MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **73** (2015) 545-558

ISSN 0340 - 6253

Position–Specific Statistical Model of DNA Sequences and Its Application for Similarity Analysis

Chenkui Kuang[1,+], Xiaoqing Liu[2,+], Junru Wang[3], Yuhua Yao[1], Qi Dai[1,4,*]

1 College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China

2 College of Sciences, Hangzhou Dianzi University, Hangzhou 310018, China

3 College of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China

4 Department of Molecular and Cell Biology, University of Texas at Dallas, Richardson, TX 75080, USA

(Received August 27, 2014)

Abstract

Sequence similarities serve as a foundation for the structure-function studies, as well as evolutionary relationships among different species. Several statistical methods have been proposed to compare biological sequences, but challenge remains. This paper proposed a novel position-specific statistical model based on base-specific position matrices and Markov model to describe randomness degree and local dynamic distribution in specific positions. We illustrated its application in analysis of chromosomes 1 of eight species, human coding/ non-coding sequences and phylogenetic relationships. The experiment results demonstrate that nucleotides' distributions in the first position from right of the nucleotide C and left of the nucleotide G are extremely uneven, and most of conserved patterns in non-coding sequences are relatively longer and more conservative than that in coding sequences. This understanding can then be used to guide development of more powerful methods for sequence analysis.

Introduction

In past 10 years, sequencing technology has been undergoing dramatic changes, and consequently thousands of genomes have been sequenced until now [1-5]. Sequence data does not, by itself, increase the scientist's understanding of the biology of organisms, but creates

[+] Chenkui Kuang and Xiaoqing Liu contributed equally to this work as co-first authors.

[*] Corresponding author. E-mail addresses: daiailiu04@yahoo.com

many challenges for bio-scientists [6]. One of challenges is how to compare the biological sequences, which is a more important strategy in sequence analysis.

Up to now, many methods have been proposed to compare biological sequences, which can be categorized into two classes. One is traditional alignment method, which is a common and much empirical to select a sequence alignment based on scoring matrix and gap penalty parameters. The advantage of using these matrices is based on a defined evolutionary model, and the statistical significance of alignment scores obtained by local-alignment programs may be evaluated. However, dynamic programming in alignment makes computation more complicated, and iterative computational steps limits its utility for large datasets [7]. Therefore, the research of alignment-free methods becomes apparent and necessary [8].

Graphical representation is one of widely used alignment-free methods, which was first proposed by Hamori and Ruskin [9]. From graphical representations, many numerical features of biological sequences were extracted to facilitate comparison of genome sequences and regulatory sequences. The advantage of such approaches is providing a simple way to view, sort and compare various gene structures, helping in recognizing major characteristics among similar biological sequences [10]. Since introduction of the graphical representation, various approaches have been proposed and achieved promising results in sequence comparison. But some 2D graphical representations are accompanied with information loss due to overlapping and crossing of the curve with itself [11, 12]. In order to reduce information loss, some 2D and 3D representations have been proposed [13-18]. According to the handling bases of biological sequences, they can be classified as single nucleotide-based and dual nucleotide-based representations. Although the above graphical representation methods have achieved promising results, there are still some problems in graphical representations. First, many graphical representations were designed by assigning the single bases or dual nucleotides to corresponding direction/points/cells in Cartesian coordinates, so little attention has been paid to the whole distribution of the single nucleotide or dual nucleotides in biological sequences. Second, the choice of the direction/points/cells for the single base or dual nucleotides is arbitrary. Finally, feature-based similarity measures are always associated with the invariants of the distance matrices that are gotten by complex repetitive computation. When the sequences are long, these kinds of feature-based similarity measures become less useful [19-25].

In addition, several statistical methods have been proposed to compare different biological sequences [26-29]. They used *k*-word distribution and Markov model [30, 31] to numerically characterize compositional characteristics and pseudo-periodic sequence patterns,

and used different distance measures to evaluate the similarity among different sequences. For example, Euclidean distance [32], Euclidean and Mahalanobis distance [33], Markov chain models and Kullback-Leiber discrepancy [34], cosine distance [35], Kolmogorov complexity [36] and chaos theory [37].

