Inter*net* Electronic Journal of Molecular Design

December 2002, Volume 1, Number 12, Pages 668–674

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 4

Guest Editor: Jun–ichi Aihara

Numerical Characterization of DNA Primary Sequence

Ping–an He[1] and Jun Wang[1,2]

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

Received: August 1, 2002; Revised: September 4, 2002; Accepted: November 3, 2002; Published: December 31, 2002

Citation of the article:
P. He and J. Wang, Numerical Characterization of DNA Primary Sequence, *Internet Electron. J. Mol. Des.* **2002**, *1*, 668–674, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Numerical Characterization of DNA Primary Sequence[#]

Ping–an He[1,*] and Jun Wang[1,2]

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

Received: August 1, 2002; Revised: September 4, 2002; Accepted: November 3, 2002; Published: December 31, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (12), 668–674

Abstract

In a previous paper, the authors defined a numerical characterization of DNA primary sequences. A DNA primary sequence was reduced to a few of binary sequences, based on the classifications of the four nucleic acid bases. The reduced sequences are called the characteristic sequences. For each characteristic sequence, we associated two 2×2 matrices, the elements of which are given by the frequency of occurrence of all (0,1) triplets in the characteristic sequence. In this paper, we use eigenvalues of the new matrices to characterize the biological functions of purine–pyrimidine, amino–keto groups and weak–strong H–bonds, respectively.

Keywords. DNA primary sequence; characteristic sequences; leading eigenvalue; similarity; dissimilarity.

1 INTRODUCTION

With the imminent completion of the Human Genome Project and the fast increase of many complete genomes of prokaryote and eukuaryote, fundamental questions regarding the characteristics of these sequences arise, the first of which is how to compare genomes. Hence analysis and understanding of the 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. In previous research, the comparisons of DNA primary sequences are mainly to consider the alignment of the DNA primary sequences. The alignment of sequences is performed by the computer to find the smallest number of changes (deletions, insertions, substitutions, shifts) that are necessary to match labels in two DNA primary sequences.

[#] Dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday.
* Correspondence author; E–mail: pinganhe@yahoo.com.cn.

Some researchers consider graphical representations for DNA primary sequences [1–17], in particular, Gates [1], Hamori and Ruskin [2], Leong and Morgenthaler [4], Randić [9–13], Nandy [5–8,13–15], Zhang [16,17] and others, considering a real DNA primary sequence as a curve embedded in 2–D plane or 3–D space. Using research on the graphical representations, we can derive some numerical characterization for DNA primary sequences.

An alternative approach of the comparisons for DNA sequences is suggested by Randić *et al*. [9–13], who considered mathematical invariants of DNA primary sequences rather than the sequences themselves. For chemical structure and chemical graphs we can in fact obtain numerous invariants that are applied for characterization and comparison of structures. There are hundreds of topological indices that have been used in structure–property–activity studies based on molecular graphs. In this way, we can arrive at invariants for DNA primary sequences to associate a matrix with a DNA primary sequence. Once a matrix representation of sequences is given, one can consider suitable matrix invariants as invariants of the comparison of DNA sequences.

In a previous paper [3], we introduced another representation for DNA sequences, which is based on the idea of the coarse–grained description of the DNA primary sequence: we classify the four nucleic acid bases into two groups, purine–pyrimidine, amino–keto groups and weak–strong H–bonds, respectively, and then label the bases of purine, amino and weak H–bonds by 1, and the bases of pyrimidine, keto and strong H–bonds by 0, respectively. Thus, from a DNA primary sequence we obtained three (0,1)–sequences, which are called the characteristic sequences of the DNA primary sequence. For each characteristic sequence we constructed a set of 2×2 matrices, which are based on counting of the frequency of occurrence of all (0,1) triplets of the characteristic sequence. The leading eigenvalues of these matrices are computed and considered as invariants for the comparison of DNA primary sequences.

Table 1. Exon–1 of the β–Globin genes for Eight Species.

