HIBIT'11

Proceedings of the 6th International Symposium on Health Informatics and Bioinformatics

6TH INTERNATIONAL SYMPOSIUM ON HEALTH INFORMATICS AND BIOINFORMATICS

HIBIT - 2011

hibit11.iyte.edu.tr

HIBIT'11 PROCEEDINGS

This work contains the accepted submissions to the 6th International Symposium on Health Informatics and Bioinformatics, held in Izmir, Turkey from the 2nd to 5th of May 2011. Submission types were full papers, workshops/tutorials, posters, and round table discussions

SPLICING OF THE TRIPLET PERIODICITY IN GENES FROM DIFFERENT SPECIES

Yulia M.Suvorova[1], *Eugene V. Korotkov*[1,2]

[1]Bioinfomatics laboratory, Centre of Bioengineering Russian Academy of Sciences, 117312, Prospect 60-tya Oktyabrya, 7/1;
[2]Department of Applied mathematics, Moscow Physical Engineering Institute, 115409, Russia, Moscow, Kashirskoe Shosse, 31, phone:+7-499-135-2161; fax:+7-499-135-0571; email: suvorovay@gmail.com, genekorotkov@gmail.com

ABSTRACT

We found genes which contain more than one type of the triplet periodicity (TP) during their sequences. We say that these genes contain the triplet periodicity splicing (TPS). The aim of the work is to study such genes and try to understand the nature of the (TPS) phenomenon. The mathematical algorithm for detection of TP splicing of nucleotide sequences has been developed. Gene sequences from KEGG-48 databank were analyzed with a purpose of searching for genes with a TPS. The presence of a triplet periodicity splicing has been shown for 311221 genes (~8% from the total number of genes in KEGG-48). We showed that the repetitive or low-complexity sequences are not cause of the TPS. We suppose that the TPS cases may indicate fusion of genes or domains with different types of TP. The relationship between TPS and the fusion events in genes is discussed.

1. INTRODUCTION

The main mechanisms of gene evolution as it now well established are duplication, recombination and sequence divergence. Currently divergence both nucleotide and amino acid sequences is quite well studied. But at the same time little is known for certain about the phenomenon of fusion of domains and genes in a new sequence. One of the main reasons is a difficulty of detection of such events. Nowadays fusion detection is carried out by experiment or by similarity searching between genes or protein sequences. The main limitation of this method is that evolution process could change original sequences so much that sequence comparison cannot reliably detect their similarity with the parts of fused gene. It is possible too that the original genes (or domains) are not sequenced yet and those not present in databases. So the importance of pure mathematical method for studying fusion and fission events based on the sequence structure only is clearly.

Triplet periodicity (TP) of DNA coding sequences is a common property of all known living organisms ([2], [7]), and it is associated with the gene reading frame. The reasons of relation between RF and TP are the genetic code structure (which is almost identical for prokaryotes end eukaryotes), saturation of proteins with certain amino acids, and uneven using of the synonymous codons [2]. One may suppose that the difference on the TP of the adjacent segments of the gene could be used as the fusion events marker. Currently used methods of TP searching rely on the regularities of nucleotide preferences in different positions of triplets of gene sequence. Fourier transformation, hidden Markov models and other statistical methods are used for revealing TP. The methods applied in this work were directed on searching for coding regions of genes and separating them from the noncoding regions [3]. The method of information decomposition has been proposed later for searching TP [7]. It allows introducing the definition of a class of triplet periodicity as a matrix with dimensions 4x3 ([3], [7], [8]). It was previously shown that there are about 2500 classes of the triplet periodicity in coding sequences. These classes can vary greatly. This means that if two fused genes or domains had different types of TP, then the matching point can be relatively easy to detect. This point has the maximum of a difference between two matrixes of TP on the left and right side of the point of fusion. Such points are supposed to be long time detectible, because the TP of DNA sequence is difficult to be changed by a few number of DNA substitutions [7]. So, the presence of a splicing of the TP [TPS] in gene sequence may serve as an indicator of fusion event in the analyzed gene. Two problems are being solved in the present work. Firstly, we would like to develop a TPS revealing method based on the sequence structure without any additional data and then to find all genes with TPS in KEGG database. Secondly, we would like to test an assumption the relationship between the genes with TPS and the fused genes (genes that actually consist of two or more genes).