The above approaches have achieved promising results in sequence comparison, but several critical problems still exist in their development. First, mostly methods focus on the content of sub-sequences of biological sequences, and therefore to sometimes are unaware of useful position-based information of the specific bases in sequence comparison. Second, some methods constructed Markov model to numerically characterize biological sequences with help of transition probabilities, without considering local structure among *k*-neighborhood of the given bases. Since steady local structures are strongly associated with compact structural pattern or domains, we should take them into account when comparing two biological sequences.

With above problems in mind, we proposed a position-specific statistical model to analyze biological sequences. Given a biological sequence, we constructed base-specific position matrices and further analyzed randomness degree of nucleotides' distribution with help of Shannon entropy. In order to describe local structure of the given bases, we proposed a position-specific statistical model based on the base-specific position matrices and Markov model. To evaluate performance of the proposed model, we applied it to similarity/dissimilarity studies of DNA sequences and compared its performance with the alignment method.

Method

Base-specific position matrices

A biological sequence $S = s_1 s_2 \cdots s_n$ is interpreted as a succession of symbols from the alphabet, where *n* is the length of the *S*. In the context of DNA sequences, its state space is $\omega = \{A, C, G, T\}$. There is a large body of literatures on word statistics in which the frequencies of its small segments were analyzed. Given a base χ, this paper was interested not only in its only distribution in biological sequence, but also in other bases' distribution among its *k*-neighborhood. To obtain such information, we proposed base-specific position matrices to describe the bases' distribution around the given base. Take a DNA sequence for example, we first designed two sliding windows with length of *L* and put them on both sides

of the given nucleotide χ. When merging the given nucleotide χ and two windows into one, we got a base-specific sliding window of length *2L+1*. The standard approach for analyzing nucleotides' distribution of a sequence with length *n* is to use a sliding window of length *2L+1*, shifting one base at a time from position 1 to *n-2L*. Here, the sliding window is allowed to overlap in the sequence. We then obtained two base-specific position matrices ${}^{\chi}M_l^k$ and ${}^{\chi}M_r^k$ defined as follows:

$$ {}^{\chi}M_l^k = \begin{bmatrix} f_l^{\chi}(A,L) & \cdots & f_l^{\chi}(A,2) & f_l^{\chi}(A,1) \\ f_l^{\chi}(C,L) & \cdots & f_l^{\chi}(C,2) & f_l^{\chi}(C,1) \\ f_l^{\chi}(G,L) & \cdots & f_l^{\chi}(G,2) & f_l^{\chi}(G,1) \\ f_l^{\chi}(T,L) & \cdots & f_l^{\chi}(T,2) & f_l^{\chi}(T,1) \end{bmatrix} \quad (1) $$

$$ {}^{\chi}M_r^k = \begin{bmatrix} f_r^{\chi}(A,1) & f_r^{\chi}(A,2) & \cdots & f_r^{\chi}(A,L) \\ f_r^{\chi}(C,1) & f_r^{\chi}(C,2) & \cdots & f_r^{\chi}(C,L) \\ f_r^{\chi}(G,1) & f_r^{\chi}(G,2) & \cdots & f_r^{\chi}(G,L) \\ f_r^{\chi}(T,1) & f_r^{\chi}(T,2) & \cdots & f_r^{\chi}(T,L) \end{bmatrix} \quad (2) $$

where $\sum_{y\in\{A,C,G,T\}} f_l^{\chi}(y,k)=1$ and $\sum_{y\in\{A,C,G,T\}} f_r^{\chi}(y,k)=1$, $f_l^{\chi}(y,k)$ ($f_r^{\chi}(y,k)$) is the frequency of the nucleotide y in the *k*th position starts from left (right) of the given nucleotide χ.