Species	Sequence	Length
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTG GGGCAAGGTGAACGTGGAGTAAGTTGGTGGTGAGGCCCTGGGCAG	92
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGG CAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG	86
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTG GGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG	92
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTG GTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG	92
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTG GGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG	92
Mouse	ATGGTGCACCTGACTGATGCTGAGAAGGCTGCCGTTACTGCCCTGTG GGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG	93
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGT GGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC	90
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTG GGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG	92

In this paper, through comparison of characteristic sequences we try to find the biological functions of purine–pyrimidine, amino–keto groups and weak–strong H–bonds, respectively. In Table 1, the exon–1 of the β–globin gene for eight species are listed, which were reported by Randić [9].

2 CONSTRUCTION OF THE CHARACTERISTIC SEQUENCES

Nucleic acids and proteins are all linear macromolecules. However, comparison of DNA primary sequences should be considered not only the string 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: purine R = {A, G} and pyrimidine Y = {C, T}, or 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}.

Let $S = a_1 a_2 a_3 \cdots$ be a DNA primary sequence. Using above classifications, we can transform a DNA primary sequence into three (0,1) sequences by three homomorphic maps ϕ_i, $i = 1, 2, 3$, $\phi_i(S) = \phi_i(a_1)\phi_i(a_2)\cdots\cdots$, as follows:

$$\phi_1 = \begin{cases} 1 & \text{if} \quad a_i \in R \\ 0 & \text{if} \quad a_i \in Y \end{cases}$$

$$\phi_2 = \begin{cases} 1 & \text{if} \quad a_i \in M \\ 0 & \text{if} \quad a_i \in K \end{cases}$$

$$\phi_3 = \begin{cases} 1 & \text{if} \quad a_i \in W \\ 0 & \text{if} \quad a_i \in S \end{cases}$$

Thus, we obtain three (0,1) sequences corresponding to the same DNA primary sequence, and we call them as (R,Y)–, (M,K)– and (W,S)–characteristic sequences of the DNA primary sequence, respectively.

In [3], we constructed a set of 2×2 matrix and computed their leading eigenvalues for three characteristic sequences. Using the leading eigenvalues, we compared the similarities and dissimilarities for eight species in Table 1. The results in [3] coincide with the result of Randić's papers. It demonstrates that the comparison of three characteristic sequences is the same as the comparison of DNA primary sequences.

As we have seen in [3], the three characteristic sequences contain all information of the primary sequence. On the other hand, each characteristic sequence is a coarse–grained description for the DNA primary sequence, *i.e.*, some information for DNA primary sequence may be lost in a characteristic sequence so that different DNA primary sequences may have certain similar characteristic sequences. This just reflects the functions of the classifications. Therefore, comparing

each characteristic sequences has special significance. In Table 2, we list the characteristic sequences of the eight DNA sequences of Table 1. In the next section we will compare each characteristic sequence and get some conclusions that cannot be obtained from direct comparison of DNA primary sequences.

Table 2. Characteristic Sequences of the Eight DNA Sequence from Table1

human 101101010001100000111111111000100100100100001011110111101110101111011100110110111100001110111 10000011110011011001001011001001100011001110000000111000011100001001100000000000100111000011O 1100100100101010010100101101010000111010000101000001100101100100101110110010010100000100001O
goat 1010011001001111111111001001001001100000111101111011111011101111001101001111000011101 1 100100110010010010110010011001111001001000001110000111000010011000000001001001110000110 110010101001010010110001000010100000110100000110010111010011011011001001010000001000010
gallus 10110101001110010011111111101100010010011000001111011110011010110011101011110011110000110011 1000001110001100100100101101101011011110011010000011100011100000110110000000110110111000111O 1100100101001010010100101100100101101000000010100000110010111010000011101000000011000010000 10
opossum 101101010001100000111111111110010100100100100011000111101011100110011100110110111100000110 11 100000111000110010010010110111001101110111101000010110000110000011110110000000010011100001 10 1100100101101011010100101101101001101011001101001011100100100110100101010010010100000110001 0
lemur 10110000100111010011111111010001010010000000010111101111011 101011111111001101101111000011101 1 1001100001001000100100101100101100011110101000000001110000010001010111000000000101001100000110 1101011100101010010100101110010110101001010101000001100100110110101110110010000100001100001 0
mouse 1011010100011001101001111111001001001001000010111101111011101011101111001101101111000011101 11 1000001111001100100100101100100110001100111000000011100001110000100110000000000010011100001100 11001001001010101100101011000100001110100001010000011001011001001101101100100101000000100001 00
rabbit 10110101000100011011111111110001011001001000010111101111011101011111111110011011011110000111 0 10000011010001110001001011001001000111001110000000111000011000001101100000000000100111000 01 11001001101010010101001011010100001010100001010000011001 011101001101101100100101000001000 0
rat 101101010001100110100111111100100100110110001011111111011 1000011011010011010011110000111011 1000001111011100100100101100101100001000011000000011100001111100101100000001010010011100001 10 110010010011101011001010110001101011101000010100001110010110001011111011000001010000010000 10

