DNA Research **16**, 105–114, (2009) doi:10.1093/dnares/dsp002

Using Triplet Periodicity of Nucleotide Sequences for Finding Potential Reading Frame Shifts in Genes

F.E. Frenkel*, and E.V. Korotkov

Bioengineering Centre of RAS, 60-letiya Oktyabrya prosp., 7/1, Moscow 117312, Russia

(Received 26 August 2008; accepted 5 February 2009; published online 3 March 2009)

Abstract

We introduce a novel approach for the detection of possible mutations leading to a reading frame (RF) shift in a gene. Deletions and insertions of DNA coding regions are considerable events for genes because an RF shift results in modifications of the extensive region of amino acid sequence coded by a gene. The suggested method is based on the phenomenon of triplet periodicity (TP) in coding regions of genes and its relative resistance to substitutions in DNA sequence. We attempted to extend 326 933 regions of continuous TP found in genes from the KEGG databank by considering possible insertions and deletions. We revealed totally 824 genes where such extension was possible and statistically significant. Then we generated amino acid sequences according to active (KEGG's) and hypothetically ancient RFs in order to find confirmation of a shift at a protein level. Consequently, 64 sequences have protein similarities only for ancient RF, 176 only for active RF, 3 for both and 581 have no protein similarity at all. We aimed to have revealed lower bound for the number of genes in which a shift between RF and TP is possible. Further ways to increase the number of revealed RF shifts are discussed.

Keywords: triplet periodicity; reading frame; shift

1. Introduction

Mutations in DNA sequences appear as a result of base substitution or deletions, insertions and DNA sequence inversion.[1] At the level of amino acid sequences, substitutions of DNA bases may lead to substitutions of amino acids, i.e. single nucleotide substitution changes one amino acid or none. In this sense, deletions, insertions and inversions of DNA coding regions are more dramatic events for genes because such mutational alteration may lead to a reading frame (RF) shift and to modification of the extensive region of amino acid sequence. Substitution frequencies are well investigated, whereas RF shifts in genes are less studied, which is due to the difficulty in their identification. However, the investigation of the influence of RF shifts in genes on protein structure is of great interest. A single DNA base substitution may result in only one change of amino acid in a protein, whereas an RF shift causes the change of all amino acids coded downstream of the shift point in a gene. If after this the protein does not lose its function, then it is supposed that a frame shift has occurred in a functionally insignificant site or that the RF shift generates functionally similar amino acid sequences. But if after RF shift the protein has changed its function, then it is of great interest to understand how changes in amino acid sequence determine the protein's new function.

Answers for these questions can probably be found after a detailed investigation of RF shift statistics. To that end, we need a method for finding RF shifts in existing genes. Currently, general methods

Edited by Hiroyuki Toh

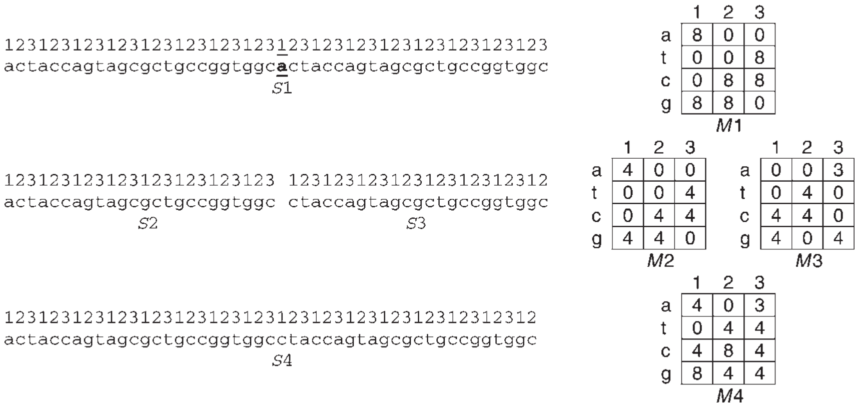
* To whom correspondence should be addressed. Tel. +7 499-135-2161. Fax. +7 499-135-0571. E-mail: felix.frenkel@gmail.com

for finding RF shifts and inversions comprise the search for similarities between amino acid sequences through the BLAST program or its analogues.[2,3] When using such procedures for finding RF shifts, we should in some way mark out the gene region where the RF shift is expected to be found. Then we recode this region of nucleotide sequence into that of the amino acid according to the new RF, thus obtaining the hypothetic amino acid sequence as a result. Thereafter, we run a similarity search for a hypothetic amino acid sequence in UniProtKB/Swiss-Prot databank. If statistically significant similarities are found, then we can be sure that the gene contains an RF shift, i.e. we see that sequences related to the hypothetic a amino acid sequence do exist. This method has allowed us to find, so far, several hundred genes in which the RF shift has really taken place.[2,3]

However, this scheme of finding RF shifts and inversions has some limitations. First, we have to choose a gene, using some criterion, where an RF shift is supposed to exist, and after that, find the probable site of RF shift and inversion in it. A full-scale search for RF shifts in all genes would require very powerful computational facilities that are not always available, although modern computer systems have made such search possible.[2] Second, even if we solve the first task, then the UniProtKB/Swiss-Prot databank will be necessary to contain the amino acid sequence having a statistically significant similarity to the hypothetical amino acid sequence. But it is possible that such a sequence will not be found due to the limitation of the UniProtKB/Swiss-Prot databank and because of too great an evolutional dissimilarity accumulated between amino acid sequences. Therefore, the concerned approach can reveal only some RF shifts accumulated to date in existing genes.

