AUSTRIAN JOURNAL OF STATISTICS
Volume 40 (2011), Number 1 & 2, 137-146

Detection of Possible Reading Frame Shifts in Genes Using Triplet Frequencies Homogeneity

Valentina Rudenko[1,2], Yulia Suvorova[1] and Eugene Korotkov[1,2]
[1]Bioengineering Centre of RAS, Moscow
[2]NRNU MEPHi, Moscow

Abstract: A new approach for detecting the reading frame shifts in coding DNA sequences has been developed. To detect the shift the hypothesis of homogeneity of triplet frequencies along the sequence was checked. The statistical significance was estimated by using Monte Carlo method. The method developed has allowed revealing 25% more cases of frame shifts in sequences with a length greater than 300, than the approach used earlier.

Keywords: Coding DNA Sequence, Reading Frame Shift, Statistical Significance, Monte Carlo Method.

1 Introduction

As it is known, DNA sequences are changed during evolution due to mutation process. Mutations can represent replacements of one base by the other one. At the level of amino acid sequence, replacement can lead to change of only one amino acid. Often even such insignificant mutation can render serious disease because of protein disability to perform its biological function Watson et al. (2007). The other possible type of mutations is insertions and deletions (indels) of nucleotides. They take place much rarer than replacements of bases, however they represent more significant evolutionary events. The indel which length is not divisible by three changes the lengthy fragment of amino acid sequence since it causes the reading frame (RF) shift.

Influence of RF shifts in genes on structure and function of proteins is poorly investigated. In some cases shift can be the cause of hereditary disease since the protein coded by the changed sequence loses its functions. For example, it has been shown that deletion in LAMA2 gene leads to a congenital muscular dystrophy (Salem et al., 2010). The mutation in a gene transcription factor NKX2.5 caused by a frame shift is associated with a congenital heart disease (Stallmeyer, Fenge, Nowak-Gottl, and Schulze-Bahr, 2010).

The extremely important task is revealing of RF shifts caused by mutations in the genomes of potential parents. It will allow making the probabilistic prediction regarding the possibility of a specific disease for their descendants. Also it opens prospects for prophylaxis and prevention of a disease. It is possible to show the examples of sequences in which RF shifts have not affected their functional properties. In this case it is assumed that a shift has occurred in insignificant region and has not affected the functional center of the molecule. The researchers who are engaged in sequencing of new DNA sequences also face the necessity of detecting RF shifts. Sequencing errors, such as indels in DNA sequence, lead to impossibility of annotating amino acid sequence which has been coded by erroneous DNA sequence. The facts listed above have led to the necessity of development of mathematical algorithms and computer programs for detection of RF shifts in

DNA sequences. Earlier some methods of detecting shifts were proposed: (Fichant and Quentin, 1995) based on statistical methods; (Posfai and Roberts, 1992; Claverie, 1993) using dynamic programming; (Schiex, Gouzy, Moisan, and Oliveira, 2003) using Markov models.

Most of the methods dedicated to detecting RF shifts were associated with the beginning of large-scale sequencing projects and the tasks of revealing sequencing errors and annotating new sequences arisen during these projects. These methods were based on searching for the homologous sequences in Swiss-prot databank. They have a number of restrictions. Firstly, they require choosing, based on some characteristic, a gene in which RF shift is supposed to occur, and then finding the possible region of shift. The general detecting of RF shifts in all genes can require rather large computational resources. Secondly, Swiss-prot databank should contain an amino acid sequence having statistically significant similarity with amino acid sequence being investigated. However, such sequence can be absent due to limited volume of Swiss-prot databank or because of too large evolutionary distinctions between amino acid sequences. Thus this approach can reveal only some part of RF shifts remained in existing genes up to the present time. Actually, the methods based on searching for homologous sequence have revealed only few hundreds of genes having RF shifts.

(Fichant and Quentin, 1995) have tried to eliminate the basic disadvantage of the previous methods. In the sequence under investigation they detect the coding frame and the positions of the sequence in which this coding frame is changed were identified as the positions of RF shifts. It have appeared that the best measures for detecting the coding frame was based on deviations of 3- and 6-mers frequencies from the expected values for the investigated genome.

