Identification of DNA Motif with Mutation

Shu, Jian-Jun

2015

Shu, J.-J. (2015). Identification of DNA Motif with Mutation. Procedia Computer Science, 51, 602-609.

https://hdl.handle.net/10356/81178

https://doi.org/10.1016/j.procs.2015.05.328

Procedia Computer Science

Volume 51, 2015, Pages 602–609

ICCS 2015 International Conference On Computational Science

Identification of DNA Motif with Mutation

Jian-Jun SHU

School of Mechanical & Aerospace Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798

mjjshu@ntu.edu.sg

Abstract

The conventional way of identifying possible motif sequences in a DNA strand is to use representative scalar weight matrix for searching good match substring alignments. However, this approach, solely based on match alignment information, is susceptible to a high number of ambiguous sites or false positives if the motif sequences are not well conserved. A significant amount of time is then required to verify these sites for the suggested motifs. Hence in this paper, the use of mismatch alignment information in addition to match alignment information for DNA motif searching is proposed. The objective is to reduce the number of ambiguous false positives encountered in the DNA motif searching, thereby making the process more efficient for biologists to use.

Keywords: DNA coding; scoring matrix; sequence analysis

1 Introduction

A DNA strand can be divided into the expressed regions (exons) and intragenic regions (introns). Exons are the substrings of the DNA sequence that contain information about genes. The introns are said to contain 'junk' sequences that do not code for any genes. However it is known that introns contain vital information as tags for ribosomal RNA to identify start and stop transcription sites. Some examples of this information include binding, donor, acceptor sites and TATA boxes, which are known as motifs. The motifs that perform the same function, such as binding sites, are highly conserved and can be represented by a sequence profile or weight matrix. During motif searching, an alignment score is calculated by adding up the individual score assigned for each position along the length of the weight matrix. The score for each position is a result of a direct substitution of base type weightage at position.

The identification of DNA binding sites for transcription factors (motifs) is important for a complete understanding of co-regulation of gene expression, but still remains to be quite challenging to achieve. Two approaches dominate motif-finding algorithms: (1) the word-based way [1-3] that relies on exhaustive enumeration or counting frequencies and (2) the probabilistic way [4-6] that relies on optimizing a scalar-based scoring matrix [7,8], which is visualized conveniently by a sequence

 Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.328

logo. However, both ways suffer from the problem of producing a high number of spurious sites, due to the sole consideration of match alignment information.

In this paper, a novel scoring method by introducing the vector representation of DNA sequences into weight matrix is introduced. In the past, two-, three-dimensional or vector representations of DNA sequences have been proposed. These representations provide a visual way of classifying genes and other motifs through DNA walks [9]. In the vector representation of DNA sequences, the four bases are placed at an equal distance from each other in a three-dimensional space. This is possible by placing each point on the vertex of a tetrahedron. Any point within the tetrahedron represents the different combinations of each base type. Therefore the weight distribution for each position can be replaced a point in the tetrahedron in space using a three-dimensional coordinate. This point is unique for the different weightage distributions of the bases. The advantages of this method are twofold: First, instead of using four numbers to represent the weightage distribution at each position, this can be reduced to three in terms of the three-dimensional coordinate; Secondly, the use of mathematical operators, namely the dot and cross products, can be adopted to describe the weightage distribution in the three-dimensional coordinate and served as the basis of measuring the quantity of match and mismatch alignment information for each alignment position.

A case study of identifying binding sites shows that, by using mismatch alignment information, a substring with the best alignment can be selected from a group of ambiguous ones and the number of false positives can be filtered down.

2 Methods

2.1 Scalar Representation of DNA Base Code

In scalar scoring scheme, each base (A, T, G and C) of a DNA sequence is represented by a set of numbers, such as 1, 2, 3 and 4. The problem with this representation is that of unequal weightage assigned to each base [10]. In order to assign the same weightage for each base, an indicator sequence using binary numbers was proposed [11]. However, DNA base coded using binary numbers are not suitable for comparing sequences. Complex and hypercomplex numbers have then been proposed as a better representation utilizing imaginary domain [12,13].

