Characteristic Sequences for DNA Primary Sequence

Ping-an He*,† and Jun Wang†,‡

Department of Applied Mathematics and College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, P.R. China

Received December 17, 2001

A DNA sequence can be identified with a word over an alphabet $\mathcal{N} = \{A, C, G, T\}$. Characteristic sequences of a DNA sequence are given in term of classifications of bases of nucleic acids. Using the characteristic sequences, we construct a set of 2 × 2 matrices to represent DNA primary sequences, which are based on counting of the frequency of occurrence of all (0,1) triplets of characteristic sequences. Furthermore, the leading eigenvalues of these matrices are computed and considered as invariants for the DNA primary sequences. Similarity and dissimilarity analysis based on the characteristic sequences are given for eight exon-1 genes of β-globin about eight species.

INTRODUCTION

With the development of the sequencing technique, a large number of DNA primary sequence data are collected into various data banks. Analysis and understanding of DNA primary sequences are very important tasks in bioinformatics.

Usually, a DNA primary sequence can be taken as a string of letters *A*, *G*, *C*, *T*, which denote the four nucleic acid bases: adenine, guanine, cytosine, and thymine, respectively. Therefore, the analysis and understanding of DNA primary sequences are performed via comparisons of such strings of the four letters. Almost all such comparisons are based on alignment of the strings: a distance function is used to represent insertion, deletion, and substitution of letters in the compared strings. Using the distance function, one can compare DNA primary sequences and resolve the questions of the homology of macromolecules.

Several researchers have considered graphical representations for the DNA primary sequences, in particular, Hamori and Ruskin,[1] Leong and Morgenthaler,[2] Randic,[6,7] Raychaudhury and Nandy,[8] Zhang[9,10] and others, considering a real DNA primary sequence as a curve embedded in 2-D plane or 3-D space. An advantage of graphical representations of DNA sequences is the possibility to derive numerical characterization for DNA primary sequences. For example, to characterize DNA primary sequences, Randic[6] considered a kind of condensed matrix called *D*/*D* matrix, the entries in which represent the quotient of the Euclidean and the graph theoretical distance between two vertices in the graphical representation of DNA sequences. And the leading eigenvalue of a D/D matrix was regarded as an index of folding of curve.

Also, Randic has introduced an alternative approach for comparison of DNA primary sequences, based on a set of invariants of DNA sequences, rather than directly using string comparisons. In a series of works,[3−7] Randic et al. have considered three kinds of condensed matrices: (1) matrices in which an individual entry corresponds to an individual pair of bases, (2) matrices in which entries sum information of different XY pairs of bases, and (3) matrices in which entries summarize information of different triplets of nucleic bases. Using these matrices, many invariants can be obtained for comparisons of DNA primary sequences.

Applying the above methods, the researchers have compared the similarities and dissimilarities of DNA primary sequences.

In this paper, based on classifications of the four nucleic acid bases, we shall reduce a DNA primary sequence into three (0,1) sequences, called the characteristic sequences of the DNA sequence. Each characteristic sequence may be regarded as a coarse-grained description of the DNA primary sequence. Via comparisons of the reduced sequences it will be easier to understand the biological function of various kinds of the nucleic acid bases.

Also, following Randic's approach, we shall construct a set of 2 × 2 matrices for the characteristic sequences of a DNA primary sequence and introduce a set of novel invariants to characterize the DNA primary sequence. Furthermore, we will make a comparison for the first exon of β-globin genes sequences belonging to eight different species. In Table 1, the exon-1 of the β-globin gene for eight species are listed, which were reported by Randic.[3]

CHARACTERISTIC SEQUENCES

Nucleic acids and proteins are all linear macromolecules. Comparison of DNA primary sequences should be considered not only the strings' structures but also their chemical structures. In DNA primary sequences, the four bases *A*, *C*, *G*, *T* can be divided into two classes according to their chemical structures, i.e., purine $R = \{A, G\}$ and pyrimidine $Y = \{C, T\}$. The bases can be divided into another two classes, amino group $M = \{A, C\}$ and keto group $K = \{G, T\}$. Besides these, the division can be also made according to the strength of the hydrogen bond, i.e., weak H-bonds $W = \{A, T\}$ and strong H-bonds $S = \{G, C\}$.

For a DNA primary sequence, using first classification, we reduce the sequence into a (0,1) sequence, that is, by 1

* Corresponding author e-mail: pinganhe@yahoo.com.cn.
† Department of Applied Mathematics.
‡ College of Advanced Science and Technology.

10.1021/ci010131z CCC: $22.00
Published on Web 07/16/2002

Table 1. Exon-1 of the β-Globin Genes for Eight Species

human β-globin 92 bases:

ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGT

GAACGTGGAGTAAGTTGGTGGTGAGGCCCTGGGCAG

goat β-globin 86 bases:

ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGT

GGATGAAGTTGGTGCTGAGGCCCTGGGCAG

gallus β-globin 92 bases:

ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGT

CAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG

opossum β-hemoglobin 92 bases:

ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGT

GCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG

lemur β-globin 92 bases:

ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGT

GGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG

mouse β-globin 94 bases:

ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGG

TGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG

rabbit β-globin 90 bases:

ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGT

GAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC

rat β-globin 92 bases:

ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGT

GAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG

(0) we denote the elements of R (Y). In this representation, some information of the DNA sequence structure may be lost; however, it does make it easier to compare sequences. Moreover, the comparison of the reduced sequences will reveal the functions of purine and pyrimidine.

We do similar operations on the sequence according to the second and third classifications to reveal the functions of amino-keto groups and weak-strong H-bonds, respectively. Thus, we obtain three (0,1) sequences corresponding to the same DNA primary sequence, and we call them (R, Y)-, (M, K)-, and (W, S)- characteristic sequences of the DNA primary sequence, respectively. The following mathematical theorem says that the three characteristic sequences give all information of the primary sequence.

Theorem. A DNA primary sequence is uniquely determined by any pair of its three characteristic sequences.

Proof: Let $G = g_1g_2\cdots$ be an arbitrary DNA primary sequence. Then we have three maps ϕ_i, $i = 1$, 2, and 3, which maps G into the (R, Y), (M, K), and (W, S)-characteristic sequences, respectively. Explicitly, $\phi_i(G) = \phi_i(g_1)\phi_i(g_2)\cdots$, where $\phi_1(g_i) = 1$ if $g_i \in R$ and $\phi_1(g_i) = 0$ if $g_i \in Y$; $\phi_2(g_i) = 1$ if $g_i \in M$ and $\phi_2(g_i) = 0$ if $g_i \in K$; $\phi_3(g_i) = 1$ if $g_i \in W$ and $\phi_3(g_i) = 0$ if $g_i \in S$. We thus obtain that every g_i correspondences a (0,1)-triplet ($\phi_1(g_i)$,$\phi_2(g_i)$,$\phi_3(g_i)$). By definition we see that $A \rightarrow (1, 1, 1)$, $C \rightarrow (0, 1, 0)$, $G \rightarrow (1, 0, 0)$, and $T \rightarrow (0, 0, 1)$, from which the theorem follows immediately.

So, we can compare their characteristic sequences to obtain the similarities/dissimilarities for DNA primary sequences. In Tables 2−4, the characteristic sequences corresponding to Table 1 are listed.

CONSTRUCTION OF THE CONDENSED MATRICES

For a DNA primary sequence, Randic et al.[5] have introduced a 4 × 4 × 4 cubic matrix based on the enumeration m_{ijk} of the occurrence of the triplets in the DNA sequence. Using these matrices, they gave a method for comparison of DNA primary sequences.

Instead of 64 possible triplets that can occur in a DNA primary sequence, there are only eight possible triplets in a characteristic sequence X: 000, 001, 010, 011, 100, 101, 110, 111. We introduce a 2 × 2 × 2 cubic matrices with eight entries $f^X_{ijk} = 100m^X_{ijk}/(N - 2)$, where m^X_{ijk} is the enumeration of the (0,1) triplet ijk in X and N is the length of X. Clearly, it represents 100 times the frequency of occurrence of the (0,1) triplet ijk in X. That we take the 100 times is for convenience of tabulation and computation.

Table 2. (R,Y)-Characteristic Sequences of the Eight DNA Sequence of Table 1

human

1011010100011000001111111110001001001001000010111101111011

1010111101110011011011110000111011

goat

1010011001001111111111001001001001100000111101111011111011

1011110011010011110000111011

gallus

101101010011100100111111111011000100100110000001111011

1010110011101011110011110000110011

opossum

1011010100011000001111111111100101001001001000110001

1110011001110011011011110000011011

lemur

10110000100111010011111111010001010010000000010111101111011

1010111111110011011011110000111011

mouse

10110010100011001101001111110001001000000010001011101111101

110000110111100110110111100001110111

rabbit

10110101000100011011111111110001011001001000010111101111011

10101111111100110110111100001110

rat

1011010100011001101001111111001001001101100010111111111011

1000011011010011010011110000111011

By F^R, F^M, and F^W we denote the cubic matrices for the (R, Y)-, (M, K)-, and (W, S)-characteristic sequences, respectively. We partition each of the cubic matrices into a pair of 2×2 condensed matrices $F^X{}_0$ and $F^X{}_1$, where $F^X{}_0 = (f^X_{0jk})$ and $F^X{}_1 = (f^X_{1jk})$ with X being R, M, or W.

In Tables 5 −7, the condensed matrices are constructed for eight exon-1 sequences of β-globin gene in Table 1, where the headers of the two 2 × 2 matrices represent the first entry i of a triplet (i, j, k) and the j, k entries of the triplets consist of j (row) and k (column) entries of the 2 × 2 matrices.

Table 3. (*M*,*K*)-Characteristic Sequences of the Eight DNA Sequence of Table 1

human
1000001111001101100100101100100110001100111000000011100001110000100110000000000001001110000110

goat
100100110010010010110010011001111001001000001110000111000010011000000001001001110000110

gallus
10000011100011001001001011011010110111100110100000111000111000001101100000001101101110001110

opossum
1000001110001100100100101101110011011101111010000101100001100000111101100000000010011100000110

lemur
1001100001001000100100101100101100011110101000000001110000010001010111000000000101001100000110

mouse
10000001111001100100100101100100100010100011000000011110000111110100110000000000010011100001100

rabbit
1000001101000111000100101100100100011100111000000011100001100000110110000000000010011100001

rat
1000001111011100100100101100101100001000011000000001110000111110010110000001010010011100000110

Table 4. (*W*,*S*)-Characteristic Sequences of the Eight DNA Sequence of Table 1

human
11001001001010100101001011010100001110100001010000011001011001001011101100100101000001000010

goat
110010101001010010110001000010100000110100000110010111010011011011001001010000001000010

gallus
1100100101001010010100101100100101101000000010100000110010111010000011101000000001100001000010

opossum
110010010110101101010010110110100110101100110100101110010010011010010101001001010000001100010

lemur
110101110010101001010010111001011010100101010100000110010011011010111011001000010000011000010

mouse
110011001001010101100101011010100101010110001010000111001011000000110110110010010100000010000100

rabbit
110010011010100101010010110101000010101000010100000110010111010011011011001001010000010000

rat
1100100100111010110010101100011010111010000101000011100101100010111110110000010100000001000010

Table 5. Frequence of Triplets *ijk* for Eight (*R*,*Y*)-Characteristic Sequences in Table 2

human					
0	0	1	1	0	1
0	10	10	0	10	13
1	9	14	1	13	20
goat					
0	0	1	1	0	1
0	6	13	0	13	8
1	7	14	1	13	25
gallus					
0	0	1	1	0	1
0	7	13	0	13	10
1	8	16	1	14	19
opossum					
0	0	1	1	0	1
0	10	13	0	13	10
1	9	14	1	13	17
lemur					
0	0	1	1	0	1
0	11	9	0	9	13
1	9	13	1	12	23
mouse					
0	0	1	1	0	1
0	12	12	0	12	11
1	8	15	1	14	16
rabbit					
0	0	1	1	0	1
0	8	9	0	9	14
0	9	14	1	14	23
rat					
0	0	1	1	0	1
0	7	12	0	12	12
1	9	16	1	14	18

Table 6. Frequency of Triplets *ijk* for Eight (*M*,*K*)-Characteristic Sequences in Table 3

human					
0	0	1	1	0	1
0	26	16	0	16	2
1	6	13	1	13	7
goat					
0	0	1	1	0	1
0	17	21	0	21	1
1	12	11	1	11	6
gallus					
0	0	1	1	0	1
0	19	12	0	12	10
1	6	17	1	17	8
opossum					
0	0	1	1	0	1
0	21	13	0	13	8
1	7	14	1	14	9
lemur					
0	0	1	1	0	1
0	24	16	0	16	8
1	14	9	1	9	4
mouse					
0	0	1	1	0	1
0	24	16	0	17	3
1	10	10	1	10	10
rabbit					
0	0	1	1	0	1
0	30	16	0	16	3
1	7	11	1	11	6
rat					
0	0	1	1	0	1
0	22	16	0	16	6
1	10	11	1	11	9

Observing Tables 5−7, we can obtain some common features of eight DNA primary sequences, respectively, that are not easily visible in Table 1.

In (*R*, *Y*)-characteristic sequences, the triplet 111(*RRR*) is the most occurring triplet, 010(*YRY*) and 000(*YYY*) are less occurring triplets. Some triplets, such as 011(*YRR*) and 110(*RRY*), show small variations in the frequency of occurrence, while others show considerable variations.

Similarly, in (*M*, *K*)-characteristic sequences, the most occurring triplet is 000(*KKK*) (goat is a expect), and the less are 101(*MKM*) and 111(*MMM*). In (*W*, *S*)-characteristic sequences, 010(*SWS*) and 111(*WWW*) is the most and least occurring sequence, respectively. The 001(*SSW*), 011(*SWW*), and 100(*WSS*) are small variations triplets, and the bigger variation is the triplet 000(*SSS*).

Since the eigenvalues of a matrix are one of the well-known matrix invariants, we consider the leading eigenvalues of the six matrices to acquire more compact information of the six matrices for each DNA primary sequence. In Table 8, the leading eigenvalues of six matrices are listed for all eight species, where each row can be considered as a six-component vector representation for one of the DNA primary sequences in Table 1. Generally, when the elements of

Table 7. Frequency of Triplets *ijk* for Eight (*W*,*S*)-Characteristic Sequences in Table 4

human					
0	0	1	1	0	1
0	13	17	0	17	13
1	23	7	1	8	2
goat					
0	0	1	1	0	1
0	17	15	0	15	13
1	20	8	1	10	1
gallus					
0	0	1	1	0	1
0	21	16	0	16	18
1	20	7	1	8	2
opossum					
0	0	1	1	0	1
0	4	16	0	16	18
1	22	11	1	12	1
lemur					
0	0	1	1	0	1
0	10	13	0	13	18
1	21	10	1	11	3
mouse					
0	0	1	1	0	1
0	12	14	0	15	16
1	21	10	1	11	1
rabbit					
0	0	1	1	0	1
0	14	14	0	15	17
1	23	8	1	9	1
rat					
0	0	1	1	0	1
0	16	13	0	13	13
1	17	10	1	11	7

Table 8. Leading Eigenvalues of the 6 Matrices F_0^X and F_1^X for the Eight DNA Sequence of Table 1

	F_0^R	F_1^R	F_0^M	F_1^M	F_0^W	F_1^W
human	21.7	28.9	31.3	18.3	30	22.2
goat	20.3	30.8	30.2	21.7	30.4	21.4
gallus	22.6	28.2	26.5	23.2	33.2	20.7
opossum	23	26.6	27.6	21.8	26.6	25
lemur	21.1	30.3	33.2	20.4	26.5	22.9
mouse	23.4	26.6	31.5	20	28.2	23
rabbit	20.5	31.7	34.7	18.6	29.2	22.2
rat	22.8	28.3	30.3	21.3	28.2	22.3

matrices do not vary strongly, the large leading eigenvalues display large average row/column sum and small leading eigenvalues reveal small average row/column sum for the 2 × 2 submatrix of triplets. Observing each row of Table 8, we can also obtain some common features about eight exon-1 gene, such as the leading eigenvalue of $F^M{}_0$ is maximal and the leading eigenvalue of $F^M{}_1$ is minimal expect goat and gallus.

In each row of Table 8, the values of $F^R{}_1$, $F^M{}_0$, and $F^W{}_0$ are large. Corresponding to the DNA primary sequence, this means that the values of $A + G$, $T + G$, and $C + G$ are large, so we can conclude that the number of base G is the largest among the frequency of occurrences of four letters for each sequence.

SIMILARITIES AND DISSIMILARITIES

In this section, we first construct a 24-component vector consisting of the frequency of occurrence of all possible triplets in the three characteristic sequences for each exon-1 gene. Using these vectors, we investigate similarities and dissimilarities for eight exon-1 gene. The frequency of triplets can be listed in any prescribed way. The underlying assumption is that if two vectors point to a similar direction in the 24-dimensional space, and then the two DNA sequences represented by the two 24-component vectors are similar.

The similarities among such vectors can be computed in two ways: (1) we calculate the Euclidean distance between the end point of the vectors, and (2) we calculate the cosine of the correlation angle of two vectors. The smaller Euclidean distance between the end points of two vectors, the more similar the DNA sequence. On the other hand, the larger the cosine of the correlation angle between two vectors, the more similar the DNA sequences.

In Table 9, the similarities and dissimilarities for eight exon-1 gene sequences that based on the Euclidean distance between the end points of the 24-component vectors are listed. Observing Table 9, we find gallus is very dissimilar to others among the eight species because its corresponding row has lager entries. On the other hand, the smaller entries are of human-rabbit, human-mouse, mouse-rat, lemur-rabbit, and mouse-rabbit.

In Table 10, we compute the magnitudes of the cosine of the angles between every pair of the 24-component vectors.

Table 9. Similarity/Dissimilarity Table for the Eight DNA Sequence of Table 1 Based on Euclidean Distance the End Point of the 24-Component Vectors

	goat	gallus	opossum	lemur	mouse	rabbit	rat
human	17.4356	17.3205	16.2481	15.748	11.4018	8.77496	14.3875
goat		20.8327	23.3666	18.6548	16.9706	19.105	15.7797
gallus			20.5913	23.7487	20.0499	21.2368	15.906
opossum				17.088	13.7113	19.105	17.9165
lemur					13.3417	13.2288	14.9332
mouse						13.2288	11.4455
rabbit							15.748

Table 10. Similarity/Dissimilarity Table for the Eight DNA Sequence of Table 1 Based on Cosine of the Angle between Two 24-Component Vectors

	goat	gallus	opossum	lemur	mouse	rabbit	rat
human	0.966373	0.966268	0.97017	0.972057	0.985443	0.992067	0.976887
goat		0.951535	0.938404	0.961071	0.967809	0.960983	0.972438
gallus			0.950952	0.935306	0.953775	0.951414	0.970704
opossum				0.966002	0.978037	0.960978	0.961688
lemur					0.979407	0.981786	0.973934
mouse						0.981871	0.984706
rabbit							0.97456

Table 11. Similarity/Dissimilarity Table for the Eight DNA Sequence of Table 1 Based on Euclidean Distance the End Point of the Six-Component Vectors

	goat	gallus	opossum	lemur	mouse	rabbit	rat
human	4.37493	7.8	7.18262	4.80416	3.87169	4.64435	3.84968
goat		6.02661	7.65441	5.38888	6.26339	5.72626	4.28019
gallus			8.25651	10.4461	8.29036	11.1045	6.75722
opossum				7.3437	4.93964	10.3005	4.4486
lemur					4.993	3.94842	4.4
mouse						6.94694	2.62107
rabbit							6.67158

Table 12. Similarity/Dissimilarity Table for the Eight DNA Sequence of Table 1 Based on Cosine of the Angle between Two Six-Component Vectors

	goat	gallus	opossum	lemur	mouse	rabbit	rat
human	0.997743	0.992511	0.993738	0.997226	0.998137	0.99809	0.998153
goat		0.99559	0.993397	0.996475	0.995336	0.996383	0.997892
gallus			0.991921	0.986676	0.991546	0.985731	0.994402
opossum				0.993888	0.997076	0.988892	0.997708
lemur					0.997057	0.998467	0.997724
mouse						0.995062	0.999138
rabbit							0.995454

Table 13. Similarity/Dissimilarity Table for the Eight DNA Sequence of Table 1 Based on Euclidean Distance the End Point of the Six-Component Vectors in Table 5

	goat	gallus	opossum	lemur	mouse	rabbit	rat
human	9.38083	6.55744	6	3.74166	5.91608	4.12311	5.19615
goat		6.85565	9.38083	9.59166	11.4455	9	8.77496
gallus			4.3589	9.32738	6.16441	8.3666	2.82843
opossum				8.94427	3.31662	9.43398	4.58258
lemur					9	3.87298	8.544
mouse						9.69536	5.65685
rabbit							7.2111

Table 14. Similarity/Dissimilarity Table for the Eight DNA Sequence of Table 1 Based on Euclidean Distance the End Point of the Six-Component Vectors in Table 6

	goat	gallus	opposum	lemur	mouse	rabbit	rat
human	13.3041	13.3417	9.27362	12.0416	7	5.19615	7.74597
goat		18.9473	15.6525	12.8062	10.7703	15.748	10.6301
gallus			5.47723	16.4012	15.2643	16.7033	12.083
opossum				12.2882	10.0499	12.2882	7.07107
lemur					8.94427	11.0454	7.54983
mouse						8	4.12311
rabbit							9.53939

Table 15. Similarity/Dissimilarity Table for the Eight DNA Sequence of Table 1 Based on Euclidean Distance the End Point of the Six-Component Vectors in Table 7

	goat	gallus	opossum	lemur	mouse	rabbit	rat
human	6.245	8.88819	11.9164	9.43398	6.78233	5.74456	10.9545
goat		5.2915	14.9545	9.59166	6.40312	6	7.681156
gallus			19.3649	14.4222	11.4455	10.0995	9.94987
opossum				7.81025	8.7178	11.1803	15.8114
lemur					4.12311	6.16441	9.64365
mouse						4.12311	9.05539
rabbit							10.247

Thus, the similarities/dissimilarities for eight exon-1 gene sequences in Table 1 are given by considering the cosine of angles between all the vectors.

Observing Table 10, we see again the dissimilarity of gallus to the others among the considered eight species because their corresponding row has little entries. Also, the more similar species pairs are human-rabbit, human-mouse, mouse-rat, lemur-rabbit, and mouse-rabbit, which coincides with the result in Table 9. Similar results have been obtained by Randic,[1,4] where similarity is based on the occurrence of the pairs and triplets of nucleic acid bases in these DNA primary sequences.

It is natural that the mouse-rabbit, mouse-rat pairs have similarity. As discussed by Randic,[3] however, the similarity of these pairs such as human-rabbit, human-mouse, and lemur-rabbit is either an artifact reflecting deficiency of the sequence invariants, or indeed the sequences may have visible similarity even though the corresponding species are not closely related in the evolutionary sense.

Similarly, for the leading eigenvalues of 2 × 2 matrices listed in Table 8, we take them as six-component vectors and offer a similarity analysis for eight exon-1 sequences in Table 1. In Tables 11 and 12, we list the Euclidean distance between the end points of the six-component vectors and the magnitudes of the cosine of angles between any all two six-component vectors.

Again, we see that gallus has the small similarity to others. This is not surprising because gallus is the only nonmammalian species among these considered species. The mouse-rat pair is still the most similar two species because its value is the smallest entry in Table 11 and the largest in Table 12. However, some disappointed values still occur, values such as the human-mouse and human-rat are found in the small entry in Table 11 and the in large value in Table 12. The reason for making these disappointed results may be as follows: (1) We make comparison on eight species only on a single gene. Each species's genome is very long and contains many exon so that the information of the species could be preserved in every exon other than one of them. (2) Information extracted in each gene sequence is not enough plenteous to comparison of eight species.

CONCLUDING REMARKS

It is well-known that the alignments of DNA sequences are computer intensive that is direct comparison for DNA sequences. Structure considered in alignment of DNA sequences is only string's structures. Here, we use an intensive approach which shall consider not only sequences' structure but also chemical structure for DNA primary sequences. The invariant of sequences is applied to compare of DNA primary sequences, rather than sequences themselves. Furthermore, considering the frequency of occurrence of (0,1) triplets in three characteristic sequences as the elements of matrix are free from the lengths of DNA primary sequences when different length of DNA sequence are compared.

It is interesting that comparing the characteristic sequences might provide a possibility to reveal the biological functions of purine-pyrimidine, amino-keto groups, and weak-strong H-bonds, respectively. For example, to compare the eight sequences in Table 1, we take data in Table 5−7 as six-component vectors, respectively. We compute the Euclidean distance between the end points of the six-component vectors in Tables 13−15, to offer a similarity analysis for eight exon-1 sequences in Table 1.

Observing Table 13−15, results of the similarities for the eight sequences in Table 1 do not coincide each other. The differences just reflect the efficiency of the classifications of the bases.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for many valuable suggestions that have improved this manuscript. This work is supported in part by the National Natural Science Foundation of China.

REFERENCES AND NOTES

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

(2) Leong, P. M.; Morgenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Applic. Biosci.* **1995**, *12*, 503−511.

(3) Randic, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50−56.

(4) Randic, M. On characterization of DNA primary sequence by a condensed matrix. *Chem. Phys. Lett.* **2000**, *317*, 29−34.

(5) Randic, M.; Guo, X. F.; Basak, S. C. Characterization of DNA based on occurrence of triplets of nucleic bases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 619−626.

(6) Randic, M.; Vracko, M. On the similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599−606.

(7) Randic, M.; Vracko, M.; Nandy, A.; Basak, S. C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1235−1244.

(8) Raychaudhury, C.; Nandy, A. Indexing scheme and similarity measures for macromolecular sequences. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243−247.

(9) Zhang, R.; Zhang C. T. Z-curve, An intutive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Stru. Dyn.* **1994**, *11 (4)*, 767−782.

(10) Zhang, C. T. A symmetrical theory of DNA sequences and its application. *J. Theor. Biol.* **1997**, *187*, 297−306.

CI010131Z