However, the expected frequencies can differ among each other even within one genome. As such, it is impossible to create an universal measure for all sequences from a genome. It is necessary to make adjustment of the method parameters depending on category to which the investigated sequence belongs, for example, the category with low or high value of CBI parameter (codon bias index proposed by (Bennetzen and Hall, 1982)).

Despite of a relative diversity of the developed methods and programs for detecting RF shifts, currently no universal approach exists which would work well for any sequences, including the ones with unknown functions, and which does not use *a priori* information. information. Besides that, earlier no full investigation regarding the presence of RF shifts in all accessible DNA sequences stored in specialized databanks has been carried out.

The purpose of the present work was the development of a new approach for detecting RF shifts which would be more powerful than existing analogues and which would not use any *a priori* information such as the presence of homologous sequences in Swiss-prot databank, or nucleotide or amino acid frequencies. Also it was required to estimate how often the mutations in the genome associated with RF shift take place. To solve the last problem we detected RF shifts in coding DNA sequences from *Kegg-46* databank `http://www.genome.ad.jp/KEGG/`.

2 Methods

2.1 The Measure of the RF Shift

Earlier we applied the statistics based on the triplet periodicity type similarity in fragments of the fixed length w to the right and to the left of the position of putative shift as the quantitative characteristic of RF shift, see Korotkov and Korotkova (2010). Since the triplet periodicity is caused by the preferences of an organism in using various triplets for coding the same amino acids and degeneracy of a genetic code, it is possible to assume that the coding sequence without RF shift has homogeneous distribution of triplet frequencies on comparatively long segments. This assumption has been used as a basis of construction of a new measure of RF shift in DNA sequences.

Let $S = s(k)$, $k = 1, 2, \ldots, L$ is a DNA sequence, where $s(k)$ belongs to a set of symbols *a, t, c, g*. DNA sequence codes the amino acid sequence of a protein, so that nonintersecting triplets code amino acids. Coding of amino acid sequence can occur in three possible ways: the first positions of triplets are located at $1 + 3n$, at $2 + 3n$, or at $3 + 3n$ positions of the DNA sequence. These ways of coding will be referenced as $T1$, $T2$ and $T3$. They represent various reading frames in a gene (Figure 1).

```
DNA sequence :  ...atggcgagagaggtgcctatagagaaattg...
     T1 :       ...123123123123123123123123123...
Am.acid seq1 :  ...M..A..R..E..V..P..I..E..K..L.....
     T2 :       ...312312312312312312312312312...
Am.acid seq2 :  ....W..R..E..R..C..L..$..R..N.......
     T3 :       ...231231231231231231231231231...
Am.acid seq3 :  .....G..E..R..G..A..Y..R..E..I......
```

Figure 1: Three possible reading frames in a gene ($ - stop codon).

To determine whether there exists a RF shift in a position k (k is divisible by 3) of the sequence we may consider a fragment of sequence with coordinates $(k - w + 1, k + w)$, where w is a width of a window before and after a tested position k. For simplicity, we will consider the values of w divisible by 3. According to our assumption, if RF shift does not exist, then different triplets are distributed uniformly along all length of a fragment. At a presence of RF shift, if there was an insertion or deletion with a length not divisible by 3 in a position k, then the uniformity of triplet frequencies should be observed for the first half of a fragment coded by the frame $T1$ and for the second half of a fragment coded by the frame $T2$ or $T3$.

Let us check the hypothesis regarding the uniformity of distributions of triplet frequencies between a fragment of a sequence with coordinates $(k - w + 1, k)$ in the frame $T1$ and fragments $(k + j, k + w + j - 1)$, in the frame ***Tj***, $j = 1..3$.

