New Invariant of DNA Sequences

Chun Li*[,†,‡] and Jun Wang[‡,§]

Department of Mathematics, Bohai University, Jinzhou 121000, P. R. China, and Department of
Applied Mathematics and College of Advanced Science and Technology, Dalian University of Technology,
Dalian 116024, P. R. China

Received April 14, 2004

For a DNA sequence with n bases, one can always associate it with an $n \times n$ nonnegative real symmetric matrix whose diagonal entries are zero. Once the matrix is given, its leading eigenvalue is usually calculated and used as an invariant to characterize the DNA sequence. Let M be such a matrix, and λ_1 its leading eigenvalue. Then $(1/n)||M||_{m1}$ and $\sqrt{(n-1)/n}||M||_F$ are the lower and upper bounds of λ_1, respectively. Since their arithmetic average is an approximate value of λ_1 and simpler for calculation, we can use it as an alternative invariant to characterize the DNA sequence. The utility of the new parameter is illustrated on the DNA sequences of five species: human, chimpanzee, mouse, rat, and gallus.

1. INTRODUCTION

Compilation of DNA primary sequence data continues unabated and tends to overwhelm us with voluminous outputs that increase daily. Comparison of different DNA primary sequences remains one of the important aspects of the analysis of DNA data banks. Usual representation of a DNA primary sequence is that of a string of letters A, G, C and T, which signify the four nucleic acid bases adenine, guanine, cytosine, and thymine, respectively. How similar/dissimilar the different sequences are may depend on how such strings of letters are encoded or characterized. The previous procedures consider differences between strings due to deletion−insertion, compression−expansion, and substitution of the string elements. These approaches, which have been hitherto widely used, are computer intensive.[1,2] Recently, an alternative approach for the comparison of sequences has been introduced. It is based on characterization of DNA by ordered sets of invariants derived from DNA sequences, rather than by a direct comparison of DNA sequences themselves.[1−9] However, as pointed in ref 2, this approach involves a number of as yet unresolved questions. In particular, questions that need our attention are as follows: how to obtain suitable invariants to characterize DNA sequences and how to select invariants suitable for sequence comparisons.

A sequence invariant is usually a real number that is independent of the labels (bases) A, G, C, and T. For example, the length of the sequence is an invariant. But the length cannot capture the main information of the sequence considered, so it is regarded as a trivial invariant. Among other sequence invariants, the leading eigenvalue of a matrix associated with a DNA sequence is an important invariant and is proved to be highly effective for characterization of DNA sequences. However, a problem we must face is that the calculation of the eigenvalue will become more and more difficult with the order of the matrix large. Are there any other suitable descriptors for DNA sequences? In this paper, we propose a new sequence invariant for the characterization of DNA sequences. Its definition is as follows:

Given a DNA sequence with n bases, we can always associate it with an $n \times n$ nonnegative real symmetric matrix whose diagonal entries are zero (see refs 1, 3−6, 8, 10, 11). Let $M = (a_{ij})_{n\times n}$ be such a matrix, i.e., $a_{ij} \geq 0$, $a_{ij} = a_{ji}$, and $a_{ii} = 0$ for $i,j = 1, 2, ..., n$. Define

$$\chi = \chi(M) = \frac{1}{2}\left(\frac{1}{n}||M||_{m1} + \sqrt{\frac{n-1}{n}}||M||_F\right) \qquad (1)$$

where $||M||_{m1} \equiv \sum_{i,j}|a_{ij}|$ and $||M||_F \equiv \sqrt{\sum_{i,j}|a_{ij}|^2} = \sqrt{\mathrm{tr}(M^TM)}$ (here tr$M$ denotes the trace of M).

In fact, if we suppose λ_1 is the leading eigenvalue of M, then it is not difficult to prove that

$$\frac{1}{n}||M||_{m1} \leq \lambda_1 \leq \sqrt{\frac{n-1}{n}}||M||_F \qquad (2)$$

Here the first inequality is proved in ref 12 (see also ref 13) and the second in ref 14, p 13. Notice that both the bounds in (2) are attained. For example, if

$$M = \begin{pmatrix} 0 & k \\ k & 0 \end{pmatrix}$$

where $k \geq 0$, then the two bounds coincide. So $\chi(M)$ can be regarded as an approximation of the leading eigenvalue of M. It is just in this sense that we call $\chi(M)$ as 'ALE-index' for short. Clearly, $\chi(M)$ is simple for calculation and thus facilitated for characterization of DNA sequences. Its utility is illustrated on the *beta*, *gamma*, *epsilon*-globin, *neurogenin*, and *neuroD* genes of five species: human, chimpanzee, mouse, rat, and gallus (see Table 1).