For example, if $S = ATGATGCACATGACTC$ is a given DNA sequence, we calculated its *A*-specific position matrices ${}^{A}M_l^3$ and ${}^{A}M_r^3$ with length 3 as follows:

$$ {}^{A}M_l^3 = \begin{bmatrix} 0.5 & 0.25 & 0 \\ 0.25 & 0 & 0.5 \\ 0 & 0.25 & 0.5 \\ 0.25 & 0.5 & 0 \end{bmatrix}, \quad {}^{A}M_r^3 = \begin{bmatrix} 0 & 0.2 & 0.4 \\ 0.4 & 0 & 0.4 \\ 0 & 0.6 & 0 \\ 0.6 & 0.2 & 0.2 \end{bmatrix}. $$

Following the same method, we also obtained base-specific position matrices ($ {}^{C}M_l^3, {}^{C}M_r^3$), ($ {}^{G}M_l^3, {}^{G}M_r^3$) and ($ {}^{T}M_l^3, {}^{T}M_r^3$) of the specific nucleotides *C*, *G* and *T*.

Markov model

Markov model is a stochastic model that models the state of a system with a random variable that changes through time. The general idea of Markov model lies in the letters are not necessarily independent and the probabilities for a_i depend on the 'past' of the sequence, namely ($a_1, a_2, \cdots, a_{i-1}$). In 1-step Markov chain, a_i depends on the past state a_{i-1}, which can be described by transition probability p_{ij}. The transition matrix $P=[p_{ij}]$ is defined as a

matrix composed of all the transition probabilities, and each state transition probability p_{ij} was defined as follows:

$$p(i, j) = p(a_n = S_i \mid a_{n-1} = S_j) \tag{3}$$

where S_i and S_j are the *i*-th and *j*-th states of a space with *N* distinct states. The state transition probabilities are subject to

$$\sum_{j=1}^{N} p(i, j) = 1, \forall\, i \tag{4}$$

$$p(i, j) \geq 0 \;\forall\; i,\; j \tag{5}$$

Let p_{ij}^k denotes the probability that Markov chain is in state *j* after *k* steps from state *i*

$$p_{ij}^k = p(a_{n+k} = S_j \mid a_n = S_i) \tag{6}$$

we have

$$\begin{aligned} p_{ij}^k &= \sum_{r=1}^{n} p(a_k = S_j \mid a_{k-1} = r, a_0 = S_i)\, p(a_{k-1} = r \mid a_0 = S_i) \\ &= \sum_{r=1}^{n} p(a_k = S_j \mid a_{k-1} = r)\, p(a_{k-1} = r \mid a_0 = S_i) \\ &= \sum_{r=1}^{n} p_{rj} p_{ir}^{k-1} \end{aligned} \tag{7}$$

So *k*-step Markov chain can be deduced from the 1-step Markov chain, and the transition probability matrix of the *k*–step Markov model can be calculated as follows:

$$\left[p_{ij}^k \right] = P^k .$$

Position-specific statistical model

Base-specific position matrices ${}^{\chi}M_l^k$ and ${}^{\chi}M_r^k$ describe the bases' distributions around the given base χ, and Markov model indicates transition probability from one state to another. Given a base χ, not only do we want to know the bases' distribution around it, but in the process we'd like to analyze their transition among the *k*-neighborhood bases. In order to further numerically characterize local structure of the given bases, we proposed a position-specific statistical model Γ based on the base-specific position matrices and Markov model

$$\Gamma = \{ {}^{\chi}\Gamma_l^L,\ {}^{\chi}\Gamma_r^L \mid \chi \in \{A, C, G, T\} \} \tag{8}$$

$${}^{\chi}\Gamma_l^L = [{}^{\chi}\pi_l^L(y, k)]_{\chi \in \{A, C, G, T\},\, k \leq L} \tag{9}$$

$${}^{\chi}\Gamma_r^L = [{}^{\chi}\pi_r^L(y, k)]_{\chi \in \{A, C, G, T\},\, k \leq L} \tag{10}$$

$$ {}^{\chi}\pi_l^L(y,k) = f_l^{\chi}(y,k) \times p^{L+1-k}(y,\chi) \quad (11) $$

$$ {}^{\chi}\pi_r^L(y,k) = \begin{cases} f_r^{\chi}(y,k) \Big/ p^k(y,\chi) & p^k(y,\chi) \neq 0 \\ 0 & else \end{cases} \quad (12) $$

where $f_l^{\chi}(y,k)$ ($f_r^{\chi}(y,k)$) is the frequency of the nucleotide y in the *k*th position starts from left (right) of the given nucleotide χ, and $p^k(y,\chi)$ denotes transition probability of the nucleotide y to the given nucleotide χ after *k* steps, $y \in \{A, C, G, T\}$, and $k \leq L$.