Under this formulation, detecting RF shifts in DNA sequences seems to be similar to standard change-point problem for a sequence of multinomial observations (Wolfe and Chen, 1990). Actually we have to test the hypothesis of homogeneity of two polynomial samples. To solve this problem, a lot of methods were developed which were successfully applied earlier, for example, to investigation of segmentation of rather long DNA

sequences, see Braun, Braun, and Müller (2000), Boys and Henderson (2004). But the difficulty of our task is that the width of a window w has to be as small as possible, because the average gene length is about several hundreds bases. So the length of the region near the site of RF shift in which the triplet frequencies are homogenous equals to several hundreds bases too. To apply the known methods to detecting change-points, rather long width of a window w that allows calculating triplet probabilities is required. For all 64 triplets the width should be greater or equal to 1920 bases (about 10 events for each triplet). This fact makes impossible the detection of RF shifts in short genes. So we decided to apply the methods for testing statistical hypotheses, and we used the information criterion introduced by Kullback (1959) to test the homogeneity of two samples:

$$I_j = \sum_{i=1}^{64} f_i \log f_i + \sum_{i=1}^{64} v_i^j \log v_i^j - \sum_{i=1}^{64} (f_i + v_i^j) \log(f_i + v_i^j) + (N_1 + N_2) \log(N_1 + N_2) - N_1 \log N_1 - N_2 \log N_2 \,. \quad (1)$$

Here, f_i are triplet frequencies for the sequence fragment with coordinates $(k-w+1, k+w)$; v_i^j are triplet frequencies for the fragment $(k+j, k+w+j-1)$; and $N_1 = N_2 = w/3$ is the number of triplets in each half of the fragment considered. If RF shift does not exist in the position k, then minimal I_j will be reached for $j = 1$. If I_2 or I_3 takes the small value and I_1 is large, then the shift is observed in the position k.

The statistics $2I_j$, has a χ^2 distribution with 63 degrees of freedom. However, according to Filina and Zubkov (2008) the distribution $2I_j$ is well conformed to standard distribution only if all f_i and v_i^j are equal to or greater than 10. During investigation of real DNA sequences the given requirement is not satisfied. For example, stop codons do not occur in a coding frame, while they can exist in alternative frames. This does not allow using standard formulas for calculation of the statistical significance of putative RF shift. Therefore it has been decided to apply Monte Carlo method for this purpose.

2.2 Estimation of the Statistical Significance of RF Shift by a Monte Carlo Method

Monte Carlo method for estimation of the statistical significance of putative RF shift consists of the following steps. Sequences $Sj = S(k-w+1, k)||S(k+j, k+w+j-1)$ are constructed, where $j = 1..3$, $S(m, n)$ - a fragment of the sequence S starting at $s(m)$ and finishing at $s(n)$ symbol, $||$ - concatenation of two fragments of the sequences. Thus, $S1$ coincides with the initial sequence, $S2$ is the sequence obtained from $S1$ as a result of deletion of one symbol from kth position, $S3$ is the sequence obtained from $S1$ as a result of deletion of two symbols starting from kth position.

Then on the basis of sequences Sj three sets of random sequences were generated with the same triplet frequencies as in sequences $S1$, $S2$, $S3$. Random sequences were obtained from initial ones by shuffling their triplets. Shuffling was performed using a random number generator as follows. Firstly, a random number from 1 to 10000 was associated with each triplet of the sequence. Then these numbers were ordered in ascending order, while triplets changed their position in accordance with the corresponding random numbers.

The number of random sequences in each set was equal to $M = 200$. Statistics I_j^n was calculated for each random sequence using the equation (1), here $j = 1..3$ is the index of a set, $n = 1..200$ is the index of a sequence in the set. For all sequences included into the set j the average value $I_j^{\exp}$ and the dispersion σ_j^2 of random value I_j were determined. The abbreviation $\exp$ means "expected", i.e. this value characterizes the most probable value of statistics I_j for DNA sequence with the same triplet frequencies as in the sequence Sj. We designate as I_j^{ob} the value calculated using the equation (1) for initial sequence Sj. Then let us turn from I_j to statistics Z_j:

$$Z_j = \frac{I_j^{\text{ob}} - I_j^{\exp}}{\sigma_j} \,. \tag{2}$$

The values Z_j approximately have standard normal distribution (Sprinthall, 2002). We designate as $p_j = P(N(0,1) > Z_j)$ a probability that a standard normally distributed random variable is greater than or equal to Z_j. In case of presence of a shift Z_1 have rather large value (heterogeneity), p_1 is small. At the same time, one of Z_j ($j = 2$ or $j = 3$) is small, and p_j is large enough. In practice it is convenient to merge two conditions of presence of a shift in one relation and use it for identification of a shift functions:

$$F_2 = -\log(p_1/p_2)\,, \qquad F_3 = -\log(p_1/p_3)\,. \tag{3}$$

If the value F_2 exceeds F_3 and exceeds some threshold level F_0, then the insertion of $1 + 3n$ or deletion of $2 + 3n$ nucleotides is observed in the position k. Similarly, if $F_3 > F_2$ and $F_3 > F_0$, it is possible to consider that the insertion of $2 + 3n$ or deletion of $1 + 3n$ nucleotides takes place in the position k. The method of threshold level F_0 determination will be presented in one of the following sections.

2.3 The Algorithm of Revealing Reading Frame Shifts in DNA Coding Sequences

The following algorithm was used for searching for RF shifts in the set of DNA sequences. We considered all possible shift positions k which were divisible by 3, starting from 150. For these positions the functions F_2 and F_3 for the same window width to the right and to the left from a position k were calculated. The width of a window w varied from 150 up to 600 symbols with a step 30. It was considered that RF shift in the position k of the sequence was found if for any window width w one of the functions F_2 or F_3 exceeded the threshold value F_0.

The algorithm of revealing RF shifts has been implemented as a C++ program using the library of parallel programming MPI. Since calculations by Monte Carlo method require large computational efforts, scanning *Kegg-46* databank was performed on a computer cluster of the Bioengineering Center of RAS consisting of more than 110 processors.

3 Results

3.1 Testing the Algorithm Developed on the Artificial DNA Sequences with Shifts

First of all, we have tested the method developed on the artificial sequences with shifts. For this purpose some DNA sequences from *Kegg-46* databank were chosen in which the shift has not been identified. Then in random positions inside these sequences we added symbols with the purpose of simulating a shift. One example of the tested sequences is the sequence with identifier *aq_023*. It relates to *Aquifex aeolicus* genome and codes *acetylornithine aminotransferase*. For this sequence, graphs of the functions F_2 and F_3 depending on a position of putative shift k have been constructed, here the window length was equal to 390 bases (Figure 2a).

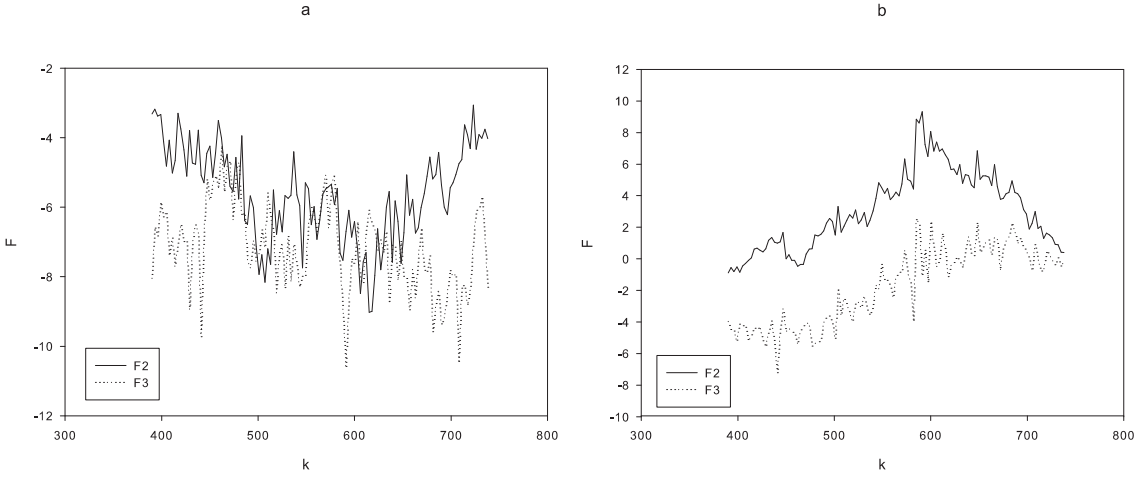

Figure 2: Graphs of the functions F_2 and F_3 for the sequences: a) aq_023 from Aquifex aeolicus genome, b) the same sequence with an insertion of a symbol "a" at position 600.