2. ALGORITHMS AND METHODS

2.1 Conditions for existing of the TPS in gene

Consider a coding nucleotide sequence $S = \{s(k), k = 1, 2, ..., L\}$, where each base $s(k)$ is selected from the alphabet $A=\{a,t,c,g\}$, L is the length of S, and it is divisible by 3.Let us introduce three RF in S and designate them as T_1, T_2 and T_3. The base $s(1)$ of a sequence S represents the first, second and third base of codon for RF T_1, T_2 and T_3, correspondingly. The RF T_1 really exists in the sequence while RFs T_2 and T_3 are the hypothetical ones.

Then let us introduce three matrices of the TP $M_1(i_1, i_2)$, $M_2(i_1, i_2)$ and $M_3(i_1, i_2)$, which are the matrices of TP calculated for RF T_1, T_2 and T_3 for the region of a sequence S from the base i_1 to the base i_2. Let us denote this subsequence as $S(i_1, i_2)$. The elements of matrices $m_1(i,j)$, $m_2(i,j)$ and $m_3(i,j)$

show the number of the bases of a type i in a sequence S (i=1 for a, i=2 for t, i=3 for c, i=4 for g,) which are in j position of the codon (j equals 1,2 or 3) for RF T_1, T_2 and T_3, respectively [7].

We should define the conditions which show us the existence of TP splicing after the base $s(x)$ in a sequence S. Firstly, TP should exist in the sequence S. Conditions of TP presence and quantitative measure for revealing TP in a sequence S, or in any its subsequence, are described in detail in work [8]. Secondly, we should introduce the quantitative measure of a difference between TP matrices. Let us introduce some function U and let us assume that two TP matrices are different to each other if $U \geq U_0$. Otherwise we consider two TP matrices as similar matrices. Thus we define the condition of TPS presence in the position x as:

$$U\{M_1(1,\ x),\ M_i(x+1,\ L)\} > U_0 (i=1,2,3) \qquad (1)$$

It means that $M_l(1,\ x)$ matrix differs from the TP matrixes after position x. The task is to select the function of U and the level of U_0 for finding of the TPS in genes. The detailed description of the function U is given in the next section.

2.2 Searching for the TPS in the sequence

Let x shows the position of a base $s(x)$ in the sequence S and let x be chosen as L_1+3n, where n=0,1,2,3,…(L-L_1)/3, and where L_1 is divisible by 3 and 60≤ L_1≤600. Consider the subsequence $S(x\text{-}L_l+l.\ x)$. For this subsequence we calculate the matrix of TP $M_1(x\text{-}L_1+l,\ x)$ for RF of T_1 in S. The subsequences $S(x+1.\ x+L_1)$. $S(x+2.\ x+L_1+1)$ and $S(x+3.\ x+L_1+2)$ are also considered, and for these subsequences we calculate the TP matrices $M_1(x+1,\ x+L_1)$. $M_2(x+2.\ x+L_1+1)$ and $M_3(x+3.\ x+L_1+2)$ for RF of T_1, T_2 and T_3, respectively. If a TPS occurs just after the position x, then the matrix $M_l(x\text{-}L_1+1.\ x)$ is supposed to be different from $M_l(x+1.\ x+L_l)$. $M_2(x+2.\ x+L_1+1)$ and $M_3(x+3.\ x+L_1+2)$ matrices. Then for each of the four TP matrices another matrix was calculated. Its elements were the arguments of the normal distribution. Each element of such a matrix was calculated using the following formula:

$$n(i,j) = \frac{m(i,j) - L_1 p(i,j)}{\sqrt{L_1 p(i,j)(1-p(i,j))}} \qquad (2)$$

where $p(i,j) = \frac{x(i)y(j)}{L_1^2}$, $m(i,j)$ is an element of M_1, M_2 or M_3, $n(i,j)$ - normally distributed value. As a result, we obtain for each of the matrices $M_1(x\text{-}L_l+1,\ x)$, $M_l(x+1,\ x+L_1)$, $M_2(x+2,\ x+L_1+1)$ and $M_3(x+3,\ x+L_1+2)$ the matrices V_1, W_1, W_2 and W_3 The difference between the matrix V_1 and of W_k matrix (k=1,2,3) is defined as:

$$U = D(1,k) = \sum_{i=1}^{4}\sum_{j=1}^{3}\left(\frac{v_1(i,j) - w_k(i,j)}{\sqrt{2}}\right)^2 (k=1,2,3) \qquad (3)$$