2.2 Vector Representation of DNA Base Code

The coordinates that are selected for vector representation are shown in Figure 1 [11]. The four bases are represented by A, T, G and C with O being the origin. The significance of the point O is such that the distance from each of the vertex is equal to 1, $OA = OT = OG = OC = 1$.

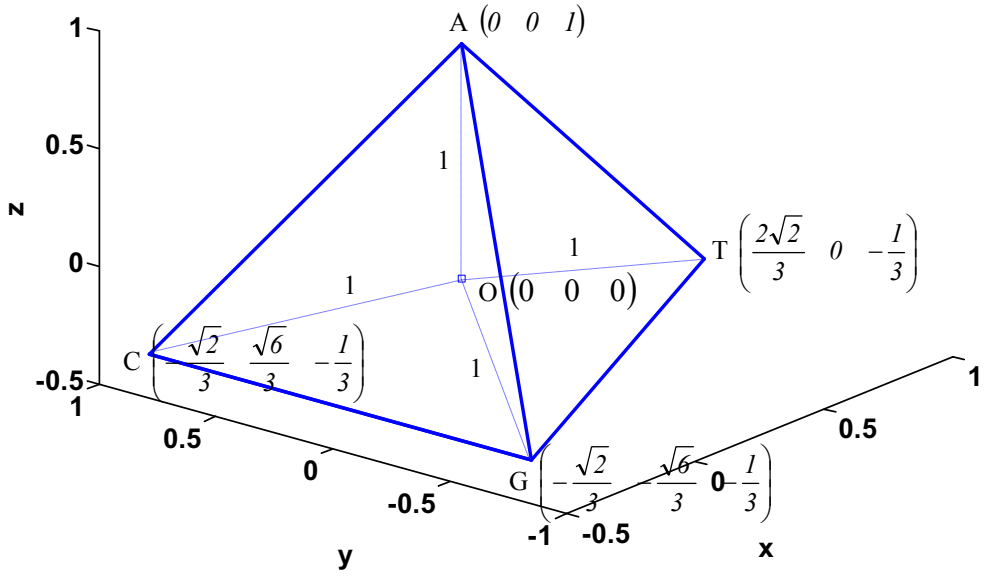

Figure 1: A tetrahedron with assigned bases in x, y and z-axes

2.3 Vector Representation of DNA Motif

The conventional scalar weight matrix [7,8] is converted to a vector weight matrix by the following: For the given weightage of $p_b \in [0,1]$ of base type $b \in \{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\}$, the vector weight matrix is represented by a point at the coordinate (x, y, z)

$$x = \sqrt{(p_{\mathrm{A}}\mathrm{A}_x)^2 + (p_{\mathrm{C}}\mathrm{C}_x)^2 + (p_{\mathrm{G}}\mathrm{G}_x)^2 + (p_{\mathrm{T}}\mathrm{T}_x)^2} \tag{1a}$$

$$y = \sqrt{(p_{\mathrm{A}}\mathrm{A}_y)^2 + (p_{\mathrm{C}}\mathrm{C}_y)^2 + (p_{\mathrm{G}}\mathrm{G}_y)^2 + (p_{\mathrm{T}}\mathrm{T}_y)^2} \tag{1b}$$

$$z = \sqrt{(p_{\mathrm{A}}\mathrm{A}_z)^2 + (p_{\mathrm{C}}\mathrm{C}_z)^2 + (p_{\mathrm{G}}\mathrm{G}_z)^2 + (p_{\mathrm{T}}\mathrm{T}_z)^2}\,, \tag{1c}$$

where $\vec{b} = (b_x, b_y, b_z)$ denote the x, y and z-coordinates of base type $b \in \{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\}$ respectively.