The triplet distribution is homogeneous along the sequence. This was verified by small values of functions F_2 and F_3 for all k, namely, their values did not exceed -2.0. After adding the symbol "a" at position 600, graphs were drawn again (Figure 2b). As it can be seen, in the position close to 600 the function F_2 reaches extremely high value, $F_2(591) = 9.32$, that corresponds to a jump from frame $T1$ to $T2$. Function F_3 still keeps small values for different k. Similar curves also were obtained for many other sequences of genes from *Kegg-46* databank before and after an artificial shift. It proves our assumption that heterogeneity of triplet frequencies can be used as a measure of RF shift.

3.2 Application of Monte Carlo Method to Detect the Threshold F_0

To determine the threshold value F_0, the algorithm described above was applied to the analysis of sequences from a random databank having the same distribution of sequence lengths and triplet frequencies as the databank of genes *Kegg-46*. Sequences of random

bank were obtained from sequences of real bank by shuffling their triplets. The level F_0 was adjusted such that the ratio of number of the sequences with RF shifts in random bank and number of sequences with RF shifts found in *Kegg-46* was rather low. This ratio characterizes the number of false positives and at level $F_0 = 5.0$ equals 5.9%. In further calculations we used the level $F_0 = 5.0$ and also the level $F_0 = 3.75$ that corresponds to the number of false positives 18.0%.

3.3 Detecting RF Shifts in the Sequences from Kegg-46 Databank

We searched RF shifts in the databank of DNA sequences of different biological species *Kegg-46*. The databank contains only sequences of genes. It is supposed that coding of amino acid sequence from DNA sequence was made using the frame $T1$. The sequences with lengths greater than 300 bases were selected. Their number in *Kegg-46* is equal to 2941437. At level $F_0 = 5.0$ we found 140138 sequences with RF shifts. It makes 4.8% from all number of investigated sequences. 81096 (58%) of all shifts corresponded to statistically significant transition from frame $T1$ to frame $T2$. Other cases of shifts were transitions from frame $T1$ to $T3$. For the threshold level $F_0 = 3.75$ a number of 225803 sequences with shifts were revealed.

Some examples of sequences having shifts are shown below. On (Figure 3a) graphs of the functions F_2 and F_3 depending on a position k for the sequence *Bcen_0873* are represented. The sequence *Bcen_0873* belongs to genome of bacterium *Burkholderia cenocepacia* and codes *lytic transclycosylase*. In the figure it is possible to see that the insertion takes place in the position close to 876 because F_2 exceeds the threshold value $F_2 = 11.87$ while F_3 is very small ($F_3 = -14.76$).

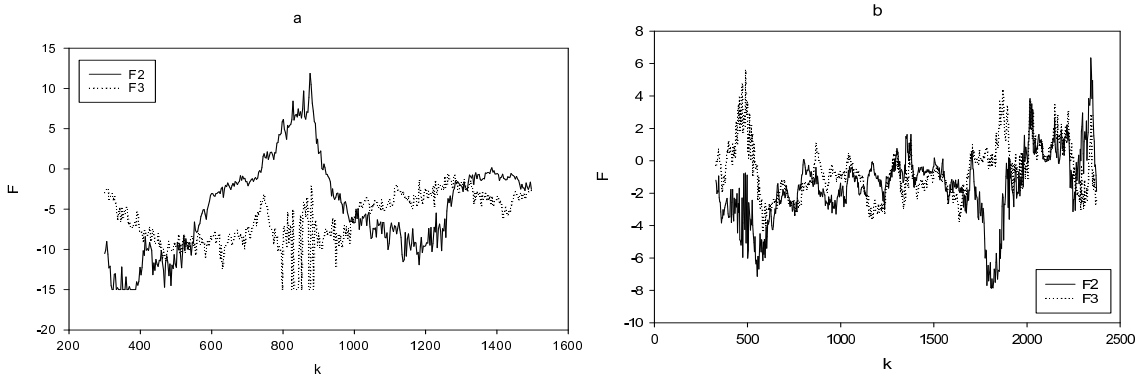

Figure 3: a) RF shift in the sequence Bcen_0873 from Burkholderia cenocepacia, b) RF shift in the sequence 452673 from P.troglodytes genome.