$D(1,k)$ is distributed as χ^2 with 6 degrees of freedom if the matrices calculated for random sequences are being compared [34]. Following the conditions (1) we chose minimum of three value $D(1,k)$, k=1,2,3. Let designate it as $D_{\min}$ The probability is $P(D_{\min} \geq x) = P^3(\chi^2(6) \geq x)$. The final measure is defined as $F = -log_{10}P(D_{\min} \geq x) = -3log_{10}(P(\chi^2(6) \geq x))$. Then introduce the level for F as $F_0 = -3log_{10}(P(\chi^2(6) \geq U_0))$.

We mowed a sliding pointer x along the sequence, for each position we have varied the length of considering subsequences L_1 within the interval from 60 to 600, with a step of variation equal to 3 bases, we searched for L_1 that maximized measure F for the position. We select the point (and corresponding L_1) at which F reaches its maximum value and then checked its statistical significance. If this value is greater than some threshold F_0, then one can consider that S has a TPS at the position x.

2.3 Monte-Carlo simulation

Let's N_0 be the total number of genes in KEGG-48 (N_0=4013150). Let us consider two alternative hypotheses. Under the first one (null hypothesis) all detected cases with $F > F_0$ are the results of some random factors. Opposite, if alternative hypothesis is true then it is not about the random factors but some other reasons. To validate this statement we performed a series of simulation studies by the Monte-Carlo method.

We generated a set of random sequences with the same length and TP level as real genes. The databank of random sequences was made by the way of mixing up an every gene sequence. It allows keeping the same distribution gene lengths and same base composition of genes as in KEGG databank. We divided the gene sequence into three subsequences for keeping the TP in a random gene sequence on the original level. The first subsequence (denoted as C_1) was obtained from a gene sequence by choosing bases which were at the positions equal to i=1+3n. The second and the third subsequences C_2 and C_3 were created by choosing bases which were at the positions i=2+3n and i=3+3n, n=0,1,2…, L/3-1. Then by mixing each of C_1, C_2 and C_3 corresponded random sequences R_1. R_2 and R_3 were obtained. And then there were produced the random sequence R which had R_1 sequence at the positions i=1+3n, R_2 sequence at the positions i=2+3n and R_3 sequence at the positions i=3+3n. i=0,1,2,…, L/3-1. The length of a sequence R was equal to L and it has the same base composition as a gene sequence. We repeated this procedure for all genes from KEGG.

The R-sequences are supposed to be TPS free because the existing TP was shuffled throughout the length of the gene. Denote the number of genes with detected $F \geq F_0$, in KEGG database as N_1, and the number of these genes among the random sequences as N_2. Probability $\alpha = N_2/N_1$ – is the significance level of our test. It is a chance that some random factors are the reason of the fusion events in KEGG genes. In our work we defined the level of F (F_0) that provides us probability α=5%, with N_1 equal to 311221 and N_2 equal to 13306. It allows us to reject the null hypothesis with probability 95% for every found event.

3. RESULTS

3.1 Searching for the genes with TPS in KEGG database

We have analyzed 4013150 genes from the KEGG/Genes (release 48) [6] databank which contains only coding sequences without introns. By the method described above we tested every sequence (denoted S) from this database. The total number of genes with $F>F_0$ was N_1= 311221 (number of false positives was less than 5%). Genes with a single TPS constituted up to 90% from the total number of genes with a TPS. Remaining 10% genes contained more than one case of TPS.

To illustrate the method performance for revealing TPS events, we want to give some examples. Firstly, we studied the possibility of creation an artificial sequence with TPS by the joining subsequences from two different real genes. To create it we split PD1767 (DNA topoisomerase I) gene from X.fastidiosa and XAC4270 (glycerol-3-phosphate acyltransferase) from X.axonopodis genome into two parts, then we took the first half of PD1767 and the second half of XAC4270 and concatenated them. In a new sequence the coordinate of the connection is located between 1223rd and 1224th bases. For this sequence we calculated $D(1,k)$ (k=1,2,and 3) for each position of x. For clarity, values were reduced to a normal distributed value (denoted $Z(k)$). Fig. 1 plots $Z(k)$ as a function of position x in the sequence. It shows that there is a maximum of Z(1) near the base 1224, when the values of Z(2) and Z(3) are relatively large at any position x. This example shows that the mathematical approach developed allows us to find the TPS in the artificial sequences.

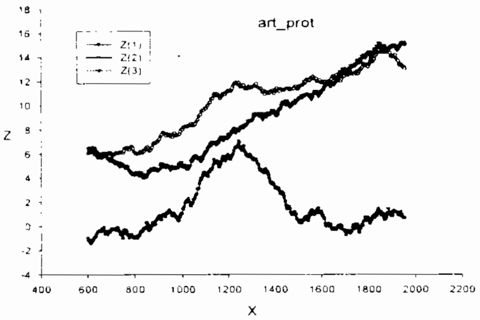

Fig. 1 Data for argument of normal distribution $Z(k)$ (k=1,2,3), x is the position in gene (L_1=600nt) for the artificial TPS gene.

The second example is a real gene with a single TPS (Fig. 2). The gene is ECP_0691 (the N-acetylglucosamine-specific IIA component) from E.coli_336 genome. As it can be seen in Fig. 2, the gene has maximum of $Z(1)$ in position 1101. It means that TP after 1101st base is different relatively to the TP existing in gene from the 1st to 1101st base.

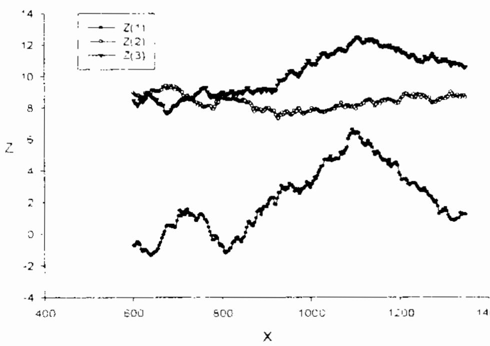

Fig. 2 Plot $Z(k)$, (k=1,2 и 3, L_1=600 bases) for ECP_0691. TPS point is about 1100 base.

The last example is the gene CFF8240_1417 (serine protease) from the C.fetus genome. As it can be seen from the Fig 3, this gene contains no less than two TPS events. It is possible to select two positions in gene sequences: 2100 and 2640, at these points TPS events are well expressed.

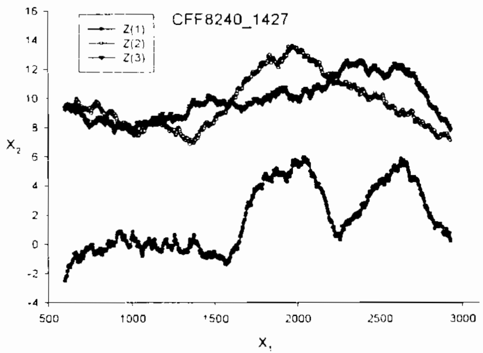

Fig. 3 Data for $D(1,k)$, (k=1,2 и 3) transformed to the argument of normal distribution $Z(k)$, where x is the position in gene sequence. L_1=600. Picture shows the situation for the gene CFF8240_1417 which has two TPS.

3.2 Further analysis of revealed TPS cases

Let us consider the genes in which we have revealed the TPS events. Let label as x_0 the position in which the TPS have occurred. If the splicing of two DNA regions with different TP took place not so long ago, then the alternative separated versions of this gene may remain in other species. Thus there are should exist single genes that correspond only to the subsequence to the left or only to the subsequence to the right of the position x_0.

To detect cases of such a similarity, we compared amino acids sequences encoded by the genes with TPS with amino acid sequences from the Swiss-prot databank. Sequence comparisons were performed by Blastx program [1]. The E-value for Blastx was selected to be equal to 10^{-5} that gives about 3,1 random similarities for comparison of 311221 of random amino acid sequences with Swiss-prot. This number shows that the fraction of random similarities in total number of found similarities is insignificant. As a result of comparisons we identified proper alignments for 131323 cases from all genes with TPS. There are 54406 cases with similarity only before the x_0 coordinate, 60333 genes have such similarity only from the x_0 position to the end of the gene. And 16584 genes have proper similarity (to the different sequences) for both parts – from the beginning to x_0 and from x_0 to the end. We did not find any proper similarity for the other genes with predicted TPS.

It is known that there are a lot of repeat regions in amino acid sequences, the main types of repeats are tandem repeats (AKRAKRAKRAKR), interspersed repeats (AKRAKRTLAKRAKR), cryptic repeats (MMMMM) [13]. All these types may affect the TP level of the corresponding DNA subsequence. If for a considering position x repeat region occupies a significant part of the subsequence $S(x$-

$L+1;\ x$) or of the subsequence $S(x+1;\ x+L)$, then it is likely to see a great difference in TP types of these subsequences, and thereby a high level of the measure F for the point. Thus we wanted to evaluate the percentage of detected cases of TPS due to the presence of repeats in one of the subsequences.

To evaluate repeat affection on the results we create a sample of 10830 sequences from all revealed genes with TPS. The sequences with TPS for the sample were randomly selected from all genomes and a number of included in the sample genes from some genome were proportional to the total number of TPS genes in the given genome. The sample genes were translated into AA sequences according to an existing reading frame and were processed by the GBA program [12], which could detect all previously described repeats, with default parameters. To find a significant level of the GBM p-value we create a control sample, by the way mixing genes (as it was done before in Monte-Carlo simulations, see 2.4) from the main sample and further translation into protein sequences (stop-codons were replaced by a random AA symbol). The control sample was also treated with the GBA. A comparison of results of processing the main and control samples gives a threshold of p-value $P_0=e^{-39}$. With p-value $\leq P_0$ ratio of the number of found repeats is about 5%.

Then we searched for repeats in the subsequences that give maximum value of F (S(x-$L+1;\ x$) and $S(x+1;\ x+L)$, x is the TPS point) using following conditions: 1) p-value$\leq P_0$;. 2) $l \leq L/3$, where l is a summary length of repeats in the subsequence, L - the subsequence length. Ander these conditions we got 750 (6.92%) sequences contain such a repeat region in at least onc of the subsequences forming TPS. The standard error of a reported proportion is $S=\sqrt{p(1-p)/n}=0,0024$. Thus, the proportion of repeats among all detected cases of TPS is (6,92±0,24)%. We also determined the proportion of repeats with p-$value\leq 1e$-5, it was (19,78±0,38)%.

4. DISCUSSION

In the present work we have revealed 311221 genes in which we may suppose the existence of TP splicing, which constitutes ~8% from the total number of the analyzed genes. We suppose that the fraction of genes with a TPS may really be much greater. The reasons why we have not revealed all the TPS are the following. Firstly, we use the relatively high level of TP for sure revealing of the TPS. The study of the TPS for the lower levels of TP probably will find more genes with a $F\geq F_0$ position. We believe that in this work we have found only the lower level of really existing number of TPS events. Note that our approach does not require any additional data and it is based on the gene sequences only.

It is also important to consider an issue whether TPS would always indicate a gene or domain splicing in a sequence. In principle, this may not be claimed with a 100% probability since there is always some small possibility (~5% for F_0) that the TPS is caused by purely random factors or that the TPS is caused by the interchange of alpha-helixes and beta-layers in the structures like αα, ßß, αß and ßα in a protein, where α is an alpha-helix, ß is a beta-layer, or by some other secondary or tertiary protein structures. However, if the last statement was true, we would observe a much greater number of TPS events.

The gene recombination and gene fusion events and transposon insertions could be the reasons of the presence of TPS in genes. In this case parts of genes having different TP could be joined in other sequence. The substantial part of previously revealed fusion events are the cases of joining functionally similar or interacting genes (proteins) ([10], [11], [9], [EIK99]). And if it implies the TP similarity then it may prevent a determination of these cases by the way of TP comparisons. However, in many cases of fused genes there is a maximum of the function F near the region of joining, but its value is smaller than F_0. And in some cases we were able to see significant TPS in genes labeled as "fused" in literature. Example of long known ([EIK99], [11]) fusion is B0879 gene (macrolide transporter subunits of ABC superfamily, ATP-binding component/ membrane component) from the E.coli K-12 genome. The fusion point in the gene is about 630th base and using our method we obtain maximum value $Z(1)$=4,2 near the 630th base in the gene.

To evaluate the percent of TP combinations that we could distinguish we performed pairwise comparisons of existing TP types. Previously a classification of TP found in different genes reveals 2520 various types (classes of different size) [5] where each TP class is defined by a corresponding triple matrix M. It permits to estimate a total number of TPS which could be revealed by the method. This procedure is a modeling of arising of a new gene by the way of random concatenating parts from different genes.

We found that only β=53% from the total number of TP matrixes combinations (we took into account a TP class size) may be revealed by the method. The β value shows that only a part of TPS cases can be found because some of TP matrixes are too similar to find the difference between them using our algorithm. So, n/β estimates the true number of TPS that may exist in genes from the KEGG databank. This value shows that ~580 thousands of genes really may contain at least one TPS. All this calculations are correct under an assumption about equal chance of all TP types to be fused. But as it was mentioned above, there is another trend exists during the evolution and genes that are fused more frequently encoded functionally similar or interacting proteins. Therefore a number of fused genes may be even greater than the last estimation. And the value of 580 thousands genes may be not a limit of all fused genes really existing in different genomes.

REFERENCES

[1] S.F.Altschul, W.Gish, W. Miller, E.W. Myers and D.J. Lipman. Basic Local Alignment Search Tool. *Mol. Biol.*, v.215, pp.403-410. 1990.

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

[3] I. Bernaola-Galvan, I. Grosse, J.L. Carpena, R. Oliver, Roman-Roldan and H.E. Stanley. Finding Borders between Coding and Noncoding DNA Regions by an Entropic Segmentation Method. *Phys.Rev. Letters*. v.85, pp.1342-1345, 2000.

[4] A.J. Enright, I. Illopoulos, N.C. Kyrpldes and C.A. Ouzounis. Protein Interaction Maps for Complete Genomes Based on Gene Fusion Events. *Nature*, v.402, pp.86-90, 1999.

[5] F.E. Frenkel and E.V. Korotkov Classification Analysis of Triplet Periodicity in Protein-Coding Regions of Genes. *Gene*. v.421, pp.52-60, 2008.

[6] M.Kanehisa. and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* v.28, pp.27-30, 2008.

[7] E.V. Korotkov M.A. Korotkova and N.A. Kudryashov Information Decomposition Method to Analyze Symbolical Sequences. *Physics Lett.A*. v.312, pp.198-210, 2003.

[8] E.V. Korotkov and M.A. Korotkova. Study of the Triplet Periodicity Phase shifts in Genes *Journal of Integrative Bioinformatics*. v. 7 pp.131-142, 2010.

[9] E.M. Marcotte M. Pellegrini, HL Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg D. Detecting Detecting Proetin-Protein Interactions from Genome Sequences. *Science*, v.285, pp.751-753, 1999.

[10] S.Pasek, J.L. Risler and P.Brezellec. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics*. v.22, pp.1418-1423, 2006.

[11] M.Serres amd M. Riley. Gene fusions and gene duplications: relevance to genomic annotation and functional analysis. *BMC Genomics*. v.6 pp.33, 2005.

[12] L. Xuehui and T. Kahveci. A Novel Algorithm for Identifying Low-comlexity Regions in a Protein Sequence. *Bioinformatics*. v.22, pp.2980-2987, 2006.

[13] J. Wooton and S. Federhen. Analysis of compositionally biased regions in sequence databases. Methods in Enzymol., v.266, pp.554-557, 1996.