown against position x for the sequence BSU24300 from Kegg data bank. This gene codes the amino acid sequence of the exodeoxyribonuclease VII large subunit from *B.subtilis* genome (Swiss-Prot entry is EX7L_BACSU).

An example of a gene with a single shift of TP phase is shown in the Fig. 1. It is a gene of cytochrome C from *B.subtilis*. As it can be seen from the Fig. 1, the gene has local maximum of F_2 for position 850, whereas the values F_1 are not greater than 2.0 for all positions. It means that the shift between TP and RF takes place after 850^{th} base. This shift corresponds to the deletion of one base or insertion of two bases after position 850. 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. 1 is somewhat smooth.

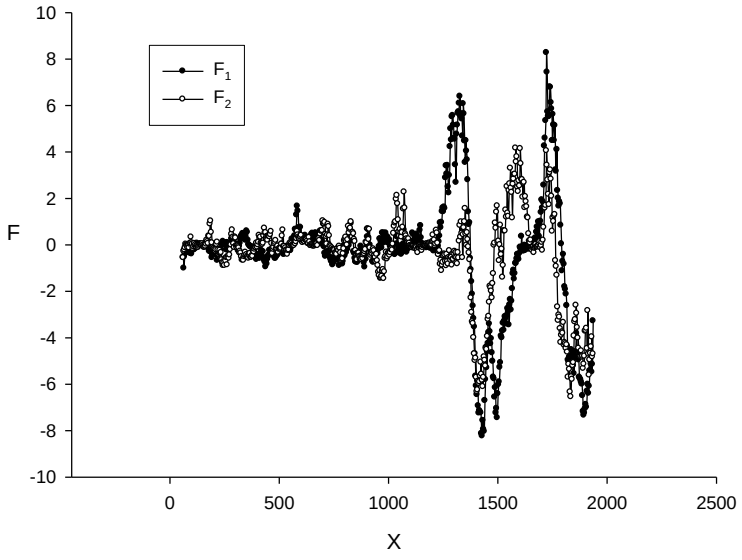

Fig.2. The functions F_1 and F_2 are shown against position x for the sequence YPK_3724 from Kegg data bank. This gene codes the amino acid sequence of the DEAD/DEAH box helicase domain protein from a genome of *Y.pseudotuberculosis (YPIII)*. (Swiss-Prot entry is B1JLX4_YERPY).

A second example of a gene with shifts of TP phase is shown in the Fig.2. The dependences of F_1 and F_2 from position x are shown in the Fig 2 for the gene sequence YPK_3724 from Kegg-46 data bank from a genome of *Y .pseudotuberculosis_YPIII*. This gene codes the DEAD/DEAH box helicase domain protein. As it can be seen from the Fig 2, this gene contains no less than 3 shifts of TP phase. It is possible to select three positions in gene sequences: 1327, 1582 and 1723. The insertion or deletion of DNA fragment with a length equal to $Q=3i+1$ or $Q=3i+2$ is possible at first and third positions. The insertion or deletion of DNA fragment with length equal to $Q=3i+2$ or $Q=3i+1$ is possible at second position, where $i=0,1,2,\ldots$. The shifts of a TP phase in this gene are well expressed and values of F_1 and F_2 are much more than 2.5.

It is very interesting to see the distribution of biological functions of the revealed genes possessing the shifts of TP phase. The list of top 20 functions is shown in Tabl.1 (for F_0). Pseudogenes have the largest number of shifts of TP phase, and this number equals 9554. It is not so strange because pseudogenes do not have a functional significance and mutations in pseudogenes are neutral events for cell genome and they can accumulate rapidly [1-3]. The second position in Table 1 has genes in which the shift of RF was observed earlier in other investigations. We have revealed 844 of these genes. This fact is additional argument that shifts of RF in real sequences can be found by the way of searching the shifts of TP phase.

Table 1. The number of genes with TP phase shifts grouped by the identical description of biological function in Kegg data bank.

№	Number of genes	Definition
1	9554	Pseudogene
2	844	frameshift
3	348	mucin-associated surface protein (MASP), putative
4	335	translation initiation factor IF-2
5	290	trans-sialidase, putative
6	282	ABC transporter related
7	274	transposase
8	263	dispersed gene family protein 1 (DGF-1), putative
9	248	major facilitator superfamily MFS_1
10	224	PE-PGRS family protein
11	221	Protein kinase, putative
12	213	serine/threonine protein kinase
13	191	Protein kinase domain containing protein
14	185	transcriptional regulator, LysR family
15	183	PPE family protein
16	175	hypothetical membrane protein
17	172	Putative transposase
18	169	Acriflavin resistance protein
19	163	Protein kinase
20	147	putative transmembrane protein

3.2 Search of similarities for amino acid sequences

Let us consider the genes in which we have revealed the shift of TP phase for F_0. Let us label as x_0 the position i in which the shift of TP phase have occurred. We can consider that there is coincidence between TP and RF for $i<x_0$. If it is so, then the shift between TP and RF is

observed for $i>x_0$ in a sequence *S*. We suppose that TP determines the RF which has existed in gene before the shift of RF, and we refer to this RF as to the ancient one. Thus in the sequence *S* there are two RFs – one really exists in a gene and the other is the hypothetical RF chosen on the base of the TP, or the ancient RF. If the shift of RF took place in a gene not so long ago, then the alternative version of this gene may remain in other species without shift of RF. For $i>x_0$ we can obtain two amino acid sequences for two RFs. The first is a really existing amino acid sequence, and the second is the hypothetical amino acid sequence which we refer to as the ancient amino acid sequence. We produced these two amino acid sequences for each gene sequence with TP phase shift. If some gene had several TP phase shifts, then we generated the ancient amino acid sequences from the first shift to the second shift, from the second shift to the third shift and so on. If the return to existing RF took place upon some shift, then such shifts were skipped. It means that we performed full reconstruction of TP shifts in gene. All these sequences were compared with amino acid sequences from Swiss-prot data bank [36] by using the Blast program [37]. Consequently, for 179427 pairs of sequences we have not found any similar amino acid sequences. The similarities only for amino acid sequences obtained for real gene's RF were observed for 188924 pairs. 9309 pairs of sequences have the similarity for hypothetical amino acid sequences, and 17811 pairs of sequences have similarity for both amino acid sequences. Database *E*-value for Blast program (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html) was selected to be equal 10^{-4} that gives about 32 random similarities for comparison of 318329 random amino acid sequences with Swiss-prot. This number shows that the fraction of random similarities in the total number of the similarities found is insignificant. The confidence interval (95%) for the number of random similarities is {14 - 50} when using the binomial distribution. We have also tested this number by the computer calculations. We produced 318329 random sequences with equal lengths and compared these sequences with the sequences from the Swiss-prot data bank. We found 58 random similarities and this value is close to 32 similarities expected from calculations using the database *E*-value.

3.3 Discussion

In the present work we show that the study of TP phase shifts can reveal the shifts of RF in genes. We have revealed 318329 prokaryotic and eukaryotic genes in which we may suppose the existence of TP phase shift, which constitutes ~10% from the total number of the analyzed genes. The triplet periodicity and RF were uniquely related to each other [20,21] and some shift between them could be created due to insertions and deletions. We reveal this shift as a shift of TP phase in a gene. We think that the fraction of genes with a TP frame shift may really be much greater. The reasons why we have revealed not all the TP phase shifts are the following. Firstly, the mathematical approach developed is able to reveal relatively short deletions and insertions. The long insertions and deletions can destroy both the TP itself (4) and a similarity of TP matrices which we estimate by the (8). Thereby the developed method misses the considerable part of genes holding long (more 50 bases) insertion or deletion that can produce the shift of RF. Secondly, the method applied works good for relatively low number of deletions and insertions in gene. If the density of insertions and deletions will be more than one event for some tens of bases, then revealing the shifts of TP phase is not always possible. It gives the situation where F_1 or F_2 are less than the threshold value F_0 for the gene. Thirdly, we use the relatively high level of triplet periodicity for sure revealing of shifts of TP phase (X_3, (5)). The study of shifts of TP for the lower levels of X_3 probably may allow finding more genes which have F_1 or F_2 larger than the threshold value F_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%) [4-8]. 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 TP phase shift would always indicate the RF shift in a gene. Firstly, errors of gene sequencing could produce the shifts between TP and RF, but it could be in prokaryotic and eukaryotic genes. Secondly, 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 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. The number of prokaryotic genes with shifts between TP and RF are shown in Table 2 for some bacterial genomes. As it can be seen, ~3.6% of prokaryotic genes possess shifts between TP and RF. It is several times 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, greater number of sequencing errors in eukaryotic genes than in prokaryotic ones, and greater speed of shift accumulations in eukaryotic genes.

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 αα, ßß, αß and ßα in a protein, where α is an alpha-helix, ß 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.

If we suppose that all or the most part of found TP phase shift is caused by the RF shift, then this process contributes more to the evolution of genes and proteins than it was considered earlier. This means that by no means all mutations arising in genes by RF shift are fatal for the proteins. From the functional point of view, the shifts of RF may be considered as events that can considerably change the function of the gene and the protein coded by this gene. Nevertheless, the genetic code should be adapted for these events [22,38-39], and new amino acid sequences should possess some new or keep an old function upon the shift of RF. Otherwise the existence of genes with RF shifts is difficult to be explained in general because these genes should lose their function and be eliminated from a genome.

Table 2. Number of genes with a shift between TP and RF in some prokaryotic genomes.

N	Genome	Number of genes in Kegg data base	Number of genes(pseudogenes) with a shift between TP and RF
1	B.avium	3510	100(15)
2	B.mallei	5508	263(86)
3	B.subtilis	4225	145(0)
4	E.coli	4667	256(14)
5	L.fermentum	1912	53(0)
6	P.aeruginosa	5651	79 (1)
7	S.aureus_col	2727	58 (16)
8	S.enterica_choleraesuis	4895	236 (1)
9	S.pneumoniae	2303	57 (28)
10	S.sonnei	4809	309 (70)
11	V.cholerae	4008	143 (0)
12	X.campestris	4242	222 (0)
13	Y.pseudotuberculosis_ypiii	4305	231 (5)
14	M.capsulatus	3052	85(0)
15	S.typhimurium	4732	216(0)
16	A.vinelandii	5222	121(24)

In the light of these proposals, the TP could be some characteristic for natural testing of gene integrity in the genome. 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] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, K. and P. Walter. Molecular Biology of the Cell. New York: Garland Science, 2002.

[2] H. Maki. Annual Review of Genetics, 36: 279-303, 2002..

[3] J.D. Watson, T.A. Baker, S.P. Bell, A. Gann, M. Levine and R. Losick *Molecular Biology of the Gene, 6th Edition.* San Francisco: CSHL Press & Benjamin Cummings, 2007.

[4] K. Okamura, L. Feuk, L., T. Marquès-Bonet, A. Navarro, and S. W. Scherer, Genomics, 88, 690-697, 2006.

[5] J. Raes and Y. van de Peer, Trends Genetics, 21, 428-431, 2005.

[6] E.M. Kramer, H-J. Su, C.C.Wu and J.M. Hu, BMC Evolutionary Biology, 6, 30-36, 2006.

[7] G.A. Fichant and Y. Quentin, Nucl. Acids Res., 23, 2900-2908, 1995.

[8] J.M.Claverie, J Mol Biol., 234,1140-1157, 1993.

[9] V.R. Chechetkin and A.Y. Turygin, J Theor Biol., 178, 205-215 ,1996.

[10] V.J. Makeev and V.G. Tumanyan, Bioinformatics, 12, 49-54, 1996,

[11] J.W. Fickett, Methods Biochem Anal. 39, 231-245, 1998.

[12] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E, 51, 5084-5091, 1995.

[13] A.D. Baxevanis, Methods Biochem Anal. 43, 233-252, 2001.

[14] G. Gutiérrez, J.L. Oliver and A. Marín, J. Theor. Biol., 167, 413-414, 1994.

[15] J. Gao, Y. Qi and Y. Cao and W.W. Tung, J. Biomed. Biotechnol., 2, 139-146 (2005).

[16] C. Yin and S.S. Yau, J. Theor. Biol., 247, 687-694, 2007.

[17] S.T. Eskesen, F.N. Eskesen, B. Kinghorn and A. Ruvinsky, BMC Mol. Biol., 5, 12 2004.

[18] I. Grosse, H. Herzel, S.V. Buldyrev and H. E. Stanley, Phys. Rev. E, 61, 5624-5629, 2000.

[19] A.K. Konopka, In: *Biocomputing: Informatics and genome projects*. Ed. Smith, San Diego: Acad. Press, 119-174, 1994.

[20] D.G. Arques and C.J. Michel, J.Theor. Biol., 182, 45-58, 1996.

[21] F.E. Frenkel and E.V. Korotkov, Gene, 421, 52-60 (2008).

[22] E.N. Trifonov, J. Mol. Biol., 194, 643-652, 1987.

[23] M. Eigen and R. Winkler-Oswatitsch, 1981. Naturwissenschaften, 68, 217-228, 1981.

[24] M. Zoltowski Conf. Proc. IEEE Eng. Med. Biol. Soc., 1, 1383-1386, 2007.

[25] M.A. Antezana and M. Kreitman, J. Mol. Evol., 49, 36-43, 1999.

[26] E.V. Korotkov, M.A. Korotkova, F.E. Frenkel and N.A. Kudryashov, Molek. Biol. (Russian) 37, 372-386, 2003.

[27] E.V. Korotkov, M.A. Korotkova and N.A. Kudryashov, Physics Lett. A., 312, 198-212, 2003.

[28] B.Issac, H. Singh, H. Kaur and G.P.S. Raghava, Bioinformatics, 18, 196-197, 2002.

[29] S.Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya and R. Ramaswamy, Comput. Appl. Bioscie., 13, 263-270, 1997.

[30] R.K. Azad and M. Borodovsky, Briefings in Bioinformatics, 5, 118-130, 2004.

[31] J. Henderson, S. Salzberg and K.H. Fasman, J. Comput. Biol. 4, 127-141, 1997.

[32] E.E. Snyder and G.D. Stormo, Nucleic Acids Res., 21, 607-616, 1993.

[33] A. Thomas and M.H. Skolnick, IMA J. Math. Appl. Med. Biol., 11, 149-160, 1994.

[34] S. Kullback S., Information Theory and Statistics, New York: Wiley, 1959.

[35] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono and M. Kanehisa, Nucleic Acids Res., 27, 29-34, 1999.

[36] UniProt Consortium, Nucl. Acids Res., 35, 193-197, 2007.

[37] S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, J. Mol. Biol., 215, 403-410, 1990.

[38] T. Bollenbach, K. Vetsigian and R. Kishony, Genome Res., 17, 405-412, 2007.

[39] E.N. Trifonov, Ann NY Acad Sci., 870, 330-338, 1999.