In order to reveal RF shifts and inversions in genes in a more reliable way, some new approach to seeking RF shifts should be developed instead of searching for similarities between hypothetic and real amino acid sequences. As shown in this work, a search for uniform triplet periodicity (TP) within the nucleotide sequence of gene's coding regions may be used as such an approach. Triplet organization of protein coding DNA sequences is a general feature of all presently known live systems.[4–12] The reason for this lies not only in the structure of genetic codes, which is virtually the same either for prokaryotes or for eukaryotes, but also in the saturation of proteins by certain amino acids.[13–16] If an RF shift occurs in genes in the presence of TP, then it will be revealed because a shift between TP and RF will also occur (Fig. 1). Since TP in a DNA sequence is unlikely to be changed by a small number of base substitutions,[17] then such shift will exist for a long period of time. The presence of such a shift between the TP of a nucleotide sequence and RF may serve as an indication of an RF shift in the concerned gene.

Currently, some methods have been developed that reveal TP by using regularity in symbol preferences over different triplet positions in the DNA sequence. They use Fourier transformation, hidden Markov chains and other statistical methods based on position-dependent preferences for nucleotides in coding sequences as a mathematical apparatus.[18–23] The methods are aimed at revealing DNA coding



	1	2	3
a	8	0	0
t	0	0	8
c	0	8	8
g	8	8	0

M1

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

M2

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

M3

	1	2	3
a	4	0	3
t	0	4	4
c	4	8	4
g	8	4	4

M4

Figure 1. Influence of one nucleotide deletion on TP of nucleotide sequence. Numbers above sequence *S*1 show positions of nucleotides in RF. Twenty-fifth base has been deleted from sequence *S*1. Consequently, sequence *S*1 can be presented as two sequences—*S*2 and *S*3, i.e. $S1 = S2aS3$. In *S*2 and *S*3 sequences, TP (matrices *M*2 and *M*3) will be the same, but in sequence *S*3, there will be cyclical shifts by one base relative to sequence *S*2 and to RF in sequence *S*1. This means that first column of matrix *M*2 corresponds to third column of matrix *M*3, second column of matrix *M*2 corresponds to first column of matrix *M*3, third column of matrix *M*2 corresponds to second column of matrix *M*3. After summing sequences *S*2 and *S*3 and formation of sequence *S*4, TPM (*M*4 matrix) comes out from addition of first column of matrix *M*2 to first column of matrix *M*3 and so on, that leads to merging of non-identical columns and considerably decreases statistical significance of TP in sequence *S*4. By accounting deletions, we get sequence *S*4 that has TPM (*M*4 matrix) in which matrices *M*2 and *M*3 are merged in consideration of cyclic permutation. Finding and accounting deletions considerably increases statistical significance of TP in sequence *S*4.

sequences and their separation from non-coding regions. A later, method of information decomposition to find TP [17,24] allows the introduction of the term of a TP class (TPC) as a 3×4 matrix. In this matrix, columns represent period positions and rows represent nucleotides.

In the current work, two problems were set. First, we wanted to find all genes where RF shifts can be identified by using TP. For each gene from the KEGG-29 databank, analysed[25] we extracted a region with TP having maximal statistical significance calculated by information decomposition without allowing any deletions or insertions of nucleotides.[17,24] Then we built a corresponding matrix of TP which was linked to the existing RF of the given DNA region. This meant that the first column of the TP matrix (TPM) corresponded to the first base of ORF presented in the DNA region having TP. Then we searched for a statistically significant extension of the TP region in the same gene in the presence of insertions and deletions of nucleotides by using modified profile analysis (Fig. 1). More than 800 genes contained a statistically significant shift between TP and ORF, which points to the presence of mutations in genes originating from the RF shift.

Second, we wished to check whether hypothetical amino acid sequences translated by using the RF of TP have homology with sequences from UniProtKB/Swiss-Prot databank (http://www.uniprot.org/). We made such a check for the genes that had mismatches between the gene's RF and TP. We confirmed the existence of such shifts for a part of the genes, since we found similarities between hypothetical amino acid sequences and amino acid sequences from the UniProtKB/Swiss-Prot databank.

2. Materials and methods

2.1. *Searching for TP in genes by method of information decomposition*

For each nucleotide sequence $S = \{s(i), i = 1, 2, \ldots, L\}$, we carried out a search for a region having maximally expressed TP by the method of information decomposition.[17,24] Let $s(i)$ be letters of the alphabet $A = \{a,t,c,g\}, A(1) = a, A(2) = t, A(3) = c$ and $A(4) = g$. Then we also make an artificial periodic sequence $U = \{u(i), i = 1, 2, \ldots, L\}$ of the same size as nucleotide sequence S. In artificial sequence, $u(i) = 1$ for $i = 1 + 3n$, $u(i) = 2$ for $i = 2 + 3n$, $u(i) = 3$ for $i = 3 + 3n$, where $n = 0, 1, 2, \ldots$. Then we choose coordinates L_1 and L_2 starting from the beginning of nucleotide sequence and fill the matrix $M^{4,3}$ for selected subsequence. Element of matrix $m(k,j)$ shows how many times symbol $A(k)$ in nucleotide subsequence from L_1 to L_2 matches the number j in artificial periodical sequence U. Each column j of the matrix $M(k,j)$ shows number of bases a, t, c and g that occur in positions $i = j + 3n$ of a sequence S, where $n = 0, 1, 2, \ldots$. We calculate mutual information as[26]

$$I = \sum_{i=1}^{4}\sum_{j=1}^{3} m(i,j) \ln m(i,j) - \sum_{i=1}^{4} x(i) \ln x(i) - \sum_{j=1}^{3} y(j) \ln y(j) + L \ln L, \qquad (1)$$

where $x(i)$ and $y(j)$ are the frequencies of nucleotide occurrences in sequence S and of numbers 1, 2 and 3 occurrences in artificial sequence, correspondingly.

We used gene sequences from KEGG-29 databank as source sequences S. All analysed sequences represent coding region (CDS) of genes without introns. That is why, when $L_1 - 1$ and L_2 are set to a multiple of 3 while defining sequences S and U, then first, second and third columns of matrix M for any values of L_1 and L_2 will always contain nucleotides corresponding to first, second and third bases of the gene's codon, respectively. In other words, matrix M is linked to RF, which exists in the analysed gene.

