APPLICATION OF A MONTE-CARLO METHOD FOR SEARCH OF THE READING FRAME SHIFTS IN DNA CODING SEQUENCES

V.M.RUDENKO, E.V.KOROTKOV
Centre of Bioengineering RAS
Moscow, RUSSIA
e-mail: v.m.rudenko@gmail.com

Abstract

The Monte Carlo method is used for search of mutations in genes arising by reading frame (RF) shifts. Developed approach is based on comparison of triplet frequencies before and after a place of putative RF shift. We revealed 192456 genes with potential RF shifts in kegg data bank version 46. The received results show, that the Monte Carlo method is more effective for detection the RF shifts in genes than earlier applied approaches.

1 Introduction

Search of the RF shifts in coding DNA sequences is important for understanding of mechanisms of gene evolution because the amino acid sequence coded by gene is completely changed below the shift [1]. If such changed sequence has biological function, it's very interesting to understand what factors were the reason for it. The understanding of these factors can occur as a result of accumulation of some examples of the RF shifts in genes. So the necessary purpose is the development of more perfect mathematical methods for search of RF shifts in genes. In this paper we suggest a new method for search of the RF shifts, based on the assumption that frequencies of nucleotide triplets are homogeneous lengthwise of a gene. Under this condition it is possible to expect that the dissimilarity of triplet frequencies before and after shift position is a suitable measure for identification the RF shift. Application of the Monte Carlo method revealed approximately in 2.5 times more number of genes with RF shifts than the analytical approach.

2 Methods

Define the DNA sequence as $S = \{s(k), k = 1, 2, ..., L\}$ where $s(k)$ is one of the symbol from alphabet $A = \{a, t, g, c\}$, L is the length of the sequence. For simplicity we assume that sequences length is multiple to 3. We consider three possible reading frames $T1, T2$ and $T3$ (fig.1) on the sequence.

