J Mol Model (2006) 12: 897–903
DOI 10.1007/s00894-005-0076-9

ORIGINAL PAPER

Na Liu · Tianming Wang

A weighted measure for the similarity analysis of DNA sequences

Received: 23 July 2005 / Accepted: 10 November 2005 / Published online: 28 March 2006

Abstract Here we propose a weighted measure for the similarity analysis of DNA sequences. It is based on LZ complexity and (0,1) characteristic sequences of DNA sequences. This weighted measure enables biologists to extract similarity information from biological sequences according to their requirements. For example, by this weighted measure, one can obtain either the full similarity information or a similarity analysis from a given biological aspect. Moreover, the length of DNA sequence is not problematic. The application of the weighted measure to the similarity analysis of β-globin genes from nine species shows its flexibility.

Keywords DNA sequence · (0,1) characteristic sequence · LZ complexity · Weighted measure

Introduction

With the completion of more genome projects, the list of genome sequences is expanding rapidly, which urges scientists to analyze them as far as possible so as to obtain more information about them. The most common method used to analyze DNA sequences is sequence alignment. This procedure generates a matrix of numbers that represents all possible alignments between the sequences. The highest set of sequential scores in the matrix defines an optimal alignment. For DNA sequences, a nucleotide-substitution matrix is used to score matches and mismatches. The advantage of using these matrices is that they are based on a defined evolutionary model and that the statistical significance of alignment scores obtained by local-alignment programs may be evaluated. However, their function is only realized with the aid of dynamic programming, which will be slow due to the large number of computational steps, and hence is not easy to use for long sequences [1]. Another method that has become popular recently relies on the quantitative characterization of DNA sequences, where many indices/invariants/descriptors have been calculated for the similarity analysis. This method is based on the fact that these indices/invariants/descriptors are a kind of characterization indices of DNA sequences. For instance, Raychaudhury and Nandy stated that graph radius g_R could effectively differentiate sequences of different size [2]. Randic et al. proposed that the leading eigenvalues of **M**/**M** and **L**/**L** matrices are invariants/descriptors that relate to DNA sequences, where the entries of the **M**/**M** matrix are given as a quotient of the Euclidean distance between two vertices of the zigzag curve of DNA sequences and the number of edges between two vertices. The entries of the **L**/**L** matrix are given as a quotient of the Euclidean distance between two vertices of the zigzag curve of DNA sequences and the sum of geometrical lengths of edges between the two vertices [3]. Work associated with invariants/descriptors can also be found in [4–9]. However, as pointed out in [4], this approach involves a series of as yet unresolved questions, which includes the questions as follows: (1) how to obtain and select suitable invariants/descriptors to characterize DNA sequences so as to compare sequences effectively; (2) the calculation of some effective invariants/descriptors become more and more difficult with the length of the sequences getting more large; (3) there is some loss of information.

In this paper, we propose a weighted measure for the similarity analysis of DNA sequences based on the (0,1) characteristic sequences and LZ complexity. The famous LZ complexity is related to the number of steps in a self-delimiting production process by which a given sequence is

N. Liu (✉)
Department of Applied Mathematics,
Dalian University of Technology,
Dalian, 116024, China
e-mail: liunasophia@163.com

N. Liu · T. Wang
College of Advanced Science and Technology,
Dalian University of Technology,
Dalian, 116024, China

T. Wang
Department of Mathematics, Hainan Normal University,
Haikou, 571158, China

presumed to be generated [10]. The virtue of this measure is its flexibility. Three weights make it convenient to emphasize a given aspect of DNA characteristics, which is useful to biologists to further analyze DNA sequences from different aspects. At the end of this paper, we illustrate the application of this weighted measure.

Special similarity measure

Let S, Q and R be sequences over a finite alphabet Δ, $L(S)$ be the length of S, $S(i)$ be the ith element of S and $S(i,j)$ be the subsequence of S that starts at position i and ends at position j. Note that $S(i, j)=\Phi$, for $i>j$. The concatenation of Q and R forms a new sequence $S=QR$, where Q is called a prefix of S, and S is called an extension of Q if there exists an integer i such that $Q=S(1,i)$.

An extension $S=QR$ of Q is reproducible from Q denoted by $Q\rightarrow S$, if there exists an integer $p\leq L(Q)$ such that $R(k) = S(p+k-1)$, for k=1, 2, …..,$L(R)$. For example: AAGA→AAGAAGA, with p=2.

A non-null sequence S is producible from its prefix $S(1,j)$, denoted by $S(1,j)\Rightarrow S$, if $S(1,j)\rightarrow S(1,L(S)-1)$. For example: AGGA⇒AGGAGGG with p=2.