Doubled mutual information $2I$ has χ^2 distribution with six degrees of freedom. This allows us to estimate statistical significance of the periodicity found. We can reduce I to standard normal distribution:[27]

$$Z = \sqrt{2\chi^2} - \sqrt{2n - 1}. \qquad (2)$$

Conformity of $2I$ to χ^2 distribution with six degrees of freedom and of value Z to standard normal distribution are reached in the case of sufficiently large size of statistical data sample, i.e. sufficiently large length of a sequence S. In order to determine the minimal length of a sequence S that makes possible the usage of function χ^2 as approximation for $2I$ value distribution, we tested conformity of $2I$ to χ^2 distribution for various lengths of sequences S. We produced a set of nucleotide sequences for each length in the range from 30 to 1000 nucleotides by using a random number generator. Each of these sets contained 10 000 sequences. Thereafter, mutual information was calculated for each sequence from each set. For each set the histogram showing distribution of $2I$ value was also built. We compared this histogram with the theoretical distribution by χ^2 criteria. It was found that for sequences of length over 60 bp, $2I$ distribution conformed to $\chi^2(6)$ with a probability more than 99%. All sequences with TP, which were found in the current work, were longer than 60 bp. This allows using χ^2 distribution for statistical estimations of hitting $2I$ into interval from some threshold value $2I_0$ to infinity.

For sequence S, we calculated the values of $2I$ for all possible values of L_1 and L_2 ($L_1 < L_2 \leq L$) and chose the pair (L_1,L_2) for which the value of mutual information was maximal. Let us refer to the nucleotide sequence found in such a way as T.

If the value of Z was >5.0 for sequence T, then we considered that the region with TP has been found. The value of $Z >5.0$ ensures the probability of incidental TP revealing in DNA sequence to be $<10^{-6}$. Thereafter, we saved the found maximal sequence for the given gene, its coordinates in the given gene and periodicity matrix M, which shows the type of TP found. We chose a threshold level $Z > 5.0$ for finding TP in order to keep the number of incidentally found TPs near 1% of all detected regions with TP found in genes from the 29th release of the KEGG databank. To choose a threshold value for Z, we generated a set of random DNA sequences with the same size and sequence length distribution as for genes from the 29th release of KEGG databank. For $Z > 5.0$, the number of found random sequences was 7200, which is ~1.5% of found regions with TP (see below). For $Z > 6.0$, we found 172 such sequences, and for $Z > 7.0$, we found no such sequences. We intentionally chose the level of $Z >$ 5.0 in order to find the most complete extension of TP regions (see Section 2.2) that exist in various genes. The point is that gene's TP can be split up by insertions and deletions into several sections that may have rather low level of Z, but which is greater than 5.0. However, matrices M for each such section in gene will be identical or very similar, but cyclically shifted against each other (Fig. 1). In this case, consequent joining of these sections into a single one can considerably increase the statistical significance of a joined region that can be found by making an alignment against matrix M (see Section 2.3). Therefore, using a relatively low threshold value of Z will allow to not miss TP regions in genes separated into several sections by insertions and deletions.

2.2. *Algorithm of TP region extension in genes from KEGG databank*

We applied this algorithm for those DNA sequences in which we have revealed TP without insertions and deletions by the method of information decomposition. Let $S = \{s(i),\ i = 1,\ 2, \ldots,\ L\}$ be the analysed nucleotide sequence from KEGG databank and t_1 and t_2 be coordinates of the left and right borders of a region T with continuous TP in sequence S. For region T, we determined TPM and used this matrix to extend the TP region by considering possible nucleotides' insertions and deletions. Then we carried out local alignment of examined DNA sequence against weight matrix w introduced on the basis of TPM. Let the coordinates of start and end of found local alignment R be r_1 and r_2. Under extension of TP region T, we mean those R such that $r_1 < t_1$ or $r_2 > t_2$. Our goal is to choose from the KEGG databank only the genes that contain statistically significant extension of the TP region T. To find statistically significant extensions, we carried out global alignment of sequence S against weight matrix w and determined the values $\Delta F_T = F(t_2) - F(t_1)$ and $\Delta F_R = F(r_2) - F(r_1)$, where F is the value of similarity function on path of global alignment (Fig. 2). We selected only those genes that had $\Delta F_1 = \Delta F_R - \Delta F_T > 0$. Thereafter, we have to determine whether the value of ΔF_1 is statistically significant. To do this, we used the Monte Carlo method. We generated a set of random nucleotide sequences Q on which the region T was left unchanged and the regions of sequence S within the range from 1 to t_1 and from t_2 to L were shuffled in a random way. The set Q contained 10^6 sequences. For each sequence from the set Q, we built global alignment and determined the value $\Delta F_R - \Delta F_T$. Then we calculated such value ΔF_0 that the probability of $\Delta F_R - \Delta F_T \geq \Delta F_0$ for the set of sequences Q was $\sim 10^{-5}$. We chose the value ΔF_0 as a threshold and considered that if $\Delta F_1 > \Delta F_0$, then we found statistically significant extension of the region with continuous TP up to the bounds of region R via considering possible insertions and