3 COMPARISON OF CHARACTERISTIC SEQUENCES

In [3], we also introduced a 2×2×2 cubic matrix with 8 entries $f_{ijk}^{X} = \left(100 m_{ijk}^{X}\right) / \left(N-2\right)$, where m_{ijk}^{X} is the enumeration of the (0,1) triplet *ijk* in characteristic sequence *X* and *N* is the length of *X*. Clearly, it represents the 100 times of 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. 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_0^X and F_1^X, where

$F_0^X = \left(f_{0jk}^X\right)$ and $F_1^X = \left(f_{1jk}^X\right)$ with X being R, M or W. In [3], we computed the leading eigenvalues condensed matrices as above. In Table 3, all leading eigenvalues of characteristic sequences are listed, as reported in [3].

Table 3. Leading eigenvalues of the 6 matrices F_0^X and F_1^X for the eight DNA sequences of Table 1.

Species	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	30.0	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	23.0	26.6	27.6	21.8	26.6	25.0
Lemur	21.1	30.3	33.2	20.4	26.5	22.9
Mouse	21.6	29.6	31.5	19.9	29.6	23.4
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

For each characteristic sequence, we take the leading eigenvalue as a two–dimensional vector (F_0^x, F_1^X), by which we compare the (R,Y)–, (M,K)– and (W,S)–characteristic sequences of DNA primary sequences based on the Euclidean distance between the end points of the two–dimensional vectors, respectively. The results of comparisons are listed on the three tables, where Table 4 reveals the information of purine–pyrimidine group, Table 5 the information of amino–keto group, and Table 6 the information of weak–strong H–bonds, respectively.

Table 4. Similarity/dissimilarity table for the eight DNA sequences of Table 1 based on their (R,Y) characteristic sequences.

Species	Goat	Gallus	Opossum	Lemmur	Mouse	Rabbit	Rat
Human	2.36008	1.14018	2.64197	1.52315	2.86007	3.04631	1.25300
Goat		3.47131	4.99300	0.94339	5.22015	0.92195	3.53553
Gallus			1.64924	2.58070	1.78885	4.08167	0.22361
Opossum				4.15933	0.40000	5.67979	1.71172
Lemmur					4.3566	1.52315	2.62488
Mouse						5.86686	1.80278
Rabbit							4.10488

Table 5. Similarity/dissimilarity table for the eight DNA sequences of Table 1 based on their (M,K) characteristic sequences.

Species	Goat	Gallus	Opossum	Lemmur	Mouse	Rabbit	Rat
Human	3.57351	6.8593	5.02096	2.83196	1.71172	3.41321	3.16228
Goat		3.99249	2.50200	3.26956	2.14009	5.46443	0.41231
Gallus			1.84391	7.26154	5.93633	9.40213	4.24853
Opossum				5.67539	4.20476	7.69675	2.64764
Lemmur					1.74642	2.34307	3.03645
Mouse						3.49285	1.76918
Rabbit							5.16236

Observing Tables 4, 5 and 6, we can obtain some information for each characteristic sequence. For example, the species gallus is the most dissimilarly with others in Table 6. However, we do not see the same result from Table 4 and 5, even the value of gallus–rat pair is the least in Table 4. The

species gallus is the only non–mammalian species among these considered species. Whether or not this means that the essential nature of the mammalian species may be revealed mainly in the characteristic sequence of weak–strong H–bonds group. On the other hand, the results in Tables 5 and 6 are very similar to that of the comparison of DNA primary sequences. This means that the information of the similarities for eight sequences may contain mainly in the reduce sequences of amino–keto groups and weak–strong H–bonds groups.

Table 6. Similarity/dissimilarity table for the eight DNA sequences of Table 1 based on their (*W*,*S*) characteristic sequences.