The second example shows the presence of two RF shifts in the sequence. In the Figure 3b, the graphs of F_2, F_3 for the sequence with the identifier 452673 coding *Rho guanine nucleotide exchange factor (GEF) 7* are presented. The sequence belongs to *Pan troglodytes (chimpanzee)* genome. The first shift was observed at coordinate 489, $F_2 = -0.75$, $F_3 = 5.63$. The second shift occurred at position 2343, $F_2 = 6.35$, $F_3 = 2.81$. The peak on the diagram near the position 1900 is not significant ($F_3 = 4.41$). Possibly,

similarity of triplet frequencies in the left and in the right fragments from this coordinate is caused by random factors. However, it is possible to assume that we see traces of ancient RF shift which have been strongly washed away by mutational process.

We classified sequences containing RF shifts according to their description (a field "description") in *Kegg-46*, Table 1. These data correspond to level $F_0 = 3.75$. It has appeared that among the sequences revealed by us 842 sequences have been annotated earlier as the genes having shifts. It makes 41% from total number of genes, annotated as "frameshift". The given result shows that the method developed is capable of revealing shifts which have been revealed earlier by other experimental and mathematical methods. Among the sequences identified by us as sequences with RF shifts there are many pseudo genes – 6073.

The interesting fact is that besides already known cases of shifts our method was able to find the large number of shifts in the genes coding PE-PGRS proteins, translation initiation factors IF-2, protein kinases. The presence of shifts in proteins which belong to the same class suggests that the mutation such as insertion or deletion has occurred in their common ancestor and has been fixed during evolution.

Table 1: Classification of genes with RF shifts based on the description in Kegg-46.

No	Description	Number with shifts	Total Number in Kegg-46	Shifts/ Total (%)
1	Pseudogene	6073	29048	21
2	Zinc finger	3090	7441	42
3	Protein kinase	2174	10042	22
4	ABC transporter related	1011	50550	2
5	Frameshift	842	2050	41
6	Transposase	802	27440	3
7	Lipoprotein	547	17018	3
8	Translation initiation factor IF-2	359	1414	25
9	PE-PGRS	288	584	49
10	Mucin-associated surface protein	249	965	26

4 Discussion

In the present work it has been shown that about 4.6% genes from *Kegg-46* databank contain RF shifts. We think that this number is the lower bound estimate of the total number of genes with shifts. Real number of these sequences should be much greater. Rather small number of shifts we can explain by the fact that for evaluation function I_j (see Equation (1)) rather long fragments of sequence are needed – minimally 150 symbols to the left and to the right from the position of the shift. Thus the shifts inside the fragments smaller than 300 nucleotides and also the pairs of shifts which had less than 150 nucleotides between them have not been revealed by the method developed.

The method performs well for revealing insertions with rather small length. Long insertion can disturb uniformity of triplet distributions in fragments to the left and to the

right from a position of insertion. Therefore, for identification of such cases it is necessary to use other methods.

We have compared results of detecting RF shifts by the method offered in the present work and earlier developed method based on triplet periodicity phase shift (Korotkov and Korotkova, 2010). The method used in given work reveals more cases of shifts (140 thousand) than the method developed earlier (112 thousand) for the sequences which length is greater than 300 bases. The number of false positives was about 6%. Thus we have increased the amount of revealed sequences with shifts by 25%. The advantage of the new approach is estimation of the statistical significance by Monte Carlo method. The statistics Z_j was calculated based on a set of triplets for the specific sequence. Therefore this estimation of statistical significance is more precise than the one determined on the basis of formulas for standard distributions.

What is the evolutionary sense of RF shifts' occurrence in DNA sequences? It is considered that indels of symbols leading to the shifts form one of evolutionary mechanisms of qualitatively new proteins' origination. But in most cases the RF shifts which have been fixed by evolutionary process do not influence the function of the original protein. Apparently, if the mutation does not occur in the active center of protein, then the probability of the protein properties conservation is high. Possibly, the idea of maintaining the stability of a cell during all possible changes in DNA sequence is intrinsic to the structure of a genetic code.

On the other hand, in the case of RF shift occurrence in the coding DNA sequence the probability that protein will not perform its functions increases. It happens due to presence of stop-codons in alternative reading frames. Such a gene with shift passes to a category of pseudogenes. Up to now the biological role of pseudogenes is not clear. Nevertheless, it is possible to tell with confidence that fragments of DNA "staying silent" at present time can be activated at any moment by means of mutation returning a reading frame to its initial state. From this point of view a mutation of RF shift can be considered as some switch between an active and passive state of a gene.

The practical tasks which are possible to be solved by means of the method developed are as follows. Firstly, they include revealing mutations in DNA sequences, namely, the insertions and deletions which lengths are not divisible by 3. On the basis of these data it is possible to restore ancient amino acid sequence of the protein. Also, the method allows detecting sequencing errors in new DNA sequences. Automatic revealing and correction of errors will increase the quality of sequencing process and will allow performing their correct annotation.

Acknowledgements

This work was supported by the Federal Targeted Program "Scientific and scientific-pedagogical personnel of the innovative Russia 2009-2013".

References

Bennetzen, J. L., and Hall, B. D. (1982). Codon selection in yeast. *The Journal of Biological Chemistry*, *257*, 3026-3031.

Boys, R. J., and Henderson, D. A. (2004). A Bayesian approach to DNA sequence segmentation. *Biometrics*, *60*, 573-588.

Braun, J. V., Braun, R. K., and Müller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, *87*, 301-314.

Claverie, J. M. (1993). Detecting frame shifts by amino acid sequence comparison. *Journal of Molecular Biology*, *234*, 1140-1157.

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

Filina, M. V., and Zubkov, A. M. (2008). Exact computation of Pearson statistics distribution and some experimental results. *Austrian Journal of Statistics*, *37*, 129-135.

Korotkov, E. V., and Korotkova, M. A. (2010). Study of the triplet periodicity phase shifts in genes. *Journal of Integrative Bioinformatics*, *7*, 131-142.

Kullback, S. (1959). *Information Theory and Statistics*. New York: John Wiley & Sons.

Posfai, J., and Roberts, R. J. (1992). Finding errors in DNA sequences. *Proceedings of the National Academy of Sciences*, *89*, 4698-4702.

Salem, I. H., Kamoun, F., Louhichi, N., Rouis, S., Mziou, M., Fendri-Kriaa, N., et al. (2010). Mutations in LAMA2 and CAPN3 genes associated with genetic and phenotypic heterogeneities within a single consanguineous family involving both congenital and progressive muscular dystrophies. *Bioscience Reports*, *31*, 125-135.

Schiex, T., Gouzy, J., Moisan, A., and Oliveira, Y. de. (2003). FrameD: A flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Research*, *31*, 3738-3741.

Sprinthall, R. C. (2002). *Basic Statistical Analysis* (7th ed.). Boston: Allyn & Bacon.

Stallmeyer, B., Fenge, H., Nowak-Gottl, U., and Schulze-Bahr, E. (2010). Mutational spectrum in the cardiac transcription factor gene NKX2.5 (CSX) associated with congenital heart disease. *Clinical Genetics*, *78*, 533-540.

Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Michael, L., Richard, L., et al. (2007). *Molecular Biology of the Gene* (6th ed.). San Francisco: Benjamin Cummings.

Wolfe, D. A., and Chen, Y. S. (1990). The change point problem in a multinomial sequence. *Communication in Statistics – Computation and Simulation*, *19*, 603-618.

Authors' address:

Valentina Rudenko, Yulia Suvorova and Eugene Korotkov
Bioinformatics Laboratory
Centre of Bioengineering
Russian Academy of Sciences
Prospect 60-tya Oktyabrya 7/1
Moscow 117312, Russia

E-Mail: genekorotkov@gmail.com