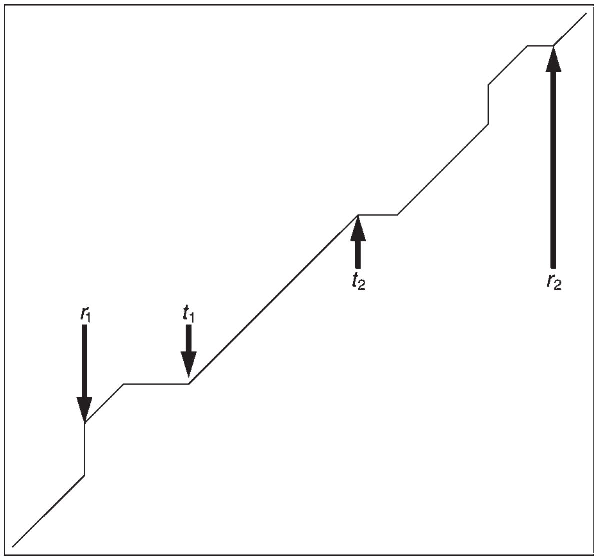


Figure 2. Example of alignment against weight matrix obtained from TPM. Here, t_1 and t_2 are coordinates of region with continuous TP found by information decomposition (T region) and r_1 and r_2 are coordinates of extended region with TP found by dynamic programming (R region). Points r_1 and r_2 on optimal alignment path have the coordinates (i_0,j_0) and (i_m,j_m) in matrix, respectively (see Section 2.3).

deletions of nucleotides. We analysed ~3 × 10^5 genes in such a way, and detected from 0 to 8 cases of region T extension due to accidental factors with a discovering probability of the given interval equal to 95%. As we revealed more than 750 genes with extension of the TP region T, then the fraction of genes with extension of region T due to merely accidental factors is ~1.1% which is a relatively small value.

Also, we estimated statistical significance Z_R of alignment R by the Monte Carlo method in a way we described earlier.[28,29] Thereto, we generated a set of sequences QR in which the region R was randomly shuffled. Then for each of these random sequences from QR set, we built global alignment and for each alignment we determined the values $\Delta G = G(r_2) - G(r_1)$. G is the similarity function for global alignment R, and it is calculated as we do for function F. For a set of ΔG values, we determined mean value $\overline{\Delta G}$ and $D(\Delta G)$, where $D(\Delta G)$ is the dispersion of ΔG. Then we determined the value of Z_R as:

$$Z_R = \frac{\Delta G_0 - \overline{\Delta G}}{D(\Delta G)}, \tag{3}$$

where $\Delta G_0 = G(r_2) - G(r_1)$ for original sequence R.

2.3. *Implementation of local and global alignment of nucleotide sequence against TPM*

During carrying out local and global alignment, we used matrices of TP M obtained for regions T of each gene from 29th release of KEGG databank. Using the matrix M, we built corrected position-specific matrix of the base weights as we suggested earlier:[28,29]

$$W'(i,l) = \log \frac{(m(i,l)+1)/\sum_{i=1}^{4}(m(i,l)+1)}{\sum_{l=1}^{3}(m(i,l)+1)/\sum_{i=1}^{4}\sum_{l=1}^{3}(m(i,l)+1)}, \tag{4}$$

where $w'(i,l)$ is the uncorrected weight of the base a_i at position l of profile. Corrected weight matrix W was calculated as:

$$w(i,l) = w'(i,l) - \tilde{N}\overline{w'}; \quad \overline{w'} = \frac{\sum_{i=1}^{4}\sum_{l=1}^{3} w'(i,l)}{12}, \tag{5}$$

where $\overline{w'}$ is the mean value of the uncorrected weight matrix. We added unity to each position of matrix $m(i,l)$ to eliminate the influence of zero elements in matrix M on the weights $w'(i,l)$. The value of C was selected for each matrix M using the QV sequences (see Section 2.4). Thereto, we varied C from interval $[-6,+6]$ with a step equal to 0.2. For each value of C from the interval, we determined alignments for N sequences from QV set, $N = 100$. For each alignment, we estimated Z_T (the statistical significance of T region alignment) by the Monte Carlo method (similar to Section 2.2). Thereto, we generated a set of sequences QT (for each sequences from QV set) in which the region T was randomly shuffled. Then we built a global alignment for each random sequence from QT set, and for each alignment, we determined the values $\Delta G(T) = G(t_2) - G(t_1)$. $G(T)$ is the similarity function for global alignment of region T for DNA sequences from QT set. For a set of $\Delta G(T)$ values, we determined the mean value $\overline{\Delta G(T)}$ and $D(\Delta G(T))$, where $D(\Delta G(T))$ is the dispersion of $\Delta G(T)$. Then we determined the value of Z_T as:

$$Z_T = \frac{\Delta G(T)_0 - \overline{\Delta G(T)}}{D(\Delta G(T))}, \tag{6}$$

where $\Delta G(T)_0 = G(t_2) - G(t_1)$ for original sequence T from QV set. Then we calculated:

$$X_C = \sum_{QV} Z_T. \tag{7}$$

We take the sum in Equation (7) for N sequences from QV set. As a result, we have a set of X_C where each C from interval $[-6,+6]$ has one X_C. We used for further calculation a value of C that has a maximum of X_C. We did the selection of C value for each matrix M.

Transition to weight matrix ensures assignment of higher weight to infrequent bases when they have high frequency in the given position of profile and, vice versa, assignment of lower weight for such bases having low frequency in the given position. To build optimal alignment of sought sequence against profile, we also introduced the weight for opening insertion or deletion ν_{do} and weight ν_{dc} for their continuation. Thereby the correlation in the formation of adjacent insertions or deletions is taken into account.

On the basis of introduced weights, we can find the optimal alignment, between analysed sequence and profile, i.e. we find such their subsequences for alignment, that maximise the similarity function. Let $S = \{s(j),\ j = 1,\ 2, \ldots,\ L\}$ be the analysed sequence. Let us create a profile matrix $q(i,j)$ of size L as:

$$\begin{aligned} q(i,j) &= w(i,l) \\ l &= j \bmod 3, \end{aligned} \tag{8}$$

where i shows the row number, $i = 1, 2, 3, 4$, and j the column index of a matrix q, $j = 1,\ 2, \ldots,\ L$.