Species	Goat	Gallus	Opossum	Lemmur	Mouse	Rabbit	Rat
Human	0.894427	3.53412	4.40454	3.56931	1.96977	0.80000	1.80278
Goat		2.88617	5.23450	4.17852	2.72029	1.44222	2.37697
Gallus			7.87718	7.05195	5.50364	4.2720	5.24976
Opossum				2.10238	2.56125	3.82099	3.13847
Lemmur					1.70294	2.78927	1.80278
Mouse						1.28062	0.70000
Rabbit							1.00499

Furthermore, we can observe the least value in each table: the gallus–rat pair in Table 4, the goat–rat pair in Table 5, and the mouse–rat pair in Table 6, respectively. Whether these results imply that the three characteristic sequences reflect some intrinsic essence of species rat from different aspect.

Generally, we can also observe the least value of all species in each table, so that we can obtain information of the (*R*,*Y*)–, (*M*,*K*)– and (*W*,*S*)–characteristic sequences, respectively. For example, the mouse species, the least value in Tables 4–6 are the opossum, human and rat, respectively. These results illuminate that the three characteristic sequences reflect some essence of mouse species from different aspect.

4 CONCLUSIONS

Comparing characteristic sequences, we can get some information that cannot be obtained from the direct comparison of DNA primary sequences and observe some special nature in species from different aspect. Although some information may be lost in characteristic sequences, we can focus our attention on the information of our interest. This is the advantage of our approach.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China.

5 REFERENCES

[1] M. A. Gates, A Simple way to look a DNA, *J. Theor. Biol*. **1986**, *119*, 319–328.
[2] E. Hamori and J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem*. **1983**, *258*, 1318–1327.

[3] P. He and J. Wang, Characteristic sequences for DNA primary sequence, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1080–1085.
[4] P. M. Leong and S. Morgenthaler, Random walk and gap plots of DNA sequences, *Comput. Applic. Biosc.* **1995**, *11*, 503–507.
[5] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, *Curr. Sci.* **1994**, *66*, 309–314.
[6] A. Nandy and P. Nandy, Graphical analysis of DNA sequences structure: II. Relative abundance of nucleotides in DNAs, gene evolution and duplication, *Curr. Sci.* **1995**, *68*, 75–85.
[7] A. Nandy, Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons, *Curr. Sci.* **1996**, *70*, 661–668.
[8] A. Nandy, P. Nandy, and S. C. Basak, Quantitative Descriptor for SNP Related Gene Sequences, *Internet Electron. J. Mol. Des.* **2002**, *1*, 367–373, http://www.biochempress.com.
[9] M. Randić, Condensed Representation of DNA Primary Sequences, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50–56.
[10] M. Randić and M. Vračko, On the Similarity of DNA Primary Sequences, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599–606.
[11] M. Randić, On characterisation of DNA primary sequences by a condensed matrix, *Chem. Phys. Lett.* **2000**, *317*, 29–34.
[12] M. Randić, X. Guo, and S. C. Basak, On the characterization of DNA primary sequences by triplet of nucleic acid bases, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 619–626.
[13] M. Randić, M. Vračko, A. Nandy, and S. C. Basak, On 3–D representation of DNA primary sequences, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235–1244.
[14] C. Raychaudhury and A. Nandy, Indexing scheme and similarity measures for macromolecular sequences, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243–247.
[15] A. Nandy, Investigations on Evolutionary Changes in Base Distributions in Gene Sequences, *Internet Electron. J. Mol. Des.* **2002**, *1*, 545–558, http://www.biochempress.com.
[16] R. Zhang and C. T. Zhang, Z–curve, An Intuitive Tool for Visualizing and Analyzing the DNA sequences, *J. Biomol. Str. Dyn.* **1994**, *11*, 767–782.
[17] C. T. Zhang, A Symmetrical Theory of DNA sequences and Its Application, *J. Theor. Biol.* **1997**, *187*, 297–306.

Biographies

Ping–an He is a PhD student of Applied Mathematics at the Dalian University of Technology. His main research interests include combinatorics, graph theory and bioinformatics.

Jun Wang is a Professor of Applied Mathematics at the Dalian University of Technology, the advisor of the first author.