2.4 Scalar Scoring – Match Alignment Information

The distance between two points in the vector weight matrix represents the degree of the similarity between them. The nearer are two points, the better is the match. This can be computed by using the dot product. For the given coordinates (x_n^s, y_n^s, z_n^s) of a DNA substring and (x_n^m, y_n^m, z_n^m) of a motif at the position n with the same length N, match alignment information is the value of the dot product of their coordinates as follows:

$$\text{Score of match alignment information} = \left| \sum_{n=1}^{N} \left[(x_n^s, y_n^s, z_n^s) \bullet (x_n^m, y_n^m, z_n^m) \right] \right|. \tag{2}$$

The resultant dot product gives a scalar. The larger the number, the greater the similarity between the DNA substring and the motif.

2.5 Vector Scoring – Mismatch Alignment Information

In this paper, a novel approach is proposed to analyze the type of mismatch contributing to the lower match alignment information. It can be used to decide whether one site has a greater chance of containing the motif over the others although they may have the same match alignment information. Mismatch alignment information can generally be classified into three types. They are transitional, complementary and transversal mismatches.

2.5.1 Types of Mismatch

Among the four bases forming the skeleton of a DNA sequence, purine bases A and G are bigger in size as compared with pyrimidine bases T and C. A bigger base pairs up with a smaller base in a manner of A to T and G to C in DNA. Transitional mismatch occurs when one base is replaced by another with the same size, *i.e.*, A by G or C by T and *vice versa*. Because the replacement is of the same size, the DNA structure is not unduly affected. Hence, transitional mismatch is considered relatively acceptable.

On the other hand, when one base is replaced by another with a different size but pairs up in DNA, complementary mismatch occurs, *i.e.*, A by T or G by C and *vice versa*.

The third kind of mismatch occurs when one base is replaced by another with a different size and does not pair up in DNA. Among three types of mismatch, this type is considered the most significant and is known as transversal mismatch, *i.e.*, A by C or T by G and *vice versa*. An alignment with a high number of transversal mismatches generally does not contain motif.

2.5.2 Algorithms

For the given coordinates $\left(x_n^s, y_n^s, z_n^s\right)$ of a DNA substring and $\left(x_n^m, y_n^m, z_n^m\right)$ of a motif at the position n with the same length N, mismatch alignment information is calculated by taking the cross product of their coordinates as follows:

$$\text{Score of mismatch alignment information} = \left\| \sum_{n=1}^{N} \left[\left(x_n^s, y_n^s, z_n^s\right) \times \left(x_n^m, y_n^m, z_n^m\right) \right] \right\|. \tag{3}$$

The resultant cross product gives a vector whose three components contain the following information:
Score of transitional mismatch alignment information

$$= \left| \sum_{n=1}^{N} \left[\left(x_n^s, y_n^s, z_n^s\right) \times \left(x_n^m, y_n^m, z_n^m\right) \right] \bullet \left(\vec{\mathrm{A}} \times \vec{\mathrm{G}} + \vec{\mathrm{C}} \times \vec{\mathrm{T}}\right) \right|; \tag{4a}$$

Score of complementary mismatch alignment information

$$= \left| \sum_{n=1}^{N} \left[\left(x_n^s, y_n^s, z_n^s\right) \times \left(x_n^m, y_n^m, z_n^m\right) \right] \bullet \left(\vec{\mathrm{A}} \times \vec{\mathrm{T}} + \vec{\mathrm{G}} \times \vec{\mathrm{C}}\right) \right|; \tag{4b}$$

Score of transversal mismatch alignment information

$$= \left| \sum_{n=1}^{N} \left[\left(x_n^s, y_n^s, z_n^s\right) \times \left(x_n^m, y_n^m, z_n^m\right) \right] \bullet \left(\vec{\mathrm{A}} \times \vec{\mathrm{C}} + \vec{\mathrm{T}} \times \vec{\mathrm{G}}\right) \right|. \tag{4c}$$

Here the respective mismatch vectors are shown in Table 1.

Table 1: Mismatch vectors

Type of mismatch	Pairs	Coordinates x	y	z
Transition	$\vec{A}\times\vec{G}$	$\frac{\sqrt{6}}{3}$	$-\frac{\sqrt{2}}{3}$	0
	$\vec{C}\times\vec{T}$	$-\frac{\sqrt{6}}{9}$	$-\frac{\sqrt{2}}{3}$	$-\frac{4\sqrt{3}}{9}$
Complement	$\vec{A}\times\vec{T}$	0	$\frac{2\sqrt{2}}{3}$	0
	$\vec{G}\times\vec{C}$	$\frac{2\sqrt{6}}{9}$	0	$-\frac{4\sqrt{3}}{9}$
Transversion	$\vec{A}\times\vec{C}$	$-\frac{\sqrt{6}}{3}$	$-\frac{\sqrt{2}}{3}$	0
	$\vec{T}\times\vec{G}$	$-\frac{\sqrt{6}}{9}$	$\frac{\sqrt{2}}{3}$	$-\frac{4\sqrt{3}}{9}$

3 Results and Discussion

3.1 Case Study of TATA Box in *Homo Sapiens H4/g* Gene

A TATA box is a DNA sequence found in the promoter region of most genes in eukaryotes. It is the binding site of either transcription factor or histone, and involved in the process of transcription by RNA polymerase. Histone acts as a spool around which DNA winds and there are five major families, namely H1, H2A, H2B, H3 and H4. H1 is known as the linker histone, while H2A, H2B, H3 and H4 are known as the core histones. They are involved in the different stages of DNA packing. In this case study, it is interesting to identify TATA box in *Homo sapiens H4*/*g* gene. The *H4*/*g* gene is responsible for producing H4 histone. By identifying the TATA box, it is possible to study the regulation of H4 histone production during DNA packing. Here, a TATA box weight matrix is generated from RNA polymerase II promoter regions [14] as shown in Table 2.

Table 2: TATA box vector weight matrix

	-2	-1	0	1	2	3	4	5
A	16	352	3	354	268	360	222	155
T	309	35	374	30	121	6	121	33
G	18	2	2	5	0	20	44	157
C	46	0	10	0	0	3	2	44
p_A	0.0411	0.9049	0.0077	0.91	0.6889	0.9254	0.5707	0.3985
p_T	0.7943	0.09	0.9614	0.0771	0.3111	0.0154	0.3111	0.0848
p_G	0.0463	0.0051	0.0051	0.0129	0	0.0514	0.1131	0.4036
p_C	0.1183	0	0.0257	0	0	0.0077	0.0051	0.1131
x	0.9428	0	0.9428	0	0.4714	0	0.4714	-0.2357
y	0	0	0	0	0	0	0	-0.4082
z	-0.3333	1	-0.3333	1	0.3333	1	0.3333	0.3333

Based on the TATA box weight matrix in Table 2, the TATA box vector weight matrix is used to identify possible sites containing the motif within *H4*/*g* gene. The vector weight matrix is aligned with *H4*/*g* gene sequence base by base until the last base. For each alignment, match alignment information is calculated and plotted in Figure 2.

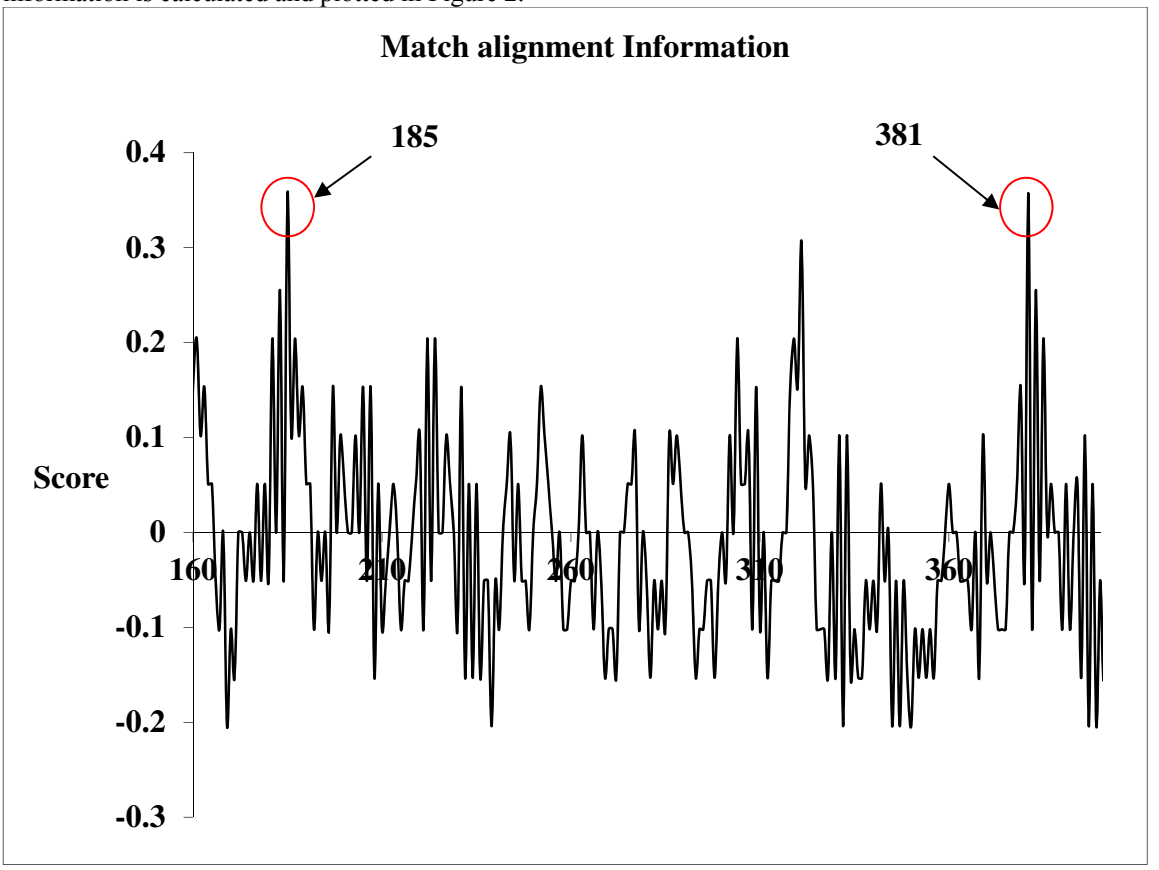

Figure 2: Match alignment information of *Homo sapiens H4*/*g* gene

3.2 Results

The plot composed of peaks and troughs. Each peak represents a local high similarity between the TATA box vector weight matrix and the regions within *H4*/*g* gene, and *vice versa* for troughs. There are two highly possible sites (positions *185* and *381* in the gene) at which the TATA box could be situated. However, it is not clear which is the real TATA box since both have the same score. In order to distinguish these two sites, mismatch alignment information should be used.

3.3 Discussion

Mismatch alignment information as discussed is classified into three categories: transition, complement and transversion. Among these three types of mismatches, transitional mismatch is considered the most acceptable. If a base type A is replaced by G or *vice versa*, it is deemed to have less effect on the process of transcription. This is because both base types have the same size. However, if a different size base type C replaces base type A in transversal mismatch, the degree of mismatch is greater than that of transitional one. Standing in between is complementary mismatch, although bases A and T may have a different size, they bind to each other in a DNA chain. Given

these three classes of mismatch alignment information, transitional mismatch is the most acceptable. This is followed by complementary mismatch and the least acceptable transversal mismatch.

A comparison of transitional mismatch alignment information shows that they are the same for both predicted TATA boxes. It is still not possible to differentiate the dissimilarity between these two sites. Hence complementary mismatch alignment information should be considered. Figure 3 shows that the motif at the site *185* has higher complementary mismatch alignment information as compared with the motif at the site *381*. This implies that a base A is replaced by a base T, as well as G by C more often in motif at the site *185*. Hence the more acceptable mismatch, complementary mismatch, occurs more often at the site *185*. Since the TATA box consists of a string of A and T bases, the complementary mismatch between A and T is unlikely to affect the function of the TATA box as a signal for transcription process. Therefore by using the complementary mismatch alignment information, the site *185* is the more likely site for the TATA box in *H4/g* gene.

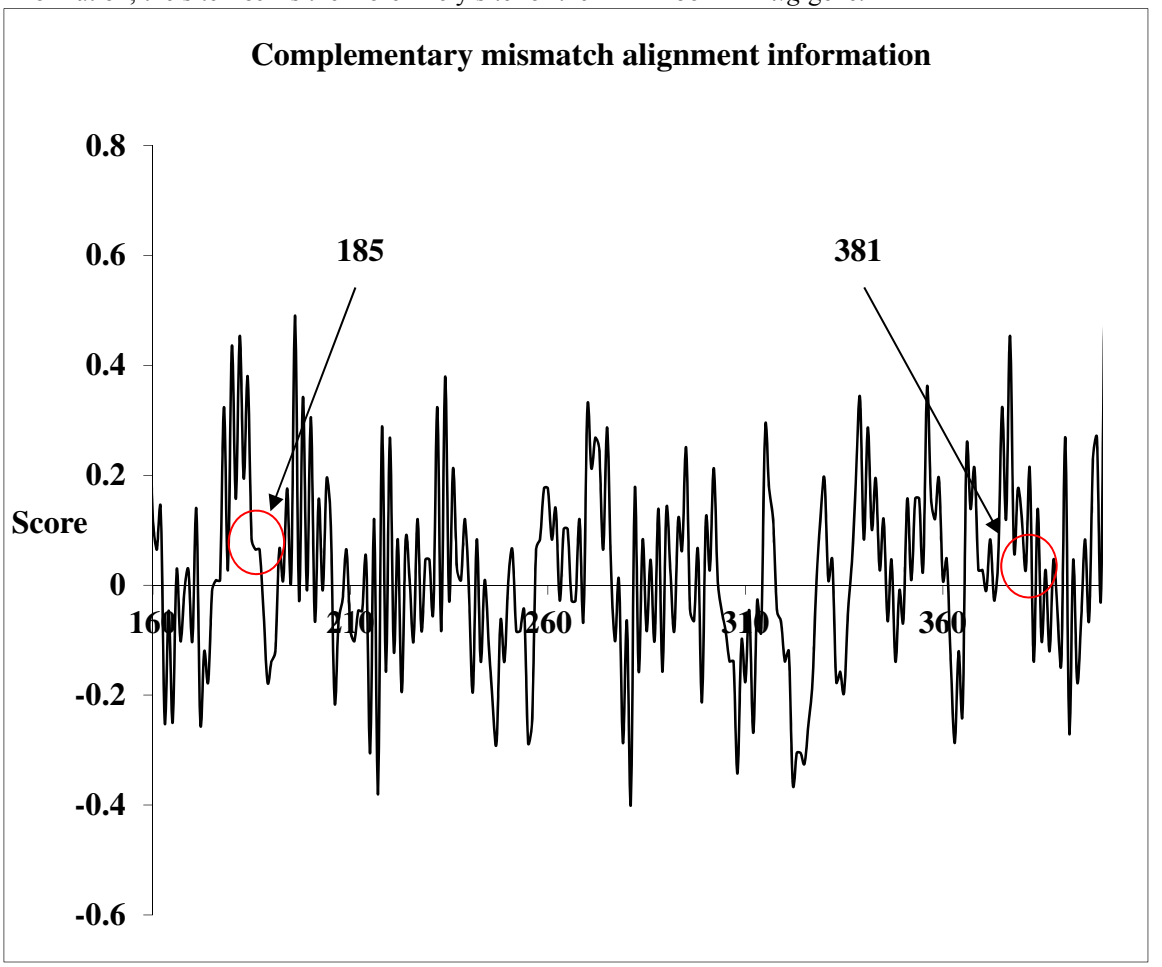

Figure 3: Complementary mismatch alignment information of *Homo sapiens H4/g* gene

4 Conclusion

DNA contains important motifs, such as binding sites. More often than not, the motifs are not exact. A weight matrix is the useful way of representing an alignment and also used as a scoring

means to predict motif sites. However by using match alignment information alone, there may be too many falsely predicted sites, especially for motif sequences that are less conserved. In order to reduce the number of false positives, the mismatch component of the alignment is considered. The vector weight matrix can be used to elaborate match and mismatch alignment information: the degree of match from the dot product and the degree of mismatch from the cross product. Using mismatch alignment information, the number of false positives can be reduced and the efficiency in identifying motif can be improved [15].

References

[1] Tompa, M. (1999) An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, 262-271.

[2] Shu, J.-J. and Li, Y. (2012) A statistical fat-tail test of predicting regulatory regions in the *Drosophila* genome. *Computers in Biology and Medicine*, **42**(9), 935-941.

[3] Shu, J.-J. and Li, Y. (2013) A statistical thin-tail test of predicting regulatory regions in the *Drosophila* genome. *Theoretical Biology and Medical Modelling*, **10**(11), 1-11.

[4] Hertz, G.Z., Hartzell, G.W. and Stormo, G.D. (1990) Identification of consensus patterns in unaligned DNA-sequences known to be functionally related. *Computer Applications in the Biosciences*, **6**(2), 81-92.

[5] Yang, C., Bolotin, E., Jiang, T., Sladek, F.M. and Martinez, E. (2007) Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, **389**(1), 52-65.

[6] Kulakovskiy, I.V., Favorov, A.V. and Makeev, V.J. (2009) Motif discovery and motif finding from genome-mapped DNase footprint data. *Bioinformatics*, **25**(18), 2318-2325.

[7] Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., Posch, S. and Grosse, I. (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics*, **21**(11), 2657-2666.

[8] Shu, J.-J., Yong, K.Y. and Chan, W.K. (2012) An improved scoring matrix for multiple sequence alignment. *Mathematical Problems in Engineering*, **2012**(490649), 1-9.

[9] Berger, J.A., Mitra, S.K., Carli, M. and Neri, A. (2004) Visualization and analysis of DNA sequences using DNA walks. *Journal of the Franklin Institute-Engineering and Applied Mathematics*, **341**(1-2), 37-53.

[10] Afreixo, V., Ferreira, P.J.S.G. and Santos, D. (2004) Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*, **14**(6), 523-530.

[11] Coward, E. (1997) Equivalence of two Fourier methods for biological sequences. *Journal of Mathematical Biology*, **36**(1), 64-70.

[12] Shu, J.-J. and Ouw, L.S. (2004) Pairwise alignment of the DNA sequence using hypercomplex number representation. *Bulletin of Mathematical Biology*, **66**(5), 1423-1438.

[13] Shu, J.-J. and Li, Y. (2010) Hypercomplex cross-correlation of DNA sequences. *Journal of Biological Systems*, **18**(4), 711-725.

[14] Bucher, P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology*, **212**(4), 563-578.

[15] Shu, J.-J., Wang, Q.-W. and Yong, K.-Y. (2011) DNA-based computing of strategic assignment problems. *Physical Review Letters*, **106**(18), 188702.