To find local optimal alignment of sequence $s(j)$ against profiles $q(i,j)$, we applied the method of dynamic programming.[30] We iteratively calculated similarity function F according to the following equation:[31]

$$F(i,j) = \max\left\{\begin{array}{l} F(i-1,j-1) + w(s(i),j) \\ \max\limits_{1\le k\le d}\{F(i-k,j) - \nu_{do} - \nu_{dc}(k-1)\} \\ \max\limits_{1\le l\le d}\{(F(i,j-l) - \nu_{do} - \nu_{dc}(l-1)\} \\ 0 \end{array}\right\}. \tag{9}$$

Here, the index i stands for nucleotide in sequence $s(i)$ and index j for the column number in profile matrix q. Initial values for similarity function F are specified as:

$$F(0,0) = 0, \tag{10}$$

$$F(i,0) = 0, \tag{11}$$

$$F(0,j) = 0, \tag{12}$$

where d is the maximal number of insertions or deletions allowed. For calculations on triplet matrix, d equals 2. During building the local alignment, we determine the maximal value of the similarity function F coordinates $(i_m j_m)$ corresponding to this maximal value. Then we determine the path from points $(i_m j_m)$ to $(i_0 j_0)$, where the value of the similarity function becomes zero for the first time. According to the path made, we built alignment between sequence S and profile matrix q. For building alignment, nucleotides of sequence S and numbers $l = \mathrm{mod}3(j)$ indicating columns of matrix w [Equation (5) were used].

We used Equation (9) for carrying out the global alignment but without using zero in the right part of the equation. All other parameters were the same as in the case of the local alignment. Initial values for similarity function F for global alignment are specified as:[31]

$$F(0,0) = 0, \tag{13}$$

$$F(i,0) = F(0,0) - \nu_{do} - \nu_{dc}(i-1), \tag{14}$$

$$F(0,j) = F(0,0) - \nu_{do} - \nu_{dc}(j-1). \tag{15}$$

2.4. *Choosing values for ν_{do} and ν_{dc}*

We chose values ν_{do} to prohibit the optimal path in local and global alignments from moving round the nucleotide having minimal weight in matrix $w(i,j)$ because of two sequential insertions. On this basis, we chose $\nu_{do}0.5\min\limits_{i,j}(w(i,j))$ At the same time, it was important to make the alignment able to find region T revealed by the method of information decomposition (Section 2.1) without any insertions and deletions. In order to choose the corresponding coefficient ν_{do} for each gene from KEGG databank, we generated a set of random sequences QV, where regions of sequence S in the range from 1 to t_1 and from t_2 to L were shuffled in a random manner. The region of sequence S in the range from t_1 to t_2 was constructed by TPM M. Then matrix M was transformed to M' as (see also Section 2.1):

$$m'(i,j) = \frac{m(i,j)}{y(j)}. \tag{16}$$

Then we generated in a random manner three nucleotide sequences $S1(k)$, $k = 1, 2, \ldots, y(1)$, $S2(k)$, $k = 1, 2, \ldots, y(2)$ and $S3(k)$, $k = 1, 2, \ldots, y(3)$. Nucleotide probabilities in sequence $S1$ are equal to probabilities $m'(i,1)$, nucleotide probabilities in sequence $S2$ are equal to probabilities $m'(i,2)$ and nucleotide probabilities in sequence $S3$ are equal to probabilities $m'(i,3)$. Index i varies from 1 to 4. Thereafter, nucleotides of sequence $S1$ took positions $k = t_1, t_1 + 3, \ldots, t_2 - 2$; nucleotides of sequence $S2$ took positions $k = t_1 + 1, t_1 + 4, \ldots, t_2 - 1$ and nucleotides of sequence $S3$ took positions $k = t_1 + 2, t_1 + 5, \ldots, t_2$. This algorithm was used for the generation of 1000 random sequences contained in QV.

Selection procedure of the ν_{do} value for each gene from the KEGG databank from which region T was extracted is described below. First, we chose the value for $\nu_{do} = 0.6\min\limits_{i,j}$ and $\nu_{dc} = 0.25\,\nu_{do}$. Then we made alignments for the set of random sequences QV and determined the number of sequences having insertions or deletions within the region from t_1 to t_2. If the fraction of sequences in set QV having at least one insertion or deletion within region from t_1 to t_2 exceeded 1%, then we increased the value of ν_{do} by $0.1\min\limits_{i,j}$ and calculated $\nu_{dc} = 0.25\,\nu_{do}$ again. Alignments of sequences from set QV were also built again using these new parameters. If the fraction of sequences with insertions or deletions was under 1%, then the process was stopped and the obtained value ν_{do} was used for finding the extended region R by local alignment (Sections 2.2 and 2.3). If the fraction of sequences with insertions or deletions was over 1%, then we increased ν_{do} and ν_{dc} again as shown earlier and alignments of sequences from set QV were built again.

3. Results and discussion

3.1. *Finding extensions of the regions with TP for genes accumulated in KEGG databank*

We analysed 578 868 genes accumulated in 29th release of the KEGG databank (http://www.genome.ad.jp/kegg/). We found 472 288 regions having continuous TP in 457 333 genes. These data indicate that 79% of genes have regions with TP. These results conform to earlier works on TP detection by either using informational methods or other techniques.[4–13,32] From these 472 288 regions containing continuous TP, we selected only those the lengths of which were significantly less than that of a gene. This means that the distance from the left and right edges of the TP region to the start and end of the gene was more than 30 bp. This criterion was satisfied for 326 933 TP regions. Then we aligned nucleotide sequences of corresponding genes against TPM (see Sections 2.2 and 2.3), which were revealed in the gene by the method of information decomposition (Section 2.1).

We revealed totally 824 genes in which TP regions were extended by using alignment against TPM. General information describing all these sequences can be found in the Section Supplementary data. Details including periodicity alignment and found protein similarities can be found in online databank installed at http://victoria.biengi.ac.ru/pertails/.

