The database of triplet periodicity change points

Y.M. Suvorova, E.V. Korotkov

Centre of Bioengineering Russian Academy of Sciences, 117312, prospect 60-tya Oktyabrya 7/1, Moscow, Russia, suvorovay@gmail.com

It is well-known that triplet periodicity (TP) is the distinguishing feature of protein coding sequences of the majority of the living organisms [1]. It is also known that in some stage of evolution mutation types such as fusions play an important role in the process of a new protein coding sequence formation [2]. If two genes with different triplet periodicity types were fused then the resulting sequence would consist of two successive parts with different TP types. So one could find a triplet periodicity change point between these parts in the sequence [3]. In order to allow one to study the phenomenon of triplet periodicity change point we collected the triplet periodicity change point events in the database. TPCPDB (Triplet periodicity Change Point Database) is an online database that contains triplet periodicity change points that were found in protein coding sequences of prokaryotic genomes from GenBank.

To study triplet periodicity change points in protein coding sequences we used a method that based on maximum difference of TP between adjacent subsequences. We used a sliding window method to find the position of maximal TP difference in a sequence. To estimate statistical significance of the found change point cases the Monte-Carlo method was used.

Moving a sliding pointer x along a sequence S of length L we considered adjacent regions of length l (from 60 to 600 nt) on the left and the right side from x. To study and compare triplet periodicity on the regions 4×3 frequency matrixes were used. An element of such a matrix is a number of nucleotides of type i (i=1 for 'a', i=2; for 't', i=3, for 'g' and i=4 for 'c'), which is in the position j of a codon (j=1,2,3), in the considered region. In order to eliminate the influence of enrichment of a certain types of nucleotides on the difference we used the following element-wise transformation

$$n_k(i,j) = \frac{m_k(i,j) - lp_k(i,j)}{\sqrt{lp_k(i,j)(1-p_k(i,j))}}$$

i=1,2,3,4; j=1,2,3, and $p_k(i,j) = ((\sum_{i=1}^{4} m_k(i,j)) \cdot (\sum_{j=1}^{3} m_k(i,j))) \Big/ l^2$

In order to take into account possible reading frame shifts near the position x we considered all three reading frames (and corresponding matrixes) after the position [4]. So we compared three matrixes on the right-side subsequence (W_k, k=1,2,3) with one on the left from x (denote it V):

$$d = \min_{k=1,2,3}(D_k(x,l)) = \sum_{i=1}^{4}\sum_{j=1}^{3}\left(\frac{v(i,j)-w_k(i,j)}{\sqrt{2}}\right)^2$$

Moving a sliding pointer x along the sequence we were looking for the position of the maximum of TP difference (d_{max}). To define the statistical significance of the found case the Monte-Carlo method was used. For each considered sequence the set of random sequences (of size N=1000) was generated by trio-shuffling of the sequence S. On this set the mean value and the deviation of d_{max} were determined. And the final value for the sequence was defined as

$$Z = \frac{d_{\max}(S) - \overline{d_{\max}(S)}}{\sqrt{D(d_{\max}(S))}}$$

We included in the database sequences with triplet periodicity change points where the Z value exceeds the threshold $Z > Z_0 = 4.0$ (that corresponds to 5% probability of the first type error).

The current version of the database includes 179 238 records of protein coding sequences with triplet periodicity change points. To access the database one of the following search options can be used: genome sequence ID or name (and if needed the concrete region on the sequence could be specified) and/or gene identification (GenBank). Search for a change points using the description of the protein is also allowed. Results are returned in a table showing information about the change points. The information for each change point record includes: internal change point number; gene identificator; genome name; product description; change point position; window size; Z-value. The database URL: http://victoria.biengi.ac.ru/tpcpdb/.

1. E.N. Trifonov (1998) 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica A*, **249**: 511–516.

2. S. Pasek, J.L. Risler, P. Brezellec (2006) Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins, *Bioinformatics*, **22**: 1418-1423.

3. Y.M. Suvorova, V.M. Rudenko, E.V. Korotkov (2012) Detection change points of triplet periodicity of gene, *Gene*, **491**: 58-64.

4. V. Rudenko, Y. Suvorova, E. Korotkov (2011) Detection of Possible Reading Frame Shifts in Genes Using Triplet Frequencies Homogeneity. *Austrian journal of statistics*, **40**: 137–146.