2. COMPARISONS WITH THE LEADING EIGENVALUE

As we know, the ALE-index χ is not a matrix invariant, but it can always be obtained from any nonnegative real

* Corresponding author e-mail: lchlmb@yahoo.com.cn. Corresponding author address: Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, P. R. China.

[†] Department of Mathematics, Bohai University.

[‡] Department of Applied Mathematics, Dalian University of Technology.

[§] College of Advanced Science and Technology, Dalian University of Technology.

10.1021/ci049874l CCC: $30.25
Published on Web 12/08/2004

Table 1. Database Source

species	sequences	database	ID/ ACCESSION	location
human	*beta*-globin	EMBL	HSHBB	62187−63610
	epsilon-globin	EMBL	HSHBB	19289−20961
	gamma globin	NCBI	M91037	
	neuroD	NCBI	U50822	
	neurogenin 1	NCBI	NM_006161	
chimpanzee	*beta*-globin	EMBL	PTGLB1	4189−5532
mouse	*beta*-globin	EMBL	MMBGL1	275−1462
	neuroD	NCBI	NM_010894	
	neurogenin 1	NCBI	BC062148	
rat	*beta*-globin	EMBL	RNGLB	310−1505
	neuroD	NCBI	D82945	
	neurogenin	NCBI	U67777	
gallus	*beta*-globin	EMBL	GGGL02	465−1810
	epsilon-globin	EMBL	GGHBBRE	20349−21873
	neuroD	NCBI	AF060885	
	neurogenin 1	NCBI	AJ012660	

Table 2. Upper Triangles of a Symmetric Matrix[a]

0	1.000	1.000	0.942	0.787	0.809	0.831	0.623	0.594	0.616
	0	1.000	0.926	0.784	0.787	0.809	0.620	0.561	0.590
		0	1.000	0.962	0.784	0.787	0.641	0.526	0.563
			0	1.000	0.707	0.745	0.604	0.483	0.526
				0	1.000	1.000	0.528	0.489	0.530
					0	1.000	0.618	0.409	0.472
						0	1.000	0.316	0.410
							0	1.000	0.866
								0	1.000
									0

[a] Taken from ref 5, Table 1.

symmetric matrix with diagonal entries zero. For example, three matrices are given in Tables 2−4. Their ALE-indices, by eq 1, are calculated as 6.77435, 6.82042, and 7.476795, while their leading eigenvalues are 6.7028,[5] 6.7634,[10] and 7.4458,[11] respectively. It is easy to see that the ALE-index is slightly bigger than the corresponding leading eigenvalue.

The Q matrix is transformed from a graphical representation of a sequence, whose off-diagonal elements are defined as the quotient of the Euclidean distance between two vertices of the curve and the sum of geometrical lengths of edges between the two vertices. By definition all diagonal entries are zero.[7,11] Sometimes, it is also denoted by L/L to avoid confusion with other matrices defined in a similar way.[5,6,10] The Q matrix of Table 4 was obtained by assigning (1,1,1) to T and (1,0,0), (0,1,0), (0,0,1) to other three bases (A, C, G).[11] For convenience, we denote such a Q matrix by Q_T. Similarly, following the method introduced in our paper,[11] we also can get other three Q matrices Q_A, Q_C, and Q_G for the same sequence. In Table 5, we list the ALE-indices and leading eigenvalues of four Q matrices for the first 16, 39 bases, and the whole sequences of the exon-I of *beta*-globin genes of three species: human, mouse, and gallus. From Table 5, we see that Δ, which is always bigger than or equal to zero, increases as the length of the sequence increases; while δ, the relative error, reduces on the whole.

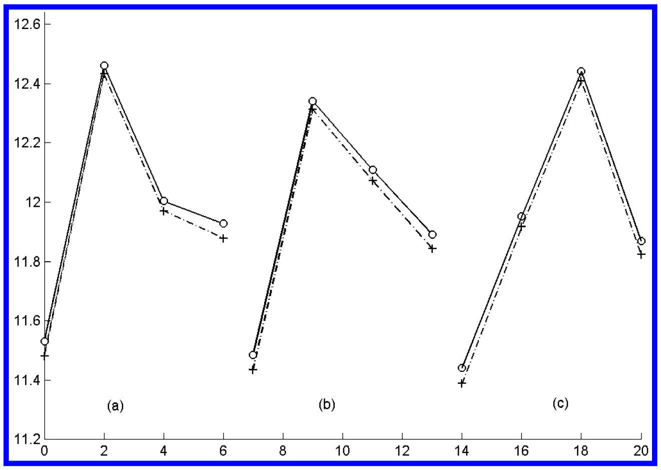

Figure 1. The ALE-indices VS leading eigenvalues. The first 16 bases: (a) human, (b) mouse, and (c) gallus.

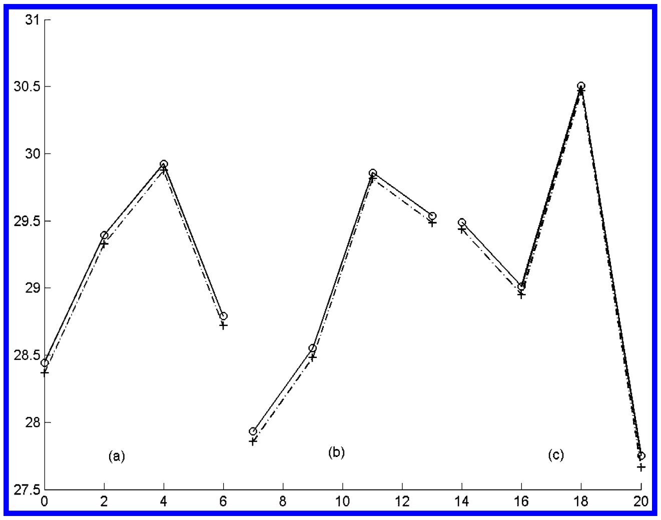

Figure 2. The ALE-indices VS leading eigenvalues. The first 39 bases: (a) human, (b) mouse, and (c) gallus.

In Figures 1−3, we show the plots of the ALE-indices of the Q matrices in Table 5 against the corresponding leading eigenvalues. The real line represents the ALE-index, while the dashed represents the leading eigenvalue. As a result, they only confirm that the new parameter of DNA appears to parallel to a great extent the leading eigenvalue.

Table 3. Upper Triangles of a Symmetric Matrix[a]

0	1.0000	0.8165	0.8077	0.6847	0.6792	0.6381	0.6145	0.6190	0.6450
	0	1.0000	0.8966	0.6720	0.6847	0.6792	0.6381	0.6390	0.6654
		0	1.0000	0.8966	0.6720	0.6847	0.6792	0.6381	0.6699
			0	1.000	0.5774	0.6383	0.6455	0.6000	0.6381
				0	1.0000	0.8165	0.6383	0.6455	0.6792
					0	1.0000	0.8165	0.6383	0.6847
						0	1.0000	0.5774	0.6720
							0	1.0000	0.8966
								0	1.0000
									0

[a] Taken from ref 10, Table 4.

Table 4. Upper Triangles of a Matrix Q Associated with the Sequence ATGGTGCACC[a]

Q	A	T	G	G	T	G	C	A	C	C
A	0	1.000	0.897	0.889	0.897	0.889	0.826	0.775	0.747	0.734
T		0	1.000	1.000	0.889	0.897	0.799	0.728	0.696	0.687
G			0	1.000	0.897	0.889	0.791	0.719	0.697	0.696
G				0	1.000	0.897	0.804	0.732	0.719	0.728
T					0	1.000	0.707	0.577	0.612	0.663
G						0	1.000	0.707	0.745	0.791
C							0	1.000	0.707	0.745
A								0	1.000	1.000
C									0	1.000
C										0

[a] Taken from ref 11, Table 2.

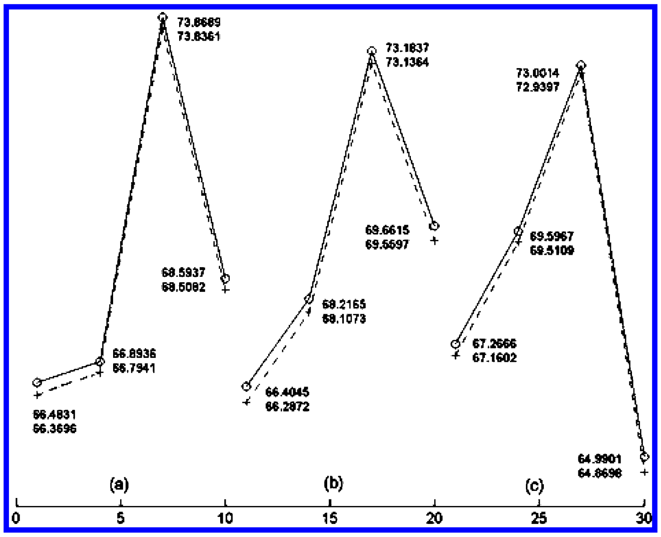

Figure 3. The ALE-indices VS leading eigenvalues. The whole sequences: (a) human, (b) mouse, and (c) gallus.

3. PROPERTIES

3.1. The Increasing Characteristic. Suppose $S = S_1S_2 \cdots S_n$ is a given DNA sequence. We append a base N (one of the four bases A, C, G, and T) to S and obtain a new DNA sequence $S^* = S_1S_2 \cdots S_nN$. Let us consider the ALE-index χ for sequence S^* via its Q matrix. Assigning (1,0,0), (0,1,0), (0,0,1), and (1,1,1) to four nucleic acid bases, we get a set of points:

$$P_1(x_1,y_1,z_1),\ P_2(x_2,y_2,z_2),\ \cdots,\ P_n(x_n,y_n,z_n),\ P_N(x,y,z)$$

From them, we construct a matrix $Q(S^*)$

$$Q(S^*) = \begin{pmatrix} & & & b_1 \\ & & & b_2 \\ & Q(S) & & \vdots \\ b_1 & b_2 & \cdots & \end{pmatrix}$$

where $Q(S)$ represents the Q matrix associated with the sequence S, and

$$b_i = \frac{|P_N - P_i|}{|P_i - P_{i+1}| + |P_{i+1} - P_{i+2}| + \cdots + |P_{n-1} - P_n| + |P_n - P_N|} \quad (i = 1, 2, ..., n)$$

Note that $0 < b_i \leq 1$, we have

$$\Delta\chi = \chi(Q(S^*)) - \chi(Q(S)) > 0$$

This implies that the ALE-index χ increases as the sequence extends. (This also can be seen from Table 5.)

3.2. The Degree of Continuity. Let $S = \cdots N_{i-1}N_i \cdots N_jN_{j+1} \cdots$. Then we call the string $B = N_i \cdots N_j$ as a "*block*" if $N_{i-1} \neq N_i = \cdots = N_j \neq N_{j+1}$, and $j-i$ the degree of continuity of this block (denoted by $dc(B)$). Thus, S can be represented in a "block" way: $B_1B_2 \cdots B_k$. We define the degree of continuity of the sequence S as follows:

$$dc(S) = \sum_{m=1}^{k} dc(B_m)$$

Clearly, n-dc(S) is just the number of blocks of S, where n is the length of the sequence S. For convenience, we denote it by NB(S), and the number of blocks with largest dc by $\mathrm{NB_L}(S)$.

Table 5. First Exons of *Beta*-Globin Genes of Three Species: Human, Mouse, and Gallus[a]

	first 16 bases				first 39 bases				whole sequence			
	χ	λ_1	Δ[b]	δ	χ	λ_1	Δ[b]	δ	χ	λ_1	Δ[b]	δ
human ATGGTGCACC TGACTCCTGA GGAGAAGTCT GCCGTTACTG CCCTGTGGGG CAAGGTGAAC GTGGATGAAG TTGGTGGTGA GGCCCTGGGC AG												
Q_A	11.5297	11.4812	0.0485	0.0042	28.4445	28.3648	0.0797	0.0028	66.4831	66.3696	0.1135	0.0017
Q_C	12.4599	12.4323	0.0276	0.0022	29.3916	29.3262	0.0654	0.0022	66.8936	66.7941	0.0995	0.0015
Q_G	12.0033	11.9703	0.0330	0.0028	29.9255	29.8740	0.0515	0.0017	73.8689	73.8361	0.0328	0.0004
Q_T	11.9270	11.8787	0.0483	0.0041	28.7920	28.7194	0.0726	0.0025	68.5937	68.5082	0.0855	0.0013
mouse ATGGTGCACC TGACTGATGC TGAGAAGTCT GCTGTCTCTT GCCTGTGGGC AAAGGTGAAC CCCGATGAAG TTGGTGGTGA GGCCCTGGGC AGG												
Q_A	11.4838	11.4349	0.0489	0.0043	27.9317	27.8567	0.0750	0.0027	66.4045	66.2872	0.1173	0.0018
Q_C	12.3398	12.3117	0.0281	0.0023	28.5527	28.4813	0.0714	0.0025	68.2165	68.1073	0.1092	0.0016
Q_G	12.1082	12.0729	0.0353	0.0029	29.8575	29.8118	0.0457	0.0015	73.1837	73.1364	0.0473	0.0007
Q_T	11.8919	11.8439	0.0480	0.0041	29.5363	29.4844	0.0519	0.0018	69.6615	69.5597	0.1018	0.0015
gallus ATGGTGCACT GGACTGCTGA GGAGAAGCAG CTCATCACCG GCCTCTGGGG CAAGGTCAAT GTGGCCGAAT GTGGGGCCGA AGCCCTGGCC AG												
Q_A	11.4416	11.3890	0.0525	0.0046	29.4872	29.4359	0.0513	0.0017	67.2666	67.1602	0.1064	0.0016
Q_C	11.9522	11.9170	0.0352	0.0029	29.0093	28.9500	0.0593	0.0020	69.5967	69.5109	0.0858	0.0012
Q_G	12.4406	12.4077	0.0329	0.0027	30.5050	30.4677	0.0373	0.0012	73.0014	72.9397	0.0617	0.0009
Q_T	11.8694	11.8222	0.0472	0.0040	27.7517	27.6655	0.0862	0.0031	64.9901	64.8698	0.1203	0.0019

[a] ALE-index χVS leading eigenvalue λ_1. [b] $\Delta = \chi - \lambda_1$, $\delta = \Delta/\lambda_1$.

Table 6. Normalized ALE-Indices of Matrices Q_A, Q_C, Q_G, and Q_T for *Beta*-Globin Genes of Five Species: Human, Chimpanzee, Mouse, Rat, and Gallus

species		human	chimpanzee	mouse	rat	gallus
CDs	χ'_A	0.706851	0.705343	0.720117	0.721493	0.715427
	χ'_C	0.742856	0.736867	0.747247	0.737042	0.776458
	χ'_G	0.763747	0.768885	0.751660	0.748930	0.742461
	χ'_T	0.734740	0.740234	0.726074	0.736247	0.719555
Introns	χ'_A	0.773660	0.773118	0.744233	0.744911	0.746021
	χ'_C	0.706172	0.705246	0.712964	0.716074	0.716587
	χ'_G	0.699275	0.700753	0.720568	0.713307	0.792600
	χ'_T	0.815062	0.817109	0.809794	0.799505	0.712577
Whole sequences	χ'_A	0.753499	0.753763	0.732444	0.734419	0.733348
	χ'_C	0.710812	0.709336	0.720937	0.719882	0.730350
	χ'_G	0.712307	0.714382	0.725045	0.720675	0.776980
	χ'_T	0.793699	0.795884	0.782648	0.779037	0.713454

For two DNA sequences S_1 and S_2 with n bases, we say the degree of continuity of S_1 is higher than that of S_2 if one of the following conditions holds.

(1) $dc(S_1) > dc(S_2)$;

(2) $dc(S_1) = dc(S_2)$, but, either $max\{dc(B_{1i})\} > max\{dc(B_{2j})\}$ or $max\{dc(B_{1i})\} = max\{dc(B_{2j})\}$ but $dc(S_1\backslash rB) > dc(S_2\backslash rB)$.

Here B_{1i} (i =1, 2, ..., k_1) and B_{2j} (j = 1, 2 ..., k_2) are the blocks of S_1 and S_2, respectively. $r = \min(NB_L(S_1), NB_L(S_2))$. $Sg\backslash rB$ (g = 1, 2) represents the sequence remained by removing r blocks with largest dc from Sg.

It is easy to see that the higher the $dc(S)$, the larger the $\chi(^bQ(S))$, and vice versa. Where

$$^bQ(S) = \lim_{t\to+\infty} {}^tQ(S)$$

is a (0,1)-matrix, and $^tQ(S)$ is the product of Hadammard multiplication of the matrix $Q(S)$ by itself t-times.

3.3. The Maximum and Minimum. Suppose the length of a DNA sequence S is n. Then we have the following:

(1) $\chi(Q_A(S)) = \chi(Q_C(S)) = \chi(Q_G(S)) = \chi(Q_T(S)) = n - 1$ if and only if $dc(S) = n - 1$, that is, S consists of a solo nucleotide acid, say $S = AA \cdots A$. Moreover, $n - 1$ is the maximum of χ's for all DNA sequences with n bases.

(2) If $dc(S) = 0$, namely $NB(S) = n$, then ^bQ, the limit of the matrices sequence $\{^tQ(S)\}$, is as follows:

$$^bQ = \begin{pmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & 0 \end{pmatrix}$$

Clearly, $\chi(^bQ) = (1/n + \sqrt{1/2n})(n - 1)$, and this is a lower bound of χ's.

In a word, for a DNA sequence S with n bases, we have

$$\left(\frac{1}{n} + \sqrt{1/2n}\right)(n - 1) \le \chi(Q(S)) \le n - 1$$

The subsections 3.2 and 3.3 show that one can not only obtain information on the form of a DNA sequence from χ but also approximately compare the χ-values of two sequences from themselves. The leading eigenvalue method seems to have no such an advantage.

4. APPLICATION

The ALE-index is very simple for calculation so that it can be directly used to deal with long DNA sequences. To reduce variations caused by a different length of sequences, one can consider the normalized ALE-index, i.e. $\chi' = \chi/n$, where n is the length of the sequence and the order of the corresponding matrix as well. Taking the whole sequence of **M*91037*** (Homo sapiens G-*gamma* globin and A-*gamma* globin genes, NCBI) as an example, although its length is 11393 bp, its four normalized ALE-indices corresponding

Table 7. Normalized ALE-Indices of Matrices Q_A, Q_C, Q_G, and Q_T for *Epsilon*-Globin Genes of Human and Gallus

species		CDs	introns	whole sequences
human	χ'_A	0.722906	0.782308	0.761616
	χ'_C	0.739140	0.677548	0.696289
	χ'_G	0.746910	0.731284	0.731807
	χ'_T	0.735290	0.774535	0.758863
gallus	χ'_A	0.718527	0.743595	0.734998
	χ'_C	0.768351	0.724440	0.733497
	χ'_G	0.749461	0.750527	0.749797
	χ'_T	0.714530	0.725275	0.721668

Table 8. Normalized ALE-Indices of Matrices Q_A, Q_C, Q_G, and Q_T for CDs of *Neurogenin* and *NeuroD* Genes

species		human	mouse	rat	gallus
neurogenin	χ'_A	0.708583	0.699440	0.697353	0.709295
	χ'_C	0.811860	0.801830	0.801563	0.829484
	χ'_G	0.783072	0.772424	0.777495	0.778177
	χ'_T	0.687691	0.702011	0.701598	0.704637
neuroD	χ'_A	0.736484	0.742439	0.743348	0.722028
	χ'_C	0.770810	0.774895	0.769644	0.809090
	χ'_G	0.745069	0.742812	0.741603	0.784720
	χ'_T	0.703437	0.702002	0.704582	0.680224

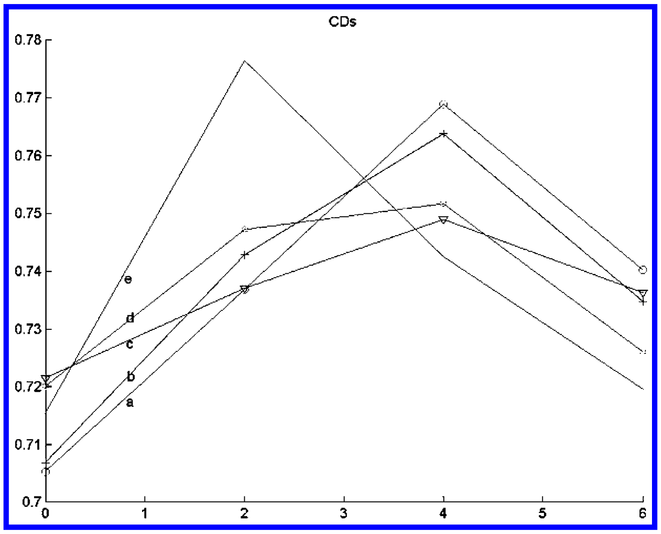

Figure 4. The plots of normalized ALE-indices for CDs of *beta*-globin genes of five species: (a) chimpanzee, (b) human, (c) rat, (d) mouse, and (e) gallus.

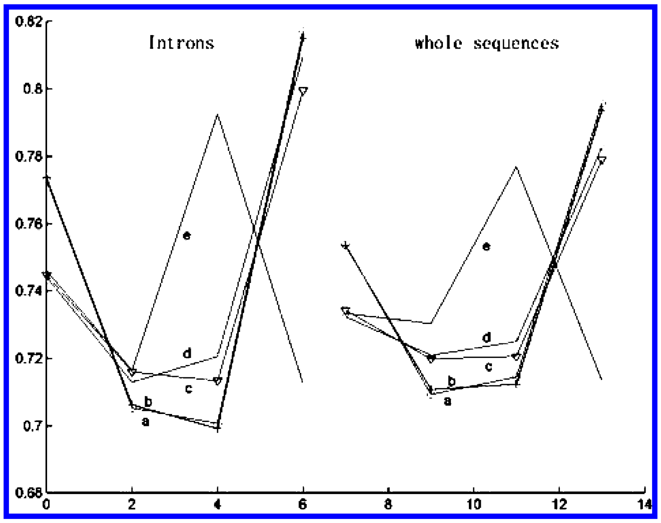

Figure 5. The plots of normalized ALE-indices for introns and the whole sequences of *beta*-globin genes of five species: (a) chimpanzee, (b) human, (c) rat, (d) mouse, and (e) gallus.

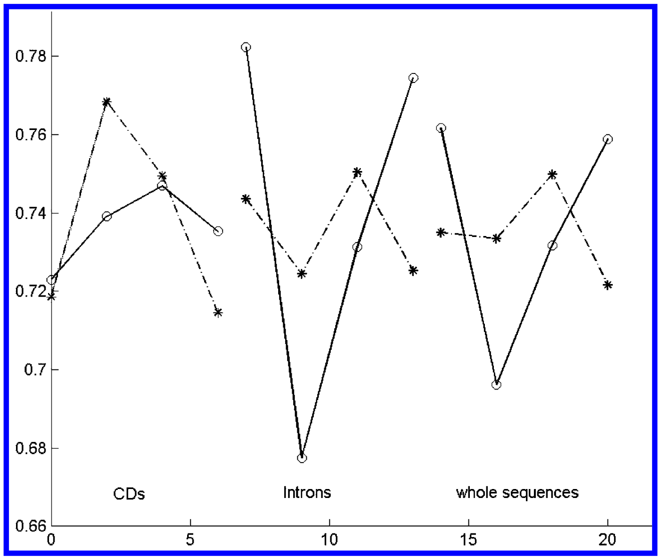

Figure 6. The plots of normalized ALE-indices for *epsilon*-globin genes of human (real line) and gallus (dashed line).

to four matrices Q_A, Q_C, Q_G, and Q_T are easily calculated as $\chi'_A = 0.757477$, $\chi'_C = 0.704654$, $\chi'_G = 0.722373$, and $\chi'_T = 0.754462$, respectively. In Table 6, we list the normalized ALE-indices for *beta*-globin genes of five species: human, chimpanzee, mouse, rat, and gallus. We also list the corresponding values for *epsilon*-globin genes of human and gallus in Table 7 and for *neurogenin* and *neuroD* genes of human, mouse, rat, and gallus in Table 8. From them, one finds that, except gallus, the only nonmammal among the five species, χ'_A and χ'_T of the CDs are roughly smaller than that of the introns, but χ'_C and χ'_G are larger.

In Figures 4−7, we show the plots of normalized ALE-indices for the five species in Tables 6−8. From them, one sees that, for each of the mammals, the four normalized ALE-indices of the CDs form a convex curve ('∩'), whereas that of the introns, even the whole globin genes, shape concave curves ('∪'). While for gallus, the curve of normalized ALE-indices of the CDs is convex, but that of the introns and the whole *beta* and *epsilon*-globin genes are no longer concave curves. The fact that gallus is a nonmammal while all others are mammals in the above tables might be a reason for this very different result. Figures 4, 5, and 7 also show that the sequences of human and chimpanzee are similar, so are mouse and rat, while gallus has great dissimilarity with others. This is analogous to the results reported by other authors.[3,6,9,15]

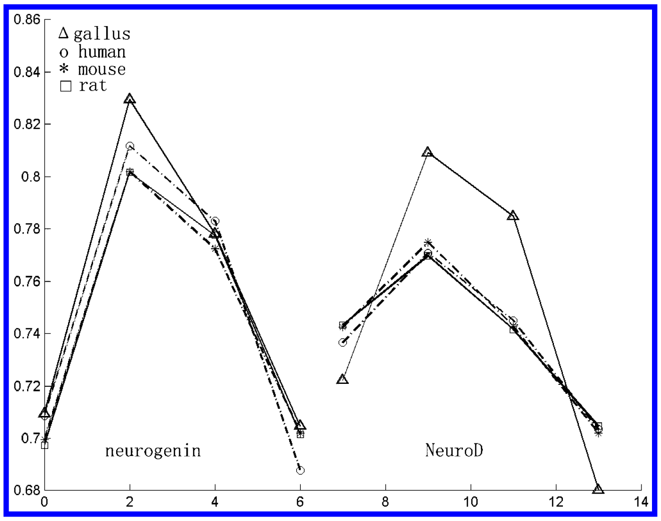

Figure 7. The plots of normalized ALE-indices for CDs of *neurogenin* and *neuroD* genes of human, mouse, rat, and gallus.

5. CONCLUSION

The 'invariant approach' has provided us with a powerful tool for characterization and comparison of DNA sequences. However, as pointed in ref 2, how to obtain suitable invariants to characterize DNA sequences and how to select invariants suitable for sequence comparisons are still questions that continue to need our attention. In this paper, we propose a new sequence invariant named 'ALE-index'. The new parameter is very simple for calculation so that it can be directly used to handle long DNA sequences. Taking the *beta*, *gamma*, *epsilon*-globin, *neurogenin*, and *neuroD* genes of five species: human, chimpanzee, mouse, rat, and gallus, as examples, we calculate the corresponding normalized ALE-indices for these sequences and their CDs and introns. As a result, the curves formed of corresponding normalized ALE-indices of the four mammals show the same regularity on the whole, while gallus, the only nonmammal among the five species, shows a very different result. Moreover, among the four mammals, we find the sequences of human and chimpanzee are similar, so are mouse and rat. These results are similar to that reported in other literature.

ACKNOWLEDGMENT

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

REFERENCES AND NOTES

(1) 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.

(2) Randic, M.; Guo, X.; Basak, S. C. On the characterization of DNA primary sequences by triplet of nucleic acid bases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 619−626.

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

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

(5) Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.* **2003**, *368*, 1−6.

(6) Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.* **2003**, *371*, 202−207.

(7) Bajzer, Z.; Randic, M.; Plasic, D.; Basak, S. C. Novel map descriptors for characterization of toxic effects in proteomics maps. *J. Mol. Graph. Model*. **2003**, *22*, 1−9.

(8) Randic, M.; Balaban, A. T. On a four-dimensional representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 532−539.

(9) He, P.; Wang, J. Characteristic sequences for DNA primary sequence. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1080−1085.

(10) Yuan, C.; Liao, B.; Wang, T. New 3D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.* **2003**, *379*, 412−417.

(11) Li, C.; Wang, J. On a 3-D representation of DNA primary sequences. *Comb. Chem. High Throughput Screening* **2004**, *7*, 23−27.

(12) London, D. Inequalities in quadratic forms. *Duke Math. J*. **1966**, *33*, 511−522.

(13) Shrock, R.; Tsai, S.-H. Upper and lower bounds for ground-state entropy of antiferromagnetic Potts models. *Phys. Rev. E* **1997**, *55*, 6791−6794.

(14) Biggs, N. *Algebraic Graph Theory*, 1st ed.; Cambridge University Press: Cambridge, England, 1974.

(15) Liu, Y. The numerical characterization and similarity analysis of DNA primary sequences. *Internet Electron. J. Mol. Des.* **2002**, *1*, 675−684.

CI049874L