3.2. *Finding protein similarities for the products obtained using active and ancient gene RFs*

Let us consider those genes in which the region of continuous TP was extended by taking into account nucleotide insertions and deletions (see Section 2.2). Further, we will discuss nucleotide sequences with coordinates from r_1 to t_1 and from t_2 to r_2 (Fig. 2). Let us call these sequences $T1$ and $T2$ (the region of continuous TP was earlier referred to as T). In sequences $T1$ and $T2$, we found TP with insertions and deletions that were very similar to continuous TP T. This is the reason why we found statistically significant alignment from r_1 to r_2 in sequence S against the weight matrix constructed on the basis of TPM M. We suppose that TP found in sequences $T1$ and $T2$ is a trace of some ancient RF that existed in these nucleotide sequences earlier. First column of the matrix M corresponds to the first codon base in sequence S, whereas due to insertions and deletions of nucleotides, in subsequences T1 and T2 matrix M corresponds to alternative ancient RF which may not match the actual RF there. Let us refer to RF specified by matrix M as 'ancient RF'. We suppose that it is possible to reveal similarity between amino acid sequences obtained by ancient RF from nucleotide sequences $T1$ and $T2$ and amino acid sequences accumulated in modern databanks like UniProtKB/Swiss-Prot. Such an assumption is based on the idea that if a gene responsible for the same genetic function existed in several genomes, then insertion or deletion of nucleotides in this gene within one genome does not ultimately lead to analogous changes in another genome. It is important that these sequences should be now known and should not have accumulated many evolutional alterations. This will allow us to see their similarity. We conducted such an investigation within the scope of the present work. Sequences of regions $T1$ and $T2$ were translated to amino acid sequences in accordance with the RF existing in the gene and in accordance with ancient RF's revealed by TPM M. For all amino acid sequences obtained in such a way, we searched for their similarities with sequences accumulated in UniProtKB/Swiss-Prot databank[33] by using BLAST program.[34] Totally we found 824 T regions in genes that were extended via joining regions $T1$ and/or $T2$. Sixty-four of TP regions T had protein similarities with regions $T1$ and/or $T2$ only by ancient the RF constructed on the basis of matrix M (see above). At the same time, for 176 T regions, protein similarities in regions $T1$ and/or $T2$ were found only for the active RF of a gene, and in three cases, for both RFs. Five hundred and eighty-one T regions have no protein similarities in regions $T1$ and/or $T2$ for amino acid sequences constructed either by the existing gene's RF or by the ancient RF constructed on the basis of matrix M.

Let us consider an example of continuous TP extension in a putative dehydratase gene (locus 'SMa0056' in KEGG databank). This gene has a length of 1134 bp and contains the region of continuous TP from 16th to 840th nucleotide. TPM for region T and weight matrix $w(i,j)$ is of the form that is shown in Fig. 3. We extended the found TP region T by adding region $T2$ that leads to the appearance of TP in this gene from the 840th to 1134th nucleotides. Statistical significance Z_R of the found alignment R

	1	2	3
A	74	78	13
T	34	83	36
C	64	63	138
G	103	51	88

1	1	2	3
A	0.26	0.31	-1.42
T	-0.43	0.45	-0.37
C	-0.35	-0.37	0.41
G	0.21	-0.48	0.05

Figure 3. TPM M and position-specific weight matrix $w(i,j)$ built for T region from 16th to 840th nucleotides of the locus 'SMa0056' (KEGG databank). For the calculation of weight matrix, we used $C = -0.3$, $\overline{W} = -0,1$.

1	Gene sequence	818 tgatcctgctcgacaccaccttctgg**ggcggatccgccctgcgtgaaggcggcgggcgtc** 877
2	TP consensus	tcatcatcatcatcatcatcatc***atcatcatcatcatcatcatcatcatcatcatcatc**
3	Existing RF in Kegg databank	tgg**ggcggatccgccctgcgtgaaggcggcgggcgtc** 23123123123123123123123123**1231231231231231231231231231231231**
4	Ancient RF obtained by TPM	23123123123123123123123***123123123123123123123123123123123123** **ggggcggatccgccctgcgtgaaggcggcgggcgtc**
5	Translation by existing RF	I L L D T T F W **G G S A L R E G G G R**
6	Translation by ancient RF	I L L D T T F **G A D P P C V K A A G V**

Figure 4. Sequence of 'SMa0056' gene from 818th to 877th nucleotide (1) aligned against weight matrix $w(i,j)$. Insertion of 841st nucleotide (nucleotide t) is shown by asterisk. Further under gene, the consensus sequence for TP (2) is shown, which is constructed by matrix M and also by RF existing in gene (3) and RF specified by TPM M named as ancient RF (4). Below translated fragments of the nucleotide sequences are shown. Translation is made according to RF existing in gene (5) and to ancient RF (6). Regions having RF shift are marked in bold.

was equal to 17.7. The value ΔF_0 equalled 10.25, whereas the value ΔF_1 equalled 13.85. This provides the probability of incidental addition of region $T2$ to region T at a level $<10^{-5}$.

During alignment of the sequence relative to found continuous TP, we revealed the insertion of the nucleotide at position 841 (nucleotide t) of the locus (Fig. 4). This means that after the 840th nucleotide (towards first nucleotide), a shift occurs between TP and RF. Thus, the first codon base in sequence $T2$ corresponds to the third column of matrix M, whereas in sequence T, the first codon base $T2$ corresponds to the first column of matrix M (Fig. 4). We suppose that in the past this gene existed without insertion of the nucleotide t in position 841 (towards first nucleotide of a gene) and mutations could not blur the TP existing in this region (Fig. 5). To check this hypothesis, we recoded $T2$ region to amino acid sequence by the actual gene's RF and by the ancient RF (Fig. 4). Then we searched for similarities with these two amino acid sequences in UniProtKB/Swiss-Prot databank. Consequently, we found five cases of similarities for hypothetic sequence only, i.e. for the amino acid sequence obtained by RF specified by TPM M (Fig. 4). At the same time, amino acid sequence constructed by RF specified in KEGG had no similarities at all (Fig. 4). Similarities for hypothetical amino acid sequence were found to dehydratases. Match ratio was >30% (e-value in range from 10^{-12} to 10^{-7}), i.e. found similarity is evolutionally distant and statistically significant. All found similarities can be accessed at http://victoria.biengi.ac.ru/pertails/perinfo.php?perid=270869.

It is also of great interest to point out that the amino acid sequence corresponding to the gene 'SMa0056' from the KEGG databank has similar amino acid sequences in the UniProtKB/Swiss-Prot databank. This amino acid sequence is designated as 'Q931B9_RHIME' in the UniProtKB/Swiss-Prot databank. Similarity is observed from the start of the protein to the 280th amino acid. This is true for a set of different mandelate racemase/muconate lactonizing enzymes and glucarate dehydratases (e.g. 'Q024C0_SOLUE', 'A9FRR8_SORC5'). This set also contains five amino acid sequences ('GUDH_STRCO', 'GUDX_ECOLI', 'GUDH_ECO57', 'GUDH_ECOLI' and 'GUDH_PSEPU'), where we found similarity to amino acid sequence produced from 842nd to the end of the gene by the ancient RF. Therefore, it can be assumed that these five sequences existed before the insertion of the 841st or independently of it. These sequences have similarity to the sequence

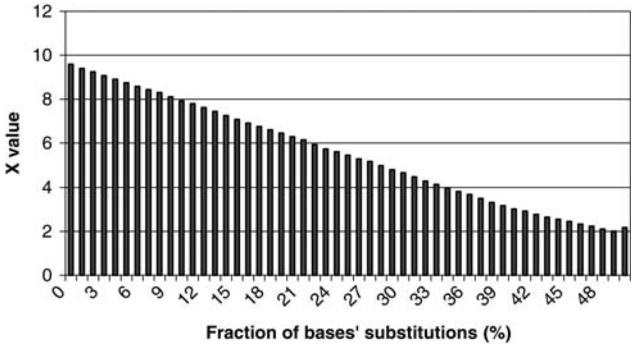


Figure 5. Resistance of sequence T with continuous TP of gene 'mlr4742' to nucleotides' substitutions. We selected a base in sequence T in a random manner and then substituted it, also in a random manner, by one of the nucleotides a, t, c or g (with probability of each being equal to 0.25). Thereafter, we estimated statistical significance of obtained sequence according to Equation (2). As follows from figure, statistical significance of TP decreases relatively slow during mutations' accumulation. For example, 10% difference from the original sequence still allows revealing TP at statistically significant level.

'Q931B9_RHIME' from 1st to 280th ±10 amino acids for the gene's (KEGG) RF and from 281st to 378th amino acids for ancient RF.

3.3. Discussion

In this work, we managed to show that investigation of shifts between TP and RF can reveal possible mutations via RF shift in a gene. We found 824 such genes in which there existed some single-type TP regions separated by deletions or insertions of nucleotides. They account for ~0.2% of the total number of analysed genes. We suppose that in the past RF and TP in these regions were explicitly linked to each other and shifts between them appeared only after deletions or insertions of nucleotides. Such relatively small fractions of genes in which an RF shift can potentially occur can be explained by the following reasons. First, we searched only for relatively short deletions or insertions having lengths up to 2 nucleotides ($d = 2$). We confined ourselves to small deletions and insertions because while the length of deletion and insertion increases, the value of function F decreases [see Equation (6)]. Costs of deletion and insertion can be too high; in this case, the value ΔF_1 will be statistically insignificant. Thus, the current method will miss the major part of genes containing long deletion and insertion of nucleotides. Second, the technique used works well for the small number of regions where deletions or insertions of nucleotides occurred. If density of deletions and insertions is more than one deletion and insertion for several tens of nucleotides (~50), then the accurate placement of deletions and insertions using this algorithm will be difficult. This will lead to the statistically insignificant value of ΔF_1 for such a gene.

In general, in this work, we aimed to reveal the lower bound for the number of genes in which a shift between RF and TP is possible. Actually, there can be much more such genes. These values are also confirmed by data in work[2] where the number of genes with RF shift obtained by BLAST program is >1%.

The technique used to find shifts between TP and RF and to reveal mutations via RF shift in genes seems to be more preferable than the usage of BLAST program for finding possible similarities. The current method of revealing mutations via an RF shift in the gene does not require finding similarities in the databank of amino acid sequences. Owing to the limited size of the databank, there will always exist a chance that similarities will not be found, although actually mutation via RF shift will exist. We suppose that complete revealing of mutations via an RF shift in genes is possible by the integration of these two techniques. That is, we also have to investigate genes having $0 \leq \Delta F_1 \leq \Delta F_0$ and to consider that we found mutations via an RF shift if regions $T1$ and $T2$ have statistically significant similarities. In this case, the TP just indicates the possibility of an RF shift and the fact of such mutations can be considered to be proved only after revealing similarities with regions $T1$ and $T2$. On the other hand, enhancement of the algorithm used in the current work can facilitate the use of more perfect algorithms for finding TP, such as Markov models. This probably will allow us to reveal RF shifts induced by a set of nucleotides insertions and deletions in various gene regions.

From the functional point of view, mutations via RF shift seem to be events that are able to cardinally change the gene function. This fact can explain the relatively small number of such events found during investigations in the past and in the current works.[2,3,35] They can make a great contribution to the formation of new genes by copying known genes and generating mutations via an RF shift[2,3,35] in them. However, the genetic code has to be adapted for this event in some way,[36] and the new amino acid sequence has to possess some biological function. Otherwise, overrun of mutational events for the creation of the new gene's function can be too great, and impossible within reasonable evolutional time.

In the light of these assumptions, TP can be some kind of test to check the gene's integrity. If the gene was duplicated in the genome, then its new copy may fail the check, which opens up possibilities for evolutional changes of the gene's copy via the RF shift and for the creation of a new gene with a new biological function as a result.

Supplementary data: Supplementary data are available online at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by Russian Academy of Sciences.

References

1. Wei, Q., Li, L. and Chen, D. J. 2007, *DNA Repair, Genetic Instability, and Cancer*, World Scientific Publishing Co. Pte, Ltd., Singapore.
2. Okamura, K., Feuk, L., Marquès-Bonet, T., Navarro, A. and Scherer, S. W. 2006, Frequent appearance of novel protein-coding sequences by frameshift translation, *Genomics*, **88**, 690–697.
3. Raes, J. and Van de Peer, Y. 2005, Functional divergence of proteins through frameshift mutations, *Trends Genet.*, **21**, 428–431.
4. Fickett, J. W. 1998, Predictive methods using nucleotide sequences, *Methods Biochem. Anal.*, **39**, 231–245.

5. Staden, R. 1994, Staden: statistical and structural analysis of nucleotide sequences, *Methods Mol. Biol.*, **25**, 69–77.
6. Baxevanis, A. D. 2001, Predictive methods using DNA sequences, *Methods Biochem. Anal.*, **43**, 233–252.
7. Gutiérrez, G., Oliver, J. L. and Marín, A. 1994, On the origin of the periodicity of three in protein coding DNA sequences, *J. Theor. Biol.*, **167** (4), 413–414.
8. Gao, J., Qi, Y., Cao, Y. and Tung, W. W. 2005, Protein coding sequence identification by simultaneously characterizing the periodic and random features of DNA sequences, *J. Biomed. Biotechnol.*, **2**, 139–146.
9. 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.
10. Eskesen, S. T., Eskesen, F. N., Kinghorn, B. and Ruvinsky, A. 2004, Periodicity of DNA in exons, *BMC Mol. Biol.*, **5**, 12.
11. Bibb, M. J., Findlay, P. R. and Johnson, M. W. 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** (1–3), 157–166.
12. Konopka, A. K. 1994. Sequences and codes: fundamentals of biomolecular cryptography. In: Smith, D. (ed.), *Biocomputing: Informatics and Genome Projects*, Academic Press: San Diego, pp. 119–174.
13. Trifonov, E. N. 1999, Elucidating sequence codes: three codes for evolution, *Ann. NY Acad. Sci.*, **870**, 330–338.
14. Eigen, M. and Winkler-Oswatitsch, R. 1981, Transfer-RNA: the early adaptor, *Naturwissenschaften*, **68**, 217–228.
15. Zoltowski, M. 2007, Is DNA code periodicity only due to CUF—codons usage frequency?, *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, **1**, 1383–1386.
16. 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** (1), 36–43.
17. Korotkov, E. V., Korotkova, M. A., Frenkel', F. E. and Kudriashov, N. A. 2003, The informational concept of searching for periodicity in symbol sequences, *Mol. Biol. (Mosk)*, **37**, 436–451.
18. Issac, B., Singh, H., Kaur, H. and Raghava, G. P. S. 2002, Locating probable genes using Fourier transform approach, *Bioinformatics*, **18** (1), 196–197.
19. Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R. 1997, Prediction of probable genes by Fourier analysis of genomic sequences, *Comput. Appl. Biosci.*, **13** (3), 263–270.
20. Azad, R. K. and Borodovsky, M. 2004, Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory, *Brief. Bioinform.*, **5** (2), 118–130.
21. Henderson, J., Salzberg, S. and Fasman, K. H. 1997, Finding genes in DNA with a hidden Markov model, *J. Comput. Biol.*, **4**, 127–141.
22. Snyder, E. E. and Stormo, G. D. 1993, Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks, *Nucleic Acids Res.*, **21**, 607–613.
23. Thomas, A. and Skolnick, M. H. 1994, A probabilistic model for detecting coding regions in DNA sequences, *IMA J. Math. Appl. Med. Biol.*, **11** (3), 149–160.
24. Korotkov, E. V., Korotkova, M. A. and Kudryshov, N. A. 2003, Information decomposition method for analysis of symbolical sequences, *Phys. Lett. A*, **312**, 198–210.
25. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. 1999, KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res.*, **27**, 29–34.
26. Kullback, S. 1959, *Information Theory and Statistics*, Wiley: New York.
27. Hudson, D. J. 1964, *Statistics: Lectures on Elementary Statistics and Probability*, CERN: Geneva.
28. Frenkel, F. E., Chaley, M. B., Korotkov, E. V. and Skryabin, K. G. 2004, Evolution of the tRNA-like sequences and genome variability, *Gene*, **335**, 57–71.
29. Chaley, M. B., Korotkov, E. V. and Kudryashov, N. A. 2003, Latent periodicity of 21 bases typical for MCP II gene is widely present in various bacterial genes, *DNA Seq.*, **14**, 37–52.
30. Needleman, S. B. and Wunsch, C. D. 1970, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.*, **48** (3), 443–453.
31. Durbin, R., Eddy, S. R., Krogh, A. and Graeme Mitchison, G. 1999, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK.
32. Frenkel, F. E. and Korotkov, E. V. 2008, Classification analysis of triplet periodicity in protein-coding regions of genes, *Gene*, **421**, 52–60.
33. UniProt Consortium 2007, The universal protein resource (UniProt), *Nucleic Acids Res.*, **35**, 193–197.
34. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215** (3), 403–410.
35. Kramer, E. M., Su, H. J., Wu, C. C. and Hu, J. M. 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.
36. Bollenbach, T., Vetsigian, K. and Kishony, R. 2007, Evolution and multilevel optimization of the genetic code, *Genome Res.*, **17** (4), 405–412.