Suppose that the RF shift has occurred in a position k (k multiple of 3) of the sequence. Introduce the quantitative characteristic reflecting possibility of presence of the RF shift in a position k of sequence S. For this purpose consider the window of length $2w+2$ which coordinates on the investigated sequence are $(k-w, k+w+2)$. Let's

```
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: Different reading frames of DNA sequence ($ - stop codon)

define the triplet frequencies f_i in a left half of the window with coordinates $(k-w,k)$ and the frequencies of triplets v_i^j in a right half of the window $(k+j,k+w+j)$ for three possible reading frames, here j is a number of the reading frame $j = 1 \ldots 3$. Let's check up the hypothesis of homogeneity of distributions of two sampling using formula [2]:

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

here $N_1 = N_2 = w/3$. If there is no RF shift in position k value I_j has minimum for $j = 1$. In the presence of the shift this minimum will be reached for $j = 2$ or $j = 3$. Values $2I_j$are distributed as χ^2 with 63 degrees of freedom, if the values of every f_i and v_i^j, $i = 1 \ldots 64, j = 1 \ldots 3$ are more or equal to 5. To avoid a problem of small samples we used a Monte Carlo method to estimate a $2I_j$distributions.

Let's consider a sequence fragment $(k - w, k + w)$ where k is a position of the putative shift, k, w multiple of 3. Under these conditions this fragment of sequence corresponds to a reading frame $T1$. We calculated the value of statistics $2I_1$ for the first frame and determine it like a $2I_1^{ob}$. Further we mixed by a random way the triplets of a fragment of sequence $(k-w, k+w)$ and each time calculated the $2I_1$. Then the average value, deviation and statistic $Z_1$for the random variable $2I_1$ were found under formulas:

$$Z_1 = \frac{2I_1^{ob} - 2I_1^{\exp}}{\sigma_1} \quad (2)$$

$$2I_1^{\exp} = \frac{\sum_{i=1}^{M} 2I_1}{M} \quad (3)$$

$$\sigma_1 = \sqrt{\frac{\sum_{i=1}^{M} (2I_1 - 2I_1^{\exp})^2}{M-1}} \quad (4)$$

here M is a number of random shuffling of sequence fragment. Variable Z_1 has approximately standard normal distribution - $N(0.1)$. The value $p_1 = P(N(0,1) > Z_1)$ shows probability of that the divergence between distributions of triplet frequencies to

the left and to the right of k position is caused by random factors. If there is a shift in the position $k$then frequencies of triplets in windows $(k-w,k)$ and $(k+1,k+1+w)$ will considerably differ from each other, Z_1 will exceed the expected value and p_1 will be relatively small. Other cases are not interesting for us. However to assert that the difference of triplets frequencies is caused by RF shift, instead of the other causes. it is necessary to check up the similarly of the distributions of triplets frequencies in a fragment of sequence before putative shift - $(k-w,k)$ and fragments of sequence after shift - $(k+j,k+j+w), j=2,3$. To do this we considered statistics for frames $T2,T3$, determined by formula (1). To estimate the degree of accident coincidence of distributions of triplet frequencies of the right part of considered window with frequencies of the left part we used a Monte Carlo method. We mixed in a random way triplets of the right part of sequence for $T2,T3$ frames and for each random sequence part calculated $2I_j$. Then under the formulas similar to (2-4) were calculated average values, deviations and also Z_2 and Z_3, for frames $T2,T3$respectively. Z_2 and Z_3 have approximately standard normal distribution. If a shift took place in the position k frequencies of triplets in windows $(k-w,k)$ and $(k+j,k+j+w)$ will be similar, I_j is small, and $Z_j$is small too. Let's consider a value p_j=$P(N(0,1)¿Z_j)$. If the RF shift presents in the sequence p_j accepts rather great value. In practice it is more convenient to use as the quantitative characteristic of RF shift values F_2 and F_3:

$$F_2 = -\log(p_1/p_2) \tag{5}$$

$$F_3 = -\log(p_1/p_3) \tag{6}$$

So the sequence under study has the shift in position k if F_2 is greater than some threshold value F_0. It means that the insertion of $1+3n$ or deletion of $2+3n$ nucleotides is possible near k position. Analogically, if $F_3 > F_0$ than the insertion of $2+3n$ or deletion of $1+3n$ nucleotides is possible near k position.

To determine the threshold value F_0 for functions F_j, $j=2,3$ the Monte Carlo method was applied. We took into consideration sequences from random databank. Databank has sequences with the same lengths and triplets distributions as Kegg-46 databank. We scanned real and random databanks with goal to find out the potential shift positions in sequences. The length of a scanning window was 602 ($w=300$) bases. In the issue of scanning the level F_0 equal 5.5 has been chosen. For this threshold a number of random sequences with RF shifts was ~6% from number of RF shifts which have been found for a real kegg-46 databank. The given number characterizes quantity of the false positives.

3 Results

The method of search of potential RF shifts has been realised as a computer program in language C++. During scanning of real sequences the window length 602 (w=300) bases was used. Window moved on 3 bases at each step of the algorithm. The number of iteration M of a Monte-Carlo method was equal to 200. With help of developed program we analyzed the databank of coding DNA sequences kegg-46 which consists of

3318627 sequences. As threshold has been chosen F_0=5.5. For calculations computer cluster was used. We revealed 224839 shifts in 192456 genes. This number makes 5.8 % from total quantity of sequences in the databank. As a result of using as a measure of shift the distinction between triplet frequencies before and after putative shift position and the Monte Carlo approach we could find out approximately in 2.5 times more RF shifts, than earlier applied method based on comparison triplet periodicity [3]. Examples of the real genes having RF shift are shown at fig.2. On fig.2a. functions F_2 and F_3 for sequence ECHS_A3663 from E.coli_HS genome are presented. The protein coding by this sequence is a cell division protein FtsY. From figure we can see that it observed a deletion of symbol in a position nearby 550 bases, F_3=8.75. The second example 2b shows sequence Acid345_1467 of genome Acidobacteria bacterium coded ribonuclease G. In this sequence it has been found two RF shifts in positions 885 and 1800. In the position 885 it is observed insertion, F_2=5.64, in the position 1800 - deletion, F_3=8.12. It is possible to assume that during evolutionary process the changed RF which had been obtained by the way of the first shift has been returned in a starting position by means of the second shift.

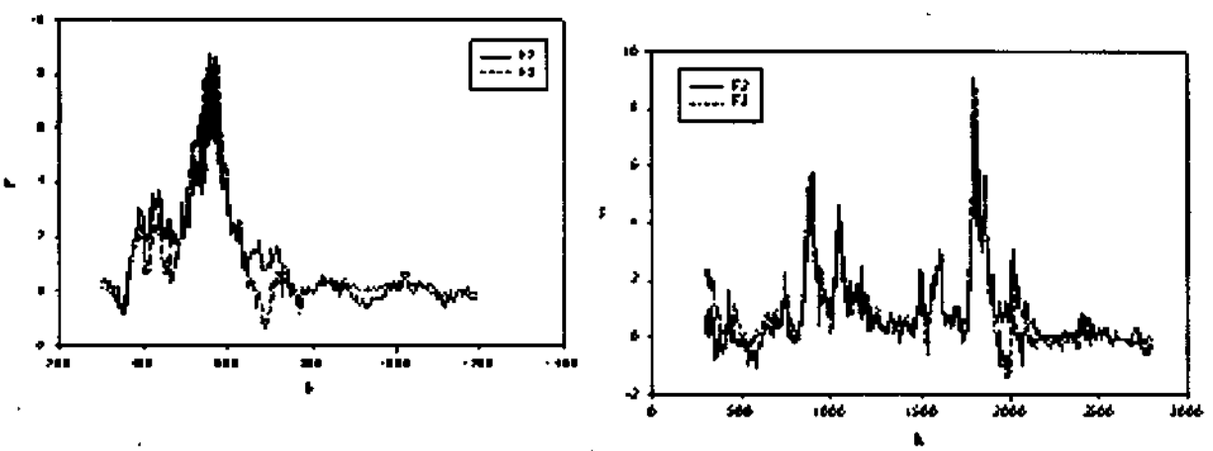

Figure 2. F_2 and $F_3$for sequences: a) ECHS_A3663, b) Acid345_1467

References

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

[2] Kullback S. (1959) Information Theory and Statistics. *New York: Wiley*

[3] Korotkov E.V., Rudenko V.M. (2009) Triplet periodicity phase shift in genes sequences. *Mathematical biology and bioinformatics (RUS).* **4, 2**, pp. 66-80