Any non-null sequence S can be built from a production process by an iterative self-deleting-building process where at the ith step $S(1,h_{i-1})\Rightarrow S(1,h_i)$, $\Phi=S(1,0)\Rightarrow S(1,1)$. An m-step production process of S leads to a parsing of S into $H(S) = S(1,h_1)\ \bullet S(h_1+1,h_2)\bullet\ldots\ldots\ \bullet S(h_{m-1}+1,h_m)$, which is called the history of S, and $H_i(S) = S(h_{i-1}+1, h_i)$ is called the ith component of $H(S)$.

A component $H_i(S)$ is called exhaustive if $S(1,h_{i-1})\rightarrow S(1,h_i)$ is not true. A history is called exhaustive if each of its components (with a possible exception of the last one) is exhaustive. What is more important, the exhaustive history of any non-null sequence is unique. For example, for the sequence S=TGGGTTGGTTTG, its exhaustive history is EH(S)=T•G•GGT•TGGT•TTG.

Let $c(S)$ be the number of components in the exhaustive history of S, then it is an important complexity indicator because the production process here is an example of a class of parsing rules. According to [10], for any given sequences S and Q, the following property always remains valid: $c(QS)\leq c(Q)+c(S)$. Furthermore, Otu and Sayood [11] have proposed that the more similar S is to Q, the smaller $c(QS)-c(Q)$ is. That is $c(QS)-c(Q)$ depends on how much S is similar to Q.

For example, let S, Q, R represent three short DNA sequences defined over the set {A, C, G, T}.S=ATTCTGAGGTACGTAAAG,Q=GGTCTGATCTAGAACGTA, R=TAGCCACGATGCAGAC. By the above mentioned rule, the corresponding exhaustive histories of S, Q, R, SQ, SR, QR, QS, RS are: A•T•TC•TG•AG•GT•AC•GTAA•AG, G•GT•C•TG•A•TCTA•GAA•CG•TA, T•A•G•C•CA•CG•AT•GCA•GAC, A•T•TC•TG•AG•GT•AC•GTAA•AGGG•TCTGAT•CTA•GAA•CGTA, A•T•TC•TG•AG•GT•AC•GTAA•AGT•AGC•CA•CGA•TGC•AGA•C, G•GT•C•TG•A•TCTA•GAA•CG•TAT•AG•CC•ACGA•TGC•AGA•C, G•GT•C•TG•A•TCTA•GAA•CG•TAA•TT• CTGAG•GTAC•GTAAA•G, T•A•G•C•CA•CG•AT•GCA•GAC•ATT•CT•GAG•GT•ACGT•AA•AG.

So $C(S)$=9, $C(Q)$=9, $C(R)$=9, $C(SQ)$=13, $C(SR)$=15, $C(QS)$=14, $C(RS)$=16. We find that we need six steps to build R from S, six steps to build R from Q, four steps to build Q from S. Thus, we say that S and Q are relatively the most similar. The reason is that S and Q share the common patterns TCTGA and ACGTA.

Therefore, for any two sequences S and Q, we take

$$rd(S,Q) = (c(SQ)-c(S)) + (c(QS)-c(Q))/c(SQ) + c(QS)$$

as the similarity measure to reflect the similarity degree between S and Q. Then in the above example, $rd(S,Q)$=1/3 and $rd(S,R)$=13/31. Our formula also shows that S is more similar to Q than to R. Otu and Sayood [11] have proved that $d(S,Q)$= 2 $rd(S,Q)$ satisfies the following four conditions:

1. $d(S,Q)\geq 0$
2. $d(S,Q) = d(Q,S)$
3. $d(S,Q)\leq d(S,T)+d(T,Q)$
4. $D(Q,T)+d(S,R)\leq \max\{d(Q,S)+d(T,R), d(Q,R)+d(T,S)\}$

Without question, $rd(S,Q)$ also satisfies the four conditions. Since $d(S,Q)$ does not satisfy that $d(S,Q)$=0 iff Q=S, $rd(S,Q)$ does not satisfy this condition, i.e. $rd(S,Q)$ is not a standard distance metric. However $d(S,Q)$ may be regarded as a distance metric approximately. In other words, if $rd(S,S)\neq 0$, then it can be regarded as an error and ignored, as Otu and Sayood have done when they successfully constructed phylogenetic tree by using mtDNA sequences [11]. The rectified $rd(S,Q)$ can be used to construct hierarchical clustering or phylogenetic tree like other distance metric. We call $rd(S,Q)$ special similarity measure.

(0,1) characