Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

Method

An Approach for Searching Insertions in Bacterial Genes Leading to the Phase Shift of Triplet Periodicity

Maria A. Korotkova[1], Nikolay A. Kudryashov[1], and Eugene V. Korotkov[1,2*]

[1]*National University of Nuclear Investigations (MIFI), Moscow 115409, Russia;*
[2]*Centre of Bioengineering, Russian Academy of Sciences, Moscow 117312, Russia.*

Genomics Proteomics Bioinformatics 2011 Oct; 9(4-5): 158-170 DOI: 10.1016/S1672-0229(11)60019-3

Received: Mar 14, 2011; Accepted: Aug 02, 2011

Abstract

The concept of the phase shift of triplet periodicity (TP) was used for searching potential DNA insertions in genes from 17 bacterial genomes. A mathematical algorithm for detection of these insertions has been developed. This approach can detect potential insertions and deletions with lengths that are not multiples of three bases, especially insertions of relatively large DNA fragments (>100 bases). New similarity measure between triplet matrixes was employed to improve the sensitivity for detecting the TP phase shift. Sequences of 17,220 bacterial genes with each consisting of more than 1,200 bases were analyzed, and the presence of a TP phase shift has been shown in ~16% of analysed genes (2,809 genes), which is about 4 times more than that detected in our previous work. We propose that shifts of the TP phase may indicate the shifts of reading frame in genes after insertions of the DNA fragments with lengths that are not multiples of three bases. A relationship between the phase shifts of TP and the frame shifts in genes is discussed.

Key words: triplet periodicity, insertion, gene sequence, reading frame, phase shift, change-point problem

Introduction

Small insertions of DNA fragments in genes can take place rather frequently (*1, 2*). If the lengths of these insertions are not multiples of three bases, it may lead to the shift in the reading frame after the insertion site. These insertions can significantly change the amino acid sequence coded by the gene and it is important to understand their contribution to the generation of reading frame changes (*3-5*). At present, the mathematical methods used to find changes of the reading frame can be divided into two groups. Both of these groups share the same feature—additional information is required besides the DNA sequence being considered. The first group of methods needs external data including the amino acid sequence data bank and uses special software for searching similarities (*6-9*). When these algorithms are used, the amino acid sequences corresponding to the alternative reading frames are created, and then are searched for their similarities in the database. If a similarity is found, then we can say that a shift of the reading frame has occurred in the analyzed gene. A data bank of amino acid sequences is necessary for this group of methods. The second group of methods uses the nucleotide sequence of the analyzed gene to find

*Corresponding author.
E-mail: genekorotkov@gmail.com

shifts of the reading frame. A set of gene sequences that have the shifts of the reading frame is used as additional information (*10-14*). As a result, a search is made in the analyzed gene for some common properties intrinsic to DNA sequences in which such shifts have already been found. Such collective properties can be described in various ways, including creation of the weight matrix, calculation of k-tuple frequencies, development of the HMM models, utilization of the neural networks and application of other mathematical approaches (*10-14*).

However, the requirement of additional information limits the application of such methods. For the methods of the first group, these limitations include the need for presence in the database of amino acid sequence, which could exist before the formation of the reading frame shift in the analyzed gene. This sequence should have significant similarity with the amino acid sequence created by using the alternative reading frame. Very often such amino acid sequence may be absent in the data bank or may show no significant similarity. In this case the search of the reading frame shifts by using this method becomes impossible. Limitations of the methods of the second group are of a different nature. These limitations are related to the fact that the search of the reading frame shifts uses the approaches connected with the revelation of some general statistical properties of gene regions, which are already known to have the reading frame shifts. However, as it was shown previously (*15*), the statistical properties of gene sequences may be different, resulting that genes belong to different classes of triplet periodicity (TP). Integration of different genes with known reading frame shifts can lead to the fact that most statistical features of the integrated sequences become poorly expressed. This can significantly decrease the power of recognition of the reading frame shifts.

Recently, we reported an approach for revealing the potential reading frame shifts in genes by searching for the phase shift of TP (*16, 17*). The advantage is that it does not require any additional information for detection of the possible positions of reading frame shifts in the gene. Only the information of TP and its phase shift is needed to identify the reading frame shifts (*15-17*). The mathematical approach developed uses the statistical test to check the homogeneity of two polynomial samples with unknown distributions. The TP matrixes to the left and to the right of position x in the analyzed coding sequences can be considered as two polynomial samples (*15-17*). This problem is the standard so-called "change-point problem" (*18-20*) that was applied to the TP of DNA sequence.

TP of the coding DNA sequences is a common feature of all currently known living systems (*21-30*) and is associated with the reading frame that exists in a gene (*15*). The formation of TP is caused by the structure of the genetic code, which is practically the same in prokaryotes and eukaryotes, by the saturation of proteins with certain amino acids (*31-33*) and by GC content of the 3rd position of codons (*34*). If a shift of the reading frame occurs in a gene with TP, then this shift could be revealed due to the shift between existing reading frame and TP (**Figure 1**). Since it is difficult to significantly change the TP of coding sequences through relatively small number of base substitutions (*35*), such shift can remain in a gene for a rather long period of time. The presence of such shift between the TP of the nucleotide sequence and the existing reading frame may indicate the existing of a reading frame shift in the analyzed gene (*17*). However, the proposed mathematical method (*17*) can only detect the TP phase shift created by insertions of relatively short DNA sequences with lengths less than several tens of DNA bases. If insertion of a longer DNA fragment occurred, then this insertion can substantially change the TP around the area of the phase shift that greatly complicates the detection of the TP phase shift by using this method.

There are two goals for the present study. Firstly, we would like to develop a new mathematical method for revealing TP phase shifts to account for possible reading frame shifts occurring due to insertions of relatively large DNA fragments (>100 bases). Secondly, we wanted to verify the presence of relatively long insertions (with length not multiple of three) in bacterial genes by applying the advanced algorithm. Our results show that approximately 16% of bacterial genes from 17 studied genomes have the TP phase shifts that may be caused by insertions of relatively long DNA fragments in the genes.

	1	2	3
A	6	0	0
T	0	6	0
C	0	0	0
G	0	0	6

	1	2	3
A	0	6	0
T	0	0	6
C	1	0	0
G	6	0	0

	1	2	3
A	6	0	0
T	0	6	0
C	0	0	1
G	0	0	6

	1	2	3
A	0	0	6
T	6	0	0
C	0	1	0
G	0	6	0

Figure 1 The influence of one DNA base insertion on the TP phase shift. The first three sequences have the reading frames T_1, T_2 and T_3, respectively. Then the coding sequence S with TP is shown. In this sequence the insertion of the nucleotide c was made in the position 19. Explicit periodicity of this sequence is chosen for clarity. In a case of "fuzzy" periodicity, the situation is the same as in the figure, but the periodicity will be difficult to be observed visually. Then we construct the TP matrices $M_1(1, 18)$, $M_1(19, 37)$, $M_2(19, 37)$ and $M_3(19, 37)$. The first matrix M_1 is constructed for the DNA region from the 1st to 18th base. Elements of these matrices $m_1(i, j)$, $m_2(i, j)$ and $m_3(i, j)$ show the number of the bases a, t, c and g (index i) for the positions in the triplet reading frames T_1, T_2 and T_3 (index j). If we compare the matrix $M_1(1, 18)$ with the matrices $M_1(19, 37)$, $M_2(19, 37)$ and $M_3(19, 37)$, it can be seen that this matrix is most similar to the matrix $M_2(19, 37)$. The initial phase of the matrices M_1, M_2 and M_3 in sequence S is equal to 1, 2 and 3 because the bases of sequence S with indices k equal to 1, 2 and 3 are the first bases of the triplet in the reading frames T_1, T_2 and T_3. Therefore, there is a TP phase shift by 1 base in sequence S after the position x=18 (the difference between the initial phases of the matrices M_2 and M_1).

Method

Algorithm for searching the TP phase shift

The algorithm has been developed on the basis of our previous study (*17*). We assume that a coding nucleotide sequence $S = \{s(k), k = 1, 2, ..., L\}$ is given, where each base $s(k)$ is chosen from the alphabet $A = \{a, t, c, g\}$, L is the length of sequence S, which is a multiple of three. Let us introduce three reading frames in sequence S and denote them as T_1, T_2 and T_3 (Figure 1). The base $s(1)$ of sequence S is the first, second and third codon base of the reading frames T_1, T_2 and T_3, respectively. T_1 actually exists in sequence S while T_2 and T_3 can be considered as hypothetical ones. We also define three TP matrices as $M_1(i_1, i_2)$, $M_2(i_1, i_2)$ and $M_3(i_1, i_2)$, which are calculated for T_1, T_2 and T_3 for a part of sequence S from i_1 to i_2, denoted as $S(i_1, i_2)$. Morerover, $m_1(i, j)$, $m_2(i, j)$ and $m_3(i, j)$ are the elements of the matrices that show the number of bases of type i in sequence S (i=1, 2, 3, 4 for a, t, c, g, respectively) in the codon position j (j can be 1, 2 or 3) for T_1, T_2 and T_3, respectively. Let x_1 and x_2 are two coordinates in sequence S that are defined as L_1+3n, where n=0, 1, 2, 3, ..., (L−L_1)/3 and L_1 is a multiple of

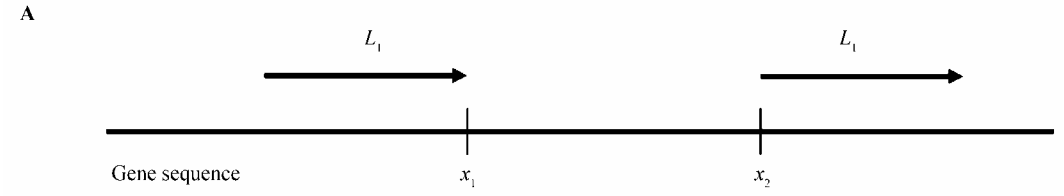
three being in the range from 60 to 600. Let us consider the fragment of the sequence $S(x_1-L_1+1, x_1)$ for which we construct the TP matrix $M_1(x_1-L_1+1, x_1)$ for T_1 of sequence S. Let us also consider the fragments $S(x_2+1, x_2+L_1)$, $S(x_2+2, x_2+L_1+1)$ and $S(x_2+3, x_2+L_1+2)$ for which we construct the TP matrices $M_1(x_2+1, x_2+L_1)$, $M_2(x_2+2, x_2+L_1+1)$ and $M_3(x_2+3, x_2+L_1+2)$ for T_1, T_2 and T_3, respectively, of sequence S. If an insertion of DNA fragment with the length of (x_2-x_1+1) or (x_2-x_1+2) DNA bases occurs right after the position x_1 in sequence S, then it creates a shift in the reading frame by one or two bases and the same shift of the TP phase. In this case the matrix $M_1(x_1-L_1+1, x_1)$ is more similar to the matrix $M_2(x_2+2, x_2+L_1+1)$ or $M_3(x_2+3, x_2+L_1+2)$, respectively. If, however, there are no insertions of nucleotides after the position x_1, then the matrix $M_1(x_1-L_1+1, x_1)$ is most similar to the matrix $M_1(x_2+1, x_2+L_1)$ for $x_1=x_2$. It is a typical problem of searching for the change point (*18-20*) in symbolical sequence. Then we added the matrix $M_1(x_1-L_1+1, x_1)$ to the matrix $M_k(x_2+k, x_2+L_1+k-1)$ to form the combined matrix M (k=1, 2, 3). The matrix M has 4 rows and 3 columns and can be considered as contingency table. The rows represent the DNA bases and the columns represent the positions of DNA bases in the corresponding reading frame. Example of filling the matrix M is shown in **Figure 2**. Then we calculated I_{1k} (the mutual information multiplied by $2L_1\ln 2$) for the combined matrix M using the following formula (*32*):

$$I_{1k} = \sum_{i=1}^{4}\sum_{j=1}^{3} m(i,j)\ln m(i,j) - \sum_{i=1}^{4} x(i)\ln x(i) - \sum_{j=1}^{3} y(j)\ln y(j) + 2L_1 \ln 2L_1 \quad (1)$$

where $x(i) = \sum_{j=1}^{3} m(i,j)$, $y(j) = \sum_{i=1}^{4} m(i,j)$.

Then we calculated the argument of the normal distribution as follows:

$$X_{1k} = \sqrt{4I_{1k}} - \sqrt{11} \quad (2)$$



B

```
123123123123123123123          123123123123123123123    T1
                                 123123123123123123123   T2
                                  123123123123123123123  T3
atcgatgatgatatcatcgatccatggtggtatcgatgatatcgatgatatca
```

$M_1(1,21)$

	1	2	3
a	3	4	0
t	0	3	4
c	0	0	3
g	4	0	0

$M_1(31,51)$

	1	2	3
a	0	3	4
t	5	0	3
c	2	0	0
g	0	4	0

$M_2(32,52)$

	1	2	3
a	3	4	0
t	0	3	4
c	0	0	3
g	4	0	0

$M_3(33,53)$

	1	2	3
a	4	0	3
t	3	4	0
c	0	3	0
g	0	0	4

Figure 2 The calculation of the matrix $M_1(x_1-L_1+1, x_1)$ and the matrices $M_k(x_2+k, x_2+L_1+k-1)$, k=1, 2, 3 in sequence S. **A**. The positions of regions with length L_1 in sequence S. **B**. The example of calculation of the matrix $M_1(x_1-L_1+1, x_1)$ and the matrix $M_k(x_2+k, x_2+L_1+k-1)$ for L_1=21, x_1=21, x_2=30. The insertion fragment begins from the 22nd base and ends by the 31st base of the DNA sequence. The insertion fragment shown in bold letters has another TP type than the rest DNA sequence. It is possible to see that matrix $M_1(1, 21)$ differs from matrix $M_1(31, 51)$ and matrix $M_3(33, 53)$ while similar to matrix $M_2(32, 52)$.

The value X_{1k} (k=1, 2, 3) is a measure that indicates the TP level in the combined matrix M=$M_1(x_1-L_1+1, x_1)$+$M_k(x_2+k, x_2+L_1+k-1)$. Sequence S (the nucleotide sequence of a gene) is not a random sequence since TP is observed in the gene. In this case we cannot use X_{1k} (k=1, 2, 3) as a measure of similarity between matrix $M_1(x_1-L_1+1, x_1)$ and matrices $M_k(x_2+k, x_2+L_1+k-1)$, k=1, 2, 3. Instead, the Monte Carlo method is used to calculate the similarity measure between two matrices $M_1(x_1-L_1+1, x_1)$ and $M_k(x_2+k, x_2+L_1+k-1)$. For this purpose, the sequences $S(x_1-L_1+1, x_1)$ and $S(x_2+1, x_2+L_1+3)$ are combined into one sequence $SS(1, 2L_1+2)$, which is shuffled with its TP retained. To achieve this, we divided the sequence $SS(1, 2L1+2)$ into three subsequences. The first of them (denoted as C_1) was obtained by choosing the bases in positions i=3n+1, where n=0, 1, 2, … from $SS(1, 2L1+2)$ sequence. The second and the third sequence C_2 and C_3 was obtained by choosing the bases in positions i=3n+2 and i=3n+3, respectively.

Next, the random sequences R_1, R_2 and R_3 were generated using a random number generator. They had the same length as the sequence C_1, C_2 and C_3, respectively. We arranged the sequences of R_1, R_2 and R_3 in ascending order and keep track of the permutation order for each sequence. Then we rearranged the bases in sequences C_1, C_2 and C_3 in the same way. Upon such shuffling of sequences C_1, C_2 and C_3, we created a random sequence C. In sequence C, positions i=3n+1, i=3n+2 and i=3n+3 were occupied by the bases of sequence C_1, C_2 and C_3, respectively. Sequence C had the equal length and the same base composition as sequence $SS(1, 2L1+2)$. We generated such random sequence C for 500 times. Each sequence C was divided back to the sequence $S(x_1-L_1+1, x_1)$ and $S(x_2+1, x_2+L_1)$, and for these two sequences we calculated X_{1k} using Formula 2. For the set of the values X_{1k}, the mean value and variance $D(X_{1k})$ were determined for k=2 and k=3. For this method of sequence SS shuffling, the values of X_{11} for SS were equal to the values of X_{11} for each of the random sequence C. As the measure of similarity between the matrices $M_1(x_1-L_1+1, x_1)$ and $M_k(x_2+k, x_2+L_1+k-1)$, we took the value:

$$Z_{1k} = \frac{X_{1k} - \bar{X}_{1k}}{\sqrt{D(X_{1k})}} \quad (3)$$

where k=2, 3.

It is possible to consider Z_{1k} as the function of L_1 for some constant values of x_1 and x_2. Then we calculated $Z_{1k}(L_1)$ for L_1=60, 90, 120, …, 600. If x_1-1 or $L-x_2$ were less than 600, we calculated $Z_{1k}(L_1)$ for L_1=60, 90, 120, …, min(x_1-1, $L-x_2$). We selected the value of L_1 that had the maximum value of Z_{1k}, which means that we selected the own value of L_1 for each x_1 and x_2. We need to find the maximum of Z_{1k} for some value L_1 since TP is not uniform along the length of a gene and may change its type (*15*). Testing of various lengths of L_1 has shown that the most effective search of the phase shifts was obtained if we did not fix any particular length L_1 and performed the search of some L_1 that had the maximal Z_{1k}. Carrying out the search of Z_{1k} maximal value for some L_1 does not interfere with the choice of a threshold Z_0 as described in the next subsection.

Maximal similarity between the matrices $M_1(x_1-L_1+1, x_1)$ and $M_k(x_2+k, x_2+L_1+k-1)$ (k=2, 3) corresponds to the maximal value of Z_{1k} for some k (k=2 or k=3 shows the insertion of 3n+1 or 3n+2 DNA fragment, correspondingly). The density distribution of Z_{1k} for different values of X_{11} is shown in **Figure 3**. It can be seen that for the different values of X_{11} we have similar distributions of Z_{12}. The same picture is observed for Z_{13}. These data show that the use of Monte Carlo method minimizes the TP influence of the sequence S on the spectrum Z_{1k}. It allows using Z_{1k} as the quantitative measure of the relationship between the matrix $M_1(x_1-L_1+1, x_1)$ and the matrices $M_k(x_2+k, x_2+L_1+k-1)$, k=2, 3. If similarity between the matrices is absent, then the values Z_{1k} (k=2, 3) are

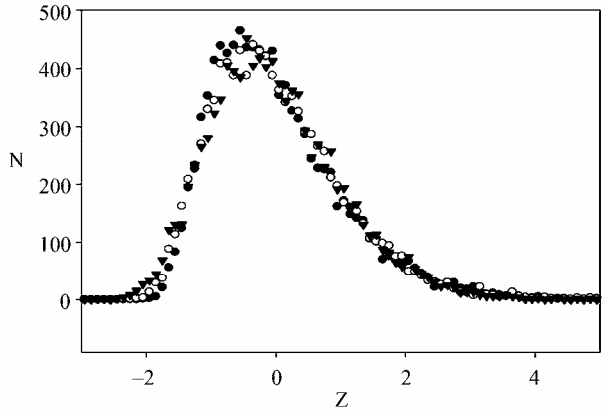


Figure 3 The density distribution of Z_{12} for different values of X_{11}. Symbols ● for X_{11}=0; ○ for X_{11}=6.0; ▼ for X_{11}=12.0.

small (typically less than 3.0), while in the presence of such similarity, Z_{1k} values will be large. We should only determine the threshold level Z_0 for Z_{1k}, k=2, 3. If $Z_0>Z_{1k}$, then it shows the absence of the statistically significant similarity between the matrices and the absence of TP phase shift between coordinates x_1 and x_2. If $Z_0<Z_{1k}$, then it shows the presence of the similarity between the matrices and the existence of TP phase shift between coordinates x_1 and x_2. Selection of Z_0 is discussed below in the next subsection.

We changed x_1 from L_1 to $L-L_1+1$ and x_2 was changed from x_1 to $L-L_1+1$. Then for each value of x_1 in sequence S we calculated value of x_2, which gives the maximum value of Z_{1k} (k=2, 3). Let such maximum be referred as mZ_{1k}. Then the plots for mZ_{1k} depending on x_1 and x_2 for k=2, 3 were obtained. In these plots we joined the neighboring points by line for more clearness. Let us assume that we have the sequence S with insertion of a DNA fragment with the length not multiple of three bases between positions x_1^0 and x_2^0. Then the maximum value for the dependence of mZ_{1k} on x_1 occurs at position x_1^0, and the maximum value for mZ_{1k} depending on x_2 will be observed near position x_2^0 (for the appropriate k). It means that the points of x_2 tend to group near and after x_2^0. If the insertion has the length equal to x_2-x_1+1, then the largest values of mZ_{1k} are observed for k=1, and if the insertion had the length x_2-x_1+2, then the largest values of mZ_{1k} are obtained for k=2. The graph of the $mZ_{1k}(x_1)$ function is a "mountain", and the apex of this graph is observed for $x_1=x_1^0$. The graph of the function of $mZ_{1k}(x_2)$ looks like a "wall" and the boundary of the wall is position x_2^0 (examples of such plots are shown below in Results). Such graphs allow to predict the approximate (with accuracy up to several tens of bases) positions x_1^0 and x_2^0 in the gene.

Application of Monte Carlo method to determine Z_0

To find the threshold value Z_0, we used the gene sequences from 17 bacterial genomes (**Table 1**) from KEGG database (*36*). We created the random data bank by shuffling the bases of each gene sequence. It

Table 1 List of the prokaryotic genomes used for searching the genes with triplet periodicity shifts

No.	Genome	No. of analyzed genes (>1,200 bp)	Q1	Q2	Q3
1	*Arcobacter butzleri*	611	5	30	3
2	*Azotobacter vinelandii*	1,306	43	140	29
3	*Bordetella avium*	885	42	108	34
4	*Burkholderia mallei*	1,380	116	240	103
5	*Bacillus subtilis*	937	50	145	35
6	*Escherichia coli*	1,158	101	237	75
7	*Lactobacillus fermentum*	444	16	49	14
8	*Methylococcus capsulatus*	854	41	111	38
9	*Pseudomonas aeruginosa*	1,566	38	162	29
10	*Staphylococcus aureus COL*	626	28	91	24
11	*Salmonella enterica* Choleraesuis	1,187	109	227	86
12	*Streptococcus pneumoniae*	507	27	82	25
13	*Shigella sonnei*	1,183	175	286	141
14	*Salmonella typhimurium*	1,200	94	220	68
15	*Vibrio cholerae*	1,047	61	176	41
16	*Xanthomonas campestris*	1,245	85	253	62
17	*Yersinia pseudotuberculosis YPIII*	1,084	119	252	98
	Total	17,220	1,150	2,809	905

Note: Q1, Q2 and Q3 are the number of the genes with a length greater than 1,200 bp that have a TP phase shift revealed by the method developed previously (*17*), the method developed in the present work, and both the method developed previously (*17*) and the method developed in the present work, respectively.

allows keeping the same length distribution for random sequences as for the studied genes from 17 bacterial genomes. To keep TP in the random sequence, the shuffling was performed in the same manner as described above. Upon shuffling of the sequences, only the TP phase shifts caused by random factors are remained. As a result, the database of random sequences was created. Sequences from this data bank had the same length and TP as in the genes of 17 bacterial genomes studied. We chose some level of Z_0 (for example, Z_0=4.0) and calculated the number of genes that had at least one value of $mZ_{1k}>Z_0$ for k=2 or 3 (as described above). This calculation was performed for the gene sequences from 17 bacterial genomes and for random sequences from the created data bank (numbers N_1 and N_2, respectively). We calculated N_2/N_1 and were increasing Z_0 until N_1/N_2 was not equal to 0.22. We did it till the level Z_0=8.0 when the number of the found TP phase shifts in random sequences was about 22% from the number of shifts that we have revealed in 17 bacterial genomes (N_2/N_1=0.22). Therefore, the level Z_0=8.0 can be chosen as the threshold level because an admixture of the TP shifts due to purely random factors can be considered as being relatively small.

We chose the ratio N_2/N_1=0.22 for two reasons. The first reason is that we wanted to find the upper limit of the number of genes with a phase shift of TP. The second reason is that we were going to compare the results obtained in this paper with the results obtained previously (*17*). The algorithm described above allows calculating Z_0 for any ratio N_2/N_1 and for any number of testing genes. The level of Z_0 depends on the number of the analyzed genes.

Results

Analysis of genes from 17 bacterial genomes

At first, we studied the TP phase shift in the artificial periodic sequence. To create the artificial sequence, we took the sequence of the transaldolase B gene from the genome of *Escherichia coli* (b0008 in KEGG database) and added a random sequence of 298 bases after the 300th base, which shared the same base frequencies as transaldolase B gene and had a TP level greater than 5.0 (Formula 2). The insertion of this fragment creates the TP phase shift (and the shift of the reading frame) after the 598th base of the gene. The sequence of b0008 was initially analyzed without insertion and it was shown that $mZ_{12}(x_1)$ (**Figure 4A**) and $mZ_{12}(x_2)$ (**Figure 4B**) did not contain the values of mZ_{12} higher than 5.0, which was even lower than the selected threshold value of 8.0. This result suggests that the analyzed sequence contains a homogeneous TP and the matrix $M_1(x_1-L_1+1, x_1)$ is always more similar to the matrix $M_1(x_2+1, x_2+L_1)$ than to the matrices $M_k(x_2+k, x_2+L_1+k-1)$, k=2, 3, indicating the absence of TP phase shifts in the gene.

A completely different pattern is observed in this gene upon artificial insertion (insertion of 298 nucleotides after the 300th base) of a random sequence, which creates the TP phase shift by one base to the right after the 598th base. The graph of the function $mZ_{12}(x_1)$ (**Figure 5A**) looks like a "mountain" and shows that while x_1 moves from the 1st to the 300th

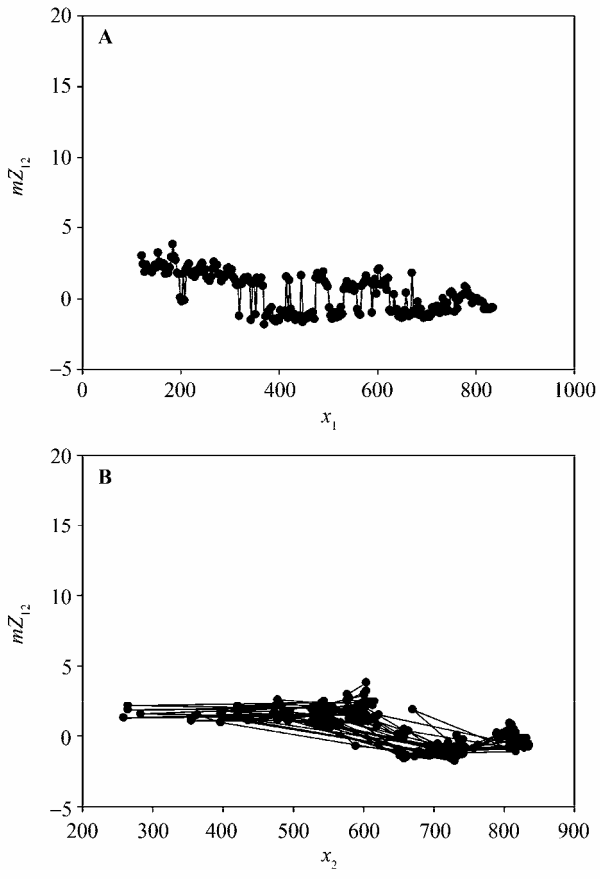


Figure 4 Dependence of mZ_{12} on x_1 (**A**) and x_2 (**B**) for the gene encoding the transaldolase B from the genome of *E. coli* (b0008 in KEGG database).

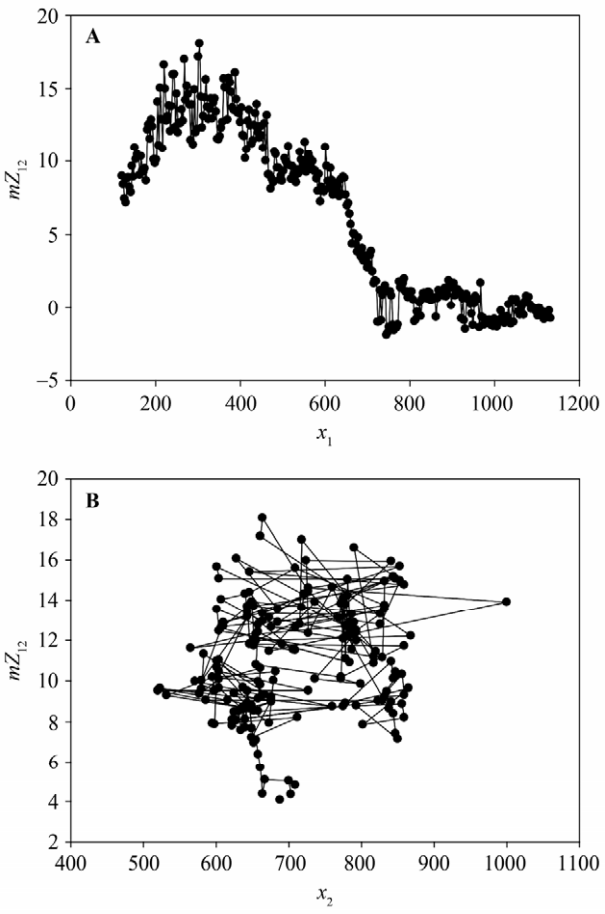


Figure 5 Dependence of mZ_{12} on x_1 (**A**) and x_2 (**B**) for the gene encoding transaldolase B from the genome of *E. coli* (b0008 in KEGG database) with insertion of 298 nucleotides after the 300th nucleotide.

base, the value of mZ_{12} increases and reaches its maximum in the 300th base. Then the value of mZ_{12} is decreased due to combination of different triplet periodicities, which are observed in this sequence after the 300th base. The graph of the dependence of mZ_{12} on x_2 (**Figure 5B**) looks like a "wall" and the values of x_2 group together after the 600th base. This graph shows that all maximal similarities of the matrix $M_1(x_1-L_1+1, x_1)$ and the matrix $M_2(x_2+2, x_2+L_1+1)$ are observed mainly for $x_2 \geq 600$. Thus, this artificial example shows that $mZ_{12}(x_1)$ and $mZ_{12}(x_2)$ graphs allow to see the TP phase shift and to predict the boundaries of the insertions x_1^0 and x_2^0 with accuracy up to several tens of DNA bases.

Then the genes with a length more than 1,200 bp from 17 bacterial genomes were analyzed by the method developed. A list of the genomes is shown in Table 1. The reason for choosing such length is explained as follows. For the correct statistical estimation, each cell of the matrices $M_1(x_1-L_1+1, x_1)$ and $M_k(x_2+k, x_2+L_1+k-1)$ should contain, on the average, no less than 10 values (*37*). We have 12 cells in each of $M_1(x_1-L_1+1, x_1)$ and $M_k(x_2+k, x_2+L_1+k-1)$ matrices, and it gives the lower limit of L_1=120. Therefore, to detect the TP phase shift, at least 240 bases are required. Accordingly, we used the genes for this study with a length of more than 1,200 nucleotides in order to reduce the influence of boundary effects on the performance of the algorithm. The total number of genes selected in the studied genomes was 17,220. There were 1,526 genes identified with insertions of the type $3n+1$ and 1,283 genes with insertions of the type $3n+2$ (n=0, 1, 2, ...), for which $mZ_{1k}>Z_0$, *i.e.*, the total number of the genes with the insertions was 2,809 (Table S1). This amount constitutes 16.3% of the studied genes with a length more than or equal to 1,200 bases. Concurrently, the random sequences from the created data bank were analyzed and 622 sequences with $mZ_{1k}>Z_0$ were found (see Method). Our analysis shows that the number of false positives in this case does not exceed 22%.

Examples of x_1 and x_2 determination

Let us consider several examples of the genes with revealed insertions for which $mZ_{1k}>Z_0$. For all examples only the values $mZ_{1k}>4.0$ are shown to reduce the influence of statistical noise on the plots. The first example is shown in Figure S1 for the gene coding the glycosyl transferase from the genome of *E. coli*. Figure S1A shows that $x_1^0 \approx 760$ and Figure S1B shows that $x_2^0 \approx 760$. According to our method, these x_1^0 and x_2^0 values indicate that this gene has a short insertion with a length equal to $3n+2$ (n=0, 1, …, 10) (function mZ_{12}) or a deletion with a length equal to $3n+1$ (n=0, 1, 2, ...). This example suggests that the mathematical method developed can detect the TP phase shifts caused by relatively short insertions or deletions. A second example of the gene with insertion having length not multiple of three bases is shown in Figure S2 for the gene coding the molybdenum cofactor biosynthesis protein A from the genome of *Burkholderia mallei*. Figure S2A shows

that $x_1^0 \approx 950$, and Figure S2B shows that $x_2^0 \approx 1100$, indicating that this method can detect the insertions that have a length about 100 nucleotides ($3n+1$, $n \approx 33$). The third example (Figure S3) is the gene encoding the ubiquinone oxidoreductase, chain G, from the genome of *E. coli* (B2283 in database KEGG), where $x_1^0 \approx 850$ and $x_2^0 \approx 1100$, *i.e.*, the insertion size is about 250 bases. This gene has the insertion with the length equals to $3n+2$, $n \approx 83$.

From these examples, it is possible to see that the accuracy of the boundaries x_1^0 and x_2^0 is no better than ±60 bp, as noted in Method.

Searching for amino acid similarity

We have also studied the similarity of the amino acid sequences created after the position x_2 till the end of the gene using the BLAST program. The amino acid sequences were created for the current reading frame in the gene and for the hypothetical reading frame, which could exist in the gene after position x_2 until the moment of a fragment insertion between positions x_1 and x_2. Thus, after x_2 position we have a pair of amino acid sequences, one of them actually exists and the other is hypothetical. For the real and hypothetical sequences, a sequence with the highest similarity was searched in the Swiss-prot database. For 803 pairs of such sequences the significant similarities in the Swiss-prot database did not exist. For 1,918 pairs of sequences the similarity was found only for the actually existing sequences, but for the hypothetical sequences the similarity was absent. For 84 pairs of sequences the similarity was found only for the hypothetical sequences. Only for 4 pairs such a similarity was observed for both the actually existing sequence and for the hypothetical sequence. These results show that the search of the insertions with a length not multiple of three bases in genes by means of revealing the TP phase shift in some cases can be confirmed by the found similarities. This result is not surprising since the insertion of DNA fragment into the gene could take place long time ago, and currently it is hard to notice the similarity at statistically significant level. Since TP is changing slowly (*35*), it allows revealing the TP phase shift that could take place long time ago.

Here we show an example in which the similarity was found for the existing and hypothetical amino acid sequences (existing and hypothetical reading frames in gene mba1516 from KEGG database, and Q62J71_BURMA amino acid sequence for existing reading frame from Swiss-prot database). This gene codes the molybdenum cofactor biosynthesis protein A from the genome of *B. mallei*. As it was noted above, this gene has the insertion with coordinates $x_1^0 \approx 870$ and $x_2^0 \approx 1100$ (Figure S2). The existing amino acid sequence has only one similarity after $x_2^0 \approx 1100$ with the amino acid sequence A4LD09_BURPS, which also encodes molybdenum cofactor biosynthesis protein A, but in the genome of pseudo *B. mallei* 305. This similarity is observed for almost 100%. However, if we study the similarity of the whole amino acid sequence Q62J71_BURMA with the sequences from Swiss-prot, it can be seen that all other similarities of this sequence are finished near the 329th amino acid. It roughly corresponds to the beginning of the insertion in the gene ($x_1^0 = 870$). An example of such similarity is shown in Figure S4A for the sequence Q2SA06_HAHCH. We assume that after the insertion of a DNA fragment, a gene similar to the gene BMA1516 was cut into two fragments F_1 and F_2 near the base 870. The fragment F_1 (from 1st to 870th bases) was added to some sequences as the beginning of the gene. Thus, multiple similarities of the region from the 1st to 320th amino acids to amino acid sequences of different proteins were arisen.

For the hypothetical sequence, the similarity search showed that the fragment from the 320th amino acid to the end of the amino acid sequence had many similarities with different proteins, only from the 1st to 260th amino acid of these proteins. The example of such similarity is shown in Figure S4B. It can be assumed that the fragment F_2 was attached to some other DNA fragment as the beginning of the gene, and the second reading frame became the coding reading frame in this fragment.

It is also possible that the gene BMA1516 was created by a fusion of three fragments. The first fragment (called E_1) is similar to the gene coding the amino acid sequence Q2SA06_HAHCH. Then it was joined to the second relatively short fragment (referred as E_2) having the length approximately equal

to $3n+2$, $n\approx83$, after which the reading frame was changed. Then the fragment E_3, which is similar to the gene coding the sequence Q63SW3_BURPS, was attached to the end of the fragment E_2. Since E_3 has a TP type similar to E_1, it is possible to reveal the TP phase shift after joining E_3. However, in the case of validity of any hypothesis from the two hypotheses considered here, the fragment of F_2 ($E_2+E_3=F_2$) or its most part E_3 may code the functionally important protein in two different reading frames.

Discussion

Deletions of DNA fragments with the length not multiple of three bases are found by this method as the insertions of one or two DNA bases. Therefore, we have developed an approach that reveals the whole set of TP phase shifts caused by deletions and insertions of DNA fragments. In this study 2,809 such genes were discovered, in which there were two regions with the same type of TP separated by insertions of nucleotides. This number constitutes approximately 16.3% from the total number of the analyzed genes while ~4% of the genes have the deletions and short insertions (*17*). Therefore, it can be assumed that the frequency of insertions of long DNA fragments is approximately few times greater than the frequency of deletions and short insertions.

In the revealed genes, the reading frame and the TP were clearly linked initially, and only after the insertions of DNA fragments the shift between them was formed. This relatively large percentage of genes with the TP phase shift may suggest that the shift of the reading frame in gene is not a very dramatic event for the encoded protein. It also means that the genetic code must somehow be adapted to these events (*33-35, 38*). If a large percentage of the genes with the shifts of the reading frame is observed, then a new amino acid sequence should often have some biological functions that can be picked up by the evolution. This may explain the relatively large percentage of genes with TP phase shift.

It is unlikely that the observed TP phase shifts are related to the sequencing errors. The mostly well-studied genomes of the bacteria, which up to date have been sequenced more than once, were chosen for this work. In this case the probability of the sequencing errors in a form of deletion or insertion of one or two DNA bases is significantly reduced, whereas replacements of DNA bases did not create the triplet phase shifts. However, the disappearance of the start or stop codons because of the sequencing errors could lead to errors in the gene identification. In this case, the accession of additional non-coding DNA fragments to actually existing gene may occur, which has relatively little effect on the TP phase shifts. This means that the connection of different types of TP may occur, but the TP phase shifts are absent. Furthermore, the insertion of long DNA fragments can hardly be induced by the sequencing errors, since the sequencing errors create deletions or insertions of only small DNA fragments (usually one or two bases).

In the present study we found a lower bound of the number of genes that contain a shift between the reading frame and TP. In reality the number of these genes may be large, since the approach works well with small numbers of insertions or deletions. If the density of the insertions and deletions is more than one insertion or deletion per a few tens of bases (~60), then discovery of the deletions and insertions by this algorithm is not always possible. As a result the statistically significant value mZ_{1k} for this gene cannot be obtained.

The mathematical approach used in this paper is the expansion of the method that was used earlier to reveal the TP phase shift (*17*). The modifications are as follows. Firstly, two triplet matrices (to the left and to the right from the position x, see Method) were compared using the level of similarity rather than using the level of a difference. It is more accurate since it allows to ignore such position x, in which the matrix $M_1(x_1-L_1+1, x_1)$ is not similar to any of the matrices $M_k(x_2+k, x_2+L_1+k-1)$, $k=1, 2, 3$. This situation may arise due to the splicing of gene fragments (*39*) and in this case the difference between the matrices $M_1(x_1-L_1+1, x_1)$ and $M_2(x_2+1, x_2+L_1)$ may be greater than the difference between the matrices $M_1(x_1-L_1+1, x_1)$ and $M_k(x_2+k, x_2+L_1+k-1)$, $k=2, 3$ due to the existence of certain classes of TP (*15*). Accordingly the splicing of genes could be identified as the TP phase shift. The use of the similarity for the matrix comparison allows

eliminating the possibility of identifying splicing of the genes with different triplet frequencies as TP phase shift. Secondly, in this paper we have developed an approach for revealing insertions of long DNA fragments with lengths not multiple of three DNA bases. It is impossible to find the TP phase shift upon insertion of long DNA fragments (>100 bp) in genes by the method proposed earlier (*17*). The method proposed in the present work allows revealing TP phase shifts after insertion of DNA fragment of any length that is not multiple of three bases. We compared the results of this study with those obtained previously (*17*) for the genes with a length greater than 1,200 bp from bacterial genomes. Comparison was performed for the same level of the false positives number (~22%, see Method). From these results, it is clear that the examination of large DNA insertions allows to reveal about 2.4 times more genes (columns Q1 and Q2) with a TP phase shift than it was revealed previously (*17*). In addition, it can identify ~80% of genes with a phase shift (columns Q1 and Q3), which were identified earlier. The remaining 20% of genes have no TP phase shift, but rather the splicing of gene fragments. This error can occur in the algorithm developed earlier (*17*) because it uses the measure of matrices difference, as noted above.

The computational complexity of the analysis of a sequence S using our method is $O(L^2)$. About 20 hours were required for the analysis of 17,220 genes with the length more than 1,200 bp. We used 5 AMD Phenom II X4 processors for calculations. This result shows that computer cluster with 100 processors or more is required for analysis of all known genes from KEGG data bank (~18×10^6 genes). In this case the time of the calculation could be greater than some months.

Search of the TP phase shifts with help of Fourier transformation was also reported in the previous study (*40*), which shows that the method is able to reveal the artificial insertions or deletions of bases in the genes. However, there is no data for the detection of the real TP phase shifts in genes from *E. coli* genome. Also, a large window was used in this study (750 bp and more) that can severely complicate the detection of the TP phase shifts in genes.

Spectral rotation measure (SRM) has been used to search for the TP in coding sequences (*41, 42*). It can also be applied to search the TP phase shifts (*41*). However, the reading frame identification was performed for a fixed window size equal to 351 bp (*41*). The use of fixed windows may lead to omission of the existing TP phase shifts, which can be found with help of larger or smaller windows (the value $2L_1$ in our work). Our approach uses the TP matrices, which determine a TP type before x_1 position and after x_2 position in the sequence S very accurately. Using of the matrices allows to find the TP phase shifts for all values of L_1 (L_1 is not fixed in our method). It will be possible to compare our method and SRM more accurately when SRM is applied for the search of the TP shifts in genes from the bacterial genomes.

The results obtained show that unexpectedly large number of genes (~16%) have a TP phase shift (change points of TP in gene sequences). Shift of the reading frame is likely to be relatively neutral mutation, which does not result in complete inactivation of the gene.

Improvement of the mathematical approach used in this paper may be implemented with help of using more advanced algorithms that were applied for searching of the change points (*18-20*). In this case it will be possible to detect the shifts of the reading frame caused by multiple insertions and deletions of DNA bases in different regions of a single gene.

Conclusion

A mathematical method has been developed for searching the TP phase shifts in genes. The method is based on a comparison of the TP matrices. Using this method, we analyzed the genes that are longer than 1,200 bp from 17 bacterial genomes. It was found that about 16% genes have the TP phase shifts. We propose that these phase shifts indicate the presence of insertions and deletions of DNA fragments in genes.

Authors' contributions

MAK prepared the software and conducted data analysis. NAK did valuable discussion and co-wrote the manuscript. EVK proposed the approach for

searching insertions in genes, conducted data analysis and co-wrote the mauscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Wei, Q., *et al*. 2007. *DNA Repair, Genetic Instability, and Cancer.* World Scientific Publishing, Singapore.

2 Watson, J.D., *et al*. 2004. *Molecular Biology of the Gene.* Benjamin-Cummings Publishing, San Francisco, USA.

3 Okamura, K., *et al*. 2006. Frequent appearance of novel protein-coding sequences by frameshift translation. *Genomics* 88: 690-697.

4 Raes, J. and Van de Peer, Y. 2005. Functional divergence of proteins through frameshift mutations. *Trends Genet.* 21: 428-431.

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 Evol. Biol.* 6: 30.

6 States, D.J. and Botstein, D. 1991. Molecular sequence accuracy and the analysis of protein coding regions. *Proc. Natl. Acad. Sci. USA* 88: 5518-5522.

7 Pearson, W.R., *et al*. 1997. Comparison of DNA sequences with protein sequences. *Genomics* 46: 24-36.

8 Birney, E., *et al*. 1996. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.* 24: 2730-2739.

9 Guan, X. and Uberbacher, E.C. 1996. Alignments of DNA and protein sequences containing frameshift errors. *Comput. Appl. Biosci.* 12: 31-40.

10 Antonov, I. and Borodovsky, M. 2010. Genetack: frameshift identification in protein-coding sequences by the Viterbi algorithm. *J. Bioinform. Comput. Biol.* 8: 535-551.

11 Kislyuk, A., *et al.* 2009. Frameshift detection in prokaryotic genomic sequences. *Int. J. Bioinform. Res. Appl.* 5: 458-477.

12 Fichant, G.A. and Quentin, Y. 1995. A frameshift error detection algorithm for DNA sequencing projects. *Nucleic Acids Res.* 23: 2900-2908.

13 Médigue, C., *et al.* 1999. Detecting and analyzing DNA sequencing errors: toward a higher quality of the *Bacillus subtilis* genome sequence. *Genome Res.* 9: 1116-1127.

14 Schiex, T., *et al.* 2003. FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Res.* 31: 3738-3741.

15 Frenkel, F.E. and Korotkov, E.V. 2008. Classification analysis of triplet periodicity in protein-coding regions of genes. *Gene* 421: 52-60.

16 Frenkel, F.E. and Korotkov, E.V. 2009. Using triplet periodicity of nucleotide sequences for finding potential reading frame shifts in genes. *DNA Res.* 16: 105-114.

17 Korotkov, E.V. and Korotkova M.A. 2010. Study of the triplet periodicity phase shifts in genes. *J. Integr. Bioinform.* 7: 131.

18 Carlstein, E., *et al* (eds). 1994. *Change-Point Problems. IMS Lecture Notes–Monograph Series*, Vol. 23. Institute of Mathematical Statistics, Hayward, USA.

19 Litton, C.D. 1998. *Statistical Analysis of Change-Point Problems. Wiley Series in Probability & Mathematical Statistics: Applied Probability & Statistics*. John Wiley & Sons, New York, USA.

20 Sinha, B. and Rukhin, A. 1995. *Applied Change Point Problems in Statistics*. Nova Science Publishers, Hauppauge, USA.

21 Fickett, J.W. 1998. Predictive methods using nucleotide sequences. *Methods Biochem. Anal.* 39: 231-245.

22 Staden, R. 1994. Staden: statistical and structural analysis of nucleotide sequences. *Methods Mol. Biol.* 25: 69-77.

23 Baxevanis, A.D. 2001. Predictive methods using DNA sequences. *Methods Biochem. Anal.* 43: 233-252.

24 Gutiérrez, G., *et al.* 1994. On the origin of the periodicity of three in protein coding DNA sequences. *J. Theor. Biol.* 167: 413-414.

25 Gao, J., *et al.* 2005. Protein coding sequence identification by simultaneously characterizing the periodic and random features of DNA sequences. *J. Biomed. Biotechnol.* 2: 139-146.

26 Yin, C. and Yau, S.S. 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *J. Theor. Biol.* 247: 687-694.

27 Eskesen, S.T., *et al*. 2004. Periodicity of DNA in exons. *BMC Mol. Biol.* 5: 12.

28 Bibb, M.J., *et al*. 1984. The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene* 30: 157-166.

29 Konopka, A.K. 1994. Sequences and codes: fundamentals of biomolecular cryptography. In *Biocomputing: Informatics and Genome Projects* (ed. Smith, D.W.), pp.119-174. Academic Press, San Diego, USA.

30 Trifonov, E.N. 1999. Elucidating sequence codes: three codes for evolution. *Ann. N. Y. Acad. Sci.* 870: 330-338.

31 Eigen, M. and Winkler-Oswatitsch, R. 1981. Transfer-RNA: the early adaptor. *Naturwissenschaften* 68: 217-228.

32 Zoltowski, M. 2007. Is DNA code periodicity only due to CUF-codons usage frequency? *Conf. Proc. IEEE Eng.*

Med. Biol. Soc. 2007: 1383-1386.

33 Antezana, M.A. and Kreitman, M. 1999. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. *J. Mol. Evol*. 49: 36-43.

34 Aota, S. and Ikemura, T. 1986. Diversity in G+C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res.* 14: 6345-6355.

35 Korotkov, E.V., *et al*. 2003. The informational concept of searching for periodicity in symbol sequences. *Mol. Biol.* (Mosk) 37: 436-451.

36 Ogata, H., *et al*. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 27: 29-34.

37 Gmurman, V.E. 1968. *Fundamentals of Probability Theory and Mathematical Statistics*. American Elsevier Publishing, New York, USA.

38 Kullback, S. 1959. *Information Theory and Statistics*. John Wiley & Sons, New York, USA.

39 Pasek, S., *et al*. 2006. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* 22: 1418-1423.

40 Masoom, H., *et al*. 2006. A fast algorithm for detecting frame shifts in DNA sequences. In *Proceedings of IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, pp.1-8. Toronto, Canada.

41 Kotlar, D. and Lavner, Y. 2003. Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. *Genome Res.* 13: 1930-1937.

42 Chen, B. and Ji, P. Visualization of the protein-coding regions with a self adaptive spectral rotation approach. *Nucleic Acids Res.* 39: e3.

Supplementary Material

Figures S1-S4; Table S1

DOI: 10.1016/S1672-0229(11)60019-3