From the statistical model Γ, it is easily to note that it consists of two kinds of distribution information in the specific positions of the given bases. One is the 'gotten' information because ${}^{\chi}\pi_l^L(y,k)$ describes the probabilities that the nucleotide y in the *k*th position starts from left becomes the given nucleotide χ after $L+1-k$ steps. On the contrary, ${}^{\chi}\pi_r^L(y,k)$ is the 'lost' information in which the given nucleotide χ is from the nucleotide y in the *k*th position starts from right after *k* steps. That is to say, the proposed statistical model describes local dynamic distribution in the specific positions around the given base. Therefore, it can be used to numerically characterize local distribution information of the bases and pseudo-periodic sequence patterns in biological sequences.

Results and Discussion

Evaluation on local distribution of the specific nucleotide

Given a nucleotide, we are interested in both its distribution in biological sequence and local distribution of the other nucleotides around it. With help of the base-specific position matrices, we analyzed the local distribution of the specific nucleotides. It is well known that Shannon's entropy measures the degree of bias/randomness of a distribution. Here, we utilized it as a measure (*RD*) to evaluate the randomness degree of nucleotides' distribution in the given position, defined as follows:

$$ RD(k) = -\sum_{y \in \{A,C,G,T\}} f(y,k) \log_2 f(y,k) \quad (12) $$

where $f(y,k)$ is the frequency of the nucleotide y in the *k*th position.

Here, we chose chromosomes 1 of eight species, in which the brassicaceae species is Arabidopsis (A. thaliana), the Chordata species is zebrafish (Brachydanio rerio), the Vertebrate and Mammal species are cat (Feliscatus) and dog (Canislupusfamiliaris), the roden(Mus musculus), the rodents species are rat (Rattus norvegicus) and mouse (Mus

musculus), and the primates species are human (Homo sapiens) and chimpanzee (Gorrilla gorilla). Figure 1 is the randomness degree of nucleotides' distribution in five positions on both sides of the specific nucleotide for eight species, in which negative values on the x-axis denote the positions on the left side of the specific nucleotides, and positive values denote the position on the right side of the specific nucleotides.

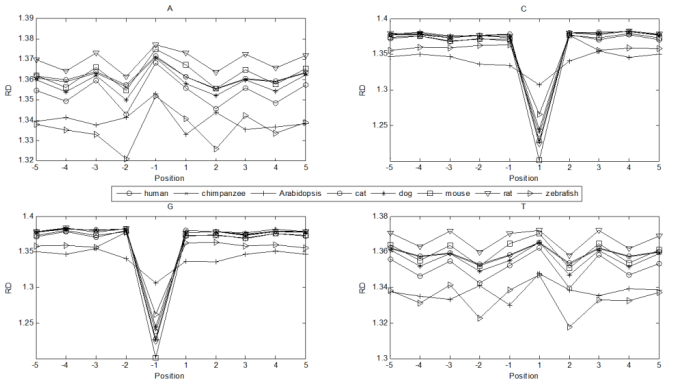


Figure 1. Randomness degree of nucleotides' distribution among five positions on the left and right of the specific nucleotide for eight species, in which negative values on the x-axis denote the position on the left side of the specific nucleotides, and positive values denote the position on the right side of the specific nucleotides.

Take a closer look at Figure 1, we found that (i) randomness degrees' changes of the five positions on the left and right of the specific nucleotides A and T (nucleotides C and G) are similar among the eight species, especially for mammals. In the double-stranded DNA, nucleotide A and nucleotide T (C and G) are complementary pairing, so their local distributions are similar; (ii) minimum randomness degrees are associated with the position 1 or -1 in eight species (Figure 1(C) and Figure 1(G)). As we all know, Shannon's entropy would reach its maximum value when the distribution is uniform one. That is to say, the distributions of four nucleotides in the first position on the right of the nucleotide C and on the left of the nucleotide G are extremely uneven; (iii) Although the randomness values possess different performances with different given nucleotides, it is interesting to note that the species belonging to the same class achieve similar performance. For example, the randomness degree of nucleotides' distributions of human is coincident with that of chimpanzee. In contrast, the randomness degrees of nucleotides' distributions of arabidopsis and zebrafish are significantly different from the others.

Human coding and non-coding sequences analysis

The initial step in genomic annotation is to identify protein coding regions of the genomes. It is a challenging problem because protein coding regions interrupted by non-coding regions are usually not continuous. So it is difficult to distinguish protein coding regions from non-coding regions by the sequences since there is no obvious sequence feature between them. Here, we used the proposed model to analyze the human coding and non-coding sequences to find some interesting features. The human coding dataset and non-coding dataset were downloaded from the nucleotide database of NCBI (National Center for Biotechnology Information), which includes 285807 coding sequences and 248713 non-coding sequences, respectively. Figures 2-3 represent nucleotides' content in the positions on the left or right of the specific nucleotides, and the colour from black to grey denotes the content from low to high.

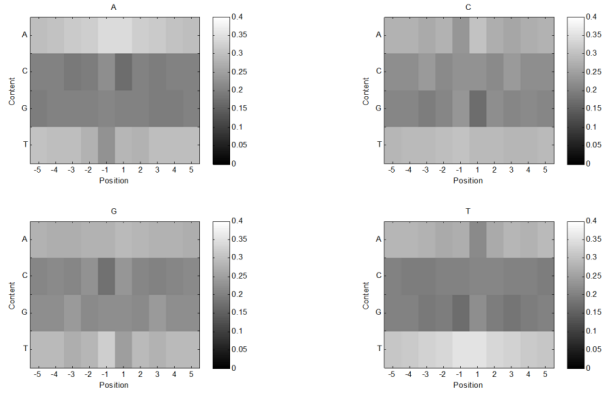


Figure 2. Nucleotides' content in the positions on the left or right of the specific nucleotides in human coding dataset, and colour from black to grey denotes the content from low to high. Negative values on the x-axis denote the position on the left side of the specific nucleotides, and positive values denote the position on the right side of the specific nucleotides.

As would be expected, the nucleotides' contents in different positions around the specific nucleotides (Figures 2-3) show three clear trends: (i) the contents of A and T are larger than others for the different specific nucleotides in human coding and non-coding sequences. As for the specific nucleotides A and T, this phenomenon is more obvious; (ii) the maximum contents are always achieved in the position 1 or -1. That is to say, the distributions of four nucleotides in the first position on the right or left of the specific nucleotides are extremely uneven. (iii) Figure 2 shows that the contents of AAA, TGA and

TTT patterns are significantly higher than that of other patterns, while Figure 3 illustrates the highest contents are associated with the AAAAA, TCA, TGA*A and TTTTT patterns. It is easy to observe that most of conserved patterns in non-coding sequences are relatively longer and more conservative than that in coding sequences. Therefore, the above features may be useful to numerically characterize the local distribution information of the bases and pseudo-periodic sequence patterns in the coding and non-coding sequences.

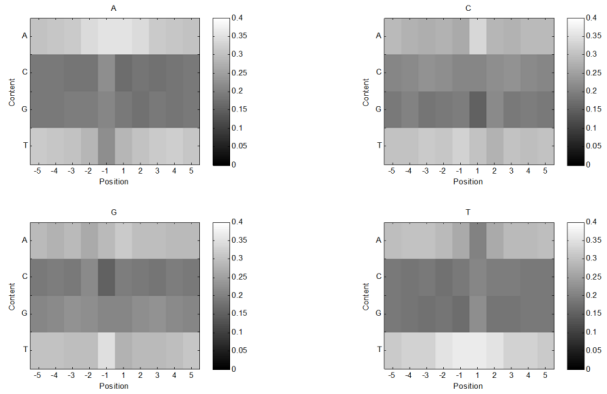


Figure 3. Nucleotides' content in the positions on the left or right of the specific nucleotides in human non-coding dataset, and color from black to gray denotes the content from low to high. Negative values on the x-axis denote the position on the left side of the specific nucleotides, and positive values denote the position on the right side of the specific nucleotides.

For a better understanding, we further calculated the randomness degree of nucleotides' distribution in the five positions on the left and right of the specific nucleotides, which represented in Figure 4. It is clear from Figure 4 that the randomness degrees' changes in the five positions on the left and right of the given nucleotides are similar between the human coding and non-coding sequences. But we should note that randomness degrees of human coding sequences are significant higher than that of human non-coding sequences. That is to say, the distributions of four nucleotides in the positions of human non-coding sequences are more uneven that of human coding sequences, which is consistent with the results in Figures 2-3.

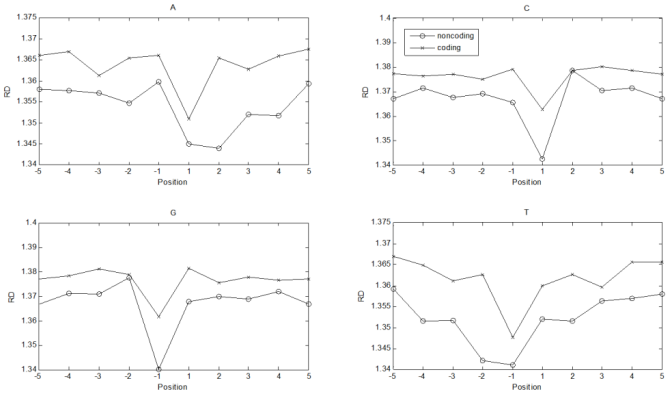


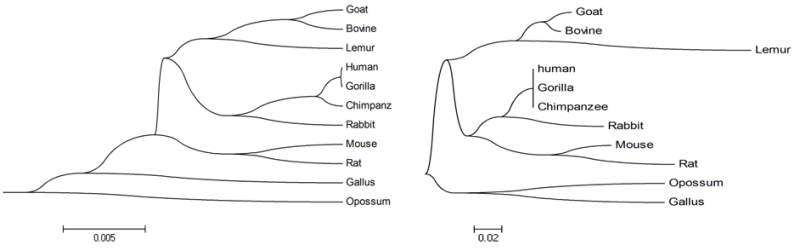
Figure 4. Randomness degree of nucleotides' distribution in five positions on the left and right of the specific nucleotide in human coding and non-coding datasets, in which negative values on the x-axis denote the position on the left side of the specific nucleotides, and positive values denote the position on the right side of the specific nucleotides.

Similarity analysis

Similarity analysis is the essential motivation of sequence analysis. Here, we further tested the validity of the proposed method by analyzing the similarities of coding sequences of the first exon of β-globin gene of 11 species: Human(U01317), Goat(M15387), North American opossum(J03643), Gallus(V00409), Black lemur(M15734), House mouse(V00722), Rabbit(V00882), Norway rat(X06701), Gorilla(X61109), Bovine(X00376) and Chimpanzee(X02345), which have been widely studied.

In order to represent relationships among 11 species more clearly, we constructed their phylogenetic trees based on the proposed statistical model Γ. The phylogenetic tree were obtained through the following main operations: we first constructed the proposed statistical models Γ with the window size 5 for 11 species and measured the similarity based on Euclidian distances among their statistical models Γ; Through arranging all the similarity degrees into a matrix, we then obtained a pair-wise distance matrix; finally, we put the pair-wise distance matrix into the neighbor-joining program in the PHYLIP package and got phylogenetic trees drawn by MEGA program. Figure 5(A) is the phylogenetic tree of 11 species using the proposed statistical models Γ.

Generally, an independent method can be used to evaluate the accuracy of a phylogenetic tree, or the validity of a phylogenetic tree can be tested by comparing it with authoritative ones. Here, we adopt the former one to test the validity of our phylogenetic tree and compared our results with ones obtained with the multiple alignment CLUSTAL X (Figure 5(B)).



(A) (B)

Figure 5. Phylogenetic trees of 11 species with proposed method and the multiple alignment CLUSTAL W. (A) is obtained by the proposed statistical models with the window size 5, and (B) is the results based on the multiple alignment CLUSTAL W.

Figure 5 shows that the results of the proposed method are quite consistent with the ones obtained by the multiple alignment CLUSTAL X (Figure 5(B)) in the following four aspects: (1) three Primates (Human, Gorilla and Chimpanzee) are clustered closely; (2) two Rodents (Mouse and Rat) are grouped closely; (3) Rabbit is clustered closely with Human, Gorilla and Chimpanzee; (4) Opossum and Gallus are less closely with other species, which is consistent with the fact that Gallus is the only non-mammal among them, and Opossum is the most remote species from the remaining mammals. In agreement with the above analysis, these results confirm that the proposed position-specific statistical model can be considered as another efficient method for similarity analysis of biological sequences.

Conclusion

Similarity analysis is one of the major goals of sequence analysis, which could serve as evidence of structural and functional conservation, as well as evolutionary relations among the sequences. This paper constructed a position-specific statistical model based on the base-specific position matrices and Markov model to analyze biological sequences. Given a biological sequence, we first constructed its base-specific position matrices and analyzed the randomness degree of nucleotides' distribution with help of Shannon entropy. We found that the distributions of four nucleotides in the first position on the right of the nucleotide C and on the left of the nucleotide G are extremely uneven, and most of conserved patterns in non-

coding sequences are relatively longer than that in coding sequences. With the properties of Markov model in mind, we proposed a position-specific statistical model to analyze the biological sequences and illustrated its applications in phylogenetic analysis. The results demonstrate that the proposed method is efficient, which highlight the necessity for sequence comparison method to describe local dynamic distribution among *k*-neighborhood of the given bases. Thus, this understanding can then be used to guide development of more powerful methods for sequence comparison with future possible improvement on genome study.

Acknowledgments: This work is supported by National Natural Science Foundation of China (61170316, 61370015 and 61272312), a research Grants (LY14F020046) from Zhejiang Provincial Natural Science Foundation of China, and 521 Talent Cultivation Plan of Zhejiang Sci-Tech University.

References

[1] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, M. Smith, Nucleotide sequence of bacteriophage phi X174 DNA, *Nature* **265** (1977) 687-695.

[2] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, Whole-genome random sequencing and assembly of Haemophilus influenzaeRd, *Science* **269** (1995) 496-512.

[3] M. L. Metzker, Emerging technologies in DNA sequencing, *Genome Res.* **15** (2005) 1767-1776.

[4] N. J. Loman, R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, M. J. Pallen, Performance comparison of benchtop high-throughput sequencing platforms, *Nat. Biotechnol.* **30** (2012) 434-439.

[5] X. Zhou, L. Ren, Q. Meng, Y. Li, Y. Yu, J. Yu, The next-generation sequencing technology and application, *Protein Cell* **1** (2010) 520-536.

[6] The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, *Nature* **489** (2012) 57-74.

[7] M. R. Kantorovitz, G. E. Robinson, S. Sinha, A statistical method for alignment-free comparison of regulatory sequences, *Bioinformatics* **23** (2007) 249-255.

[8] S. Vinga, J. Almeida, Alignment-free sequence comparison-a review, *Bioinformatics* **19** (2003) 513-523.

[9] E. Hamori, J. Ruskin, H curves a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem.* **258** (1983) 1318-1327.

[10] M. Randić, 2-D graphical representation of proteins based on virtual genetic code, *SAR QSAR Environ. Res.* **15** (2004) 147-157.

[11] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol.* **119** (1986) 319-328.

[12] P. M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, *Comput. Appl. Biosci.* **11** (1995) 503-507.

[13] B. Liao, C. Zeng, F. Q. Li, Y. Tang, Analysis of similarity/dissimilarity of DNA sequences based on dual nucleotides, *MATCH Commun. Math. Comput. Chem.* **59** (2008) 647-652.

[14] B. Liao, W. Zhu, Y. Liu, 3D graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree, *MATCH Commun. Math. Comput. Chem.* **56** (2006) 209-216.

[15] Z. B. Liu, B. Liao, W. Zhu, A new method to analyze the similarity based on dual nucleotides of the DNA sequence, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 541–552.

[16] Y. Zhang, W. Chen, Analysis of similarity/dissimilarity of long DNA sequences based on three 2DD-curves, *Comb. Chem. High Throughput Screen.* **10** (2007) 231-237.

[17] Y. Zhang, W. Chen, Invariants of DNA sequences based on 2DD-curves, *J. Theor. Biol.* **242** (2006) 382-388.

[18] Y. Zhang, On 3D graphical representation of RNA secondary structure, *MATCH Commun. Math. Comput. Chem.* **57** (2007) 157-168.

[19] G. Huang, B. Liao, Y. Li, Y. Yu, Similarity studies of DNA sequences based on a new 2D graphical representation, *Biophys. Chem.* **143** (2009) 55-59.

[20] Y. Zhang, On 2D graphical representation of RNA secondary structure, *MATCH Commun. Math. Comput. Chem.* **57** (2007) 697-710.

[21] J. Song, H. Tang, A new 2-D graphical representation of DNA sequences and their numerical characterization. *J. Biochem. Biophys. Methods* **63** (2005) 228-239.

[22] J. F. Yu, J. H. Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 493-512.

[23] J. P. Stitt, K. M. Newell, Four-component power spectral density model of steady-state isometric force, *Biol. Cybern.* **102** (2010) 137-144.

[24] S. L. Zhang , Y. Zhang , I. Gutman, Analysis of DNA sequences based on the fuzzy integral, *MATCH Commun. Math. Comput. Chem.* **70** (2013) 417-430.

[25] Y. Zhang, J. K. Hao, C. J. Zhou, K. Chang, Normalized Lempel-Ziv complexity and its application in bio-sequence analysis, *J. Math. Chem.* **46** (2009) 1203-1212.

[26] J. Van Helden, Metrics for comparing regulatory sequences on the basis of pattern counts, *Bioinformatics* **20** (2004) 399-406.

[27] Y. Zhang, A simple method to construct the similarity matrices of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **60** (2008) 313-324.

[28] M. R. Kantorovitz, G. E. Robinson, S. Sinha, A statistical method for alignment-free comparison of regulatory sequences, *Bioinformatics* **23** (2007) 249-255.

[29] Y. Zhang, W. Chen, New invariant of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **58** (2007) 197–208.

[30] Q. Dai, Y. Yang, T. Wang, Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison, *Bioinformatics* **24** (2008) 2296-2302.

[31] Y. H. Yao, F. Kong, Q. Dai, P. A. He, Sequence-segmented method applied to the similarity analysis of long protein sequence, *MATCH Commun. Math. Comput. Chem.* **70** (2013) 431-450

[32] B. E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, *Proc. Natl. Acad. Sci.* **83** (1986) 5155-5159.

[33] T. J. Wu, J. P. Burke, D. B. Davison, A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words, *Biometrics* **53** (1997) 1431-1439.

[34] T. J. Wu, Y. C. Hsieh, L.A. Li, Statistical measures of DNA dissimilarity under Markov chain models of base composition, *Biometrics* **57** (2001) 441-448.

[35] G. W. Stuart, K. Moffett, S. Baker, Integrated gene and species phylogenies from unaligned whole genome protein sequences, *Bioinformatics* **18** (2002) 100-108.

[36] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, H. Y. Zhang, An infarmation-based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics* **17** (2001) 149-154.

[37] J. S. Almeida, J. A. Carrico, A. Maretzek, P. A. Noble, M. Fletcher, Analysis of genomic sequences by chaos game representation, *Bioinformatics* **17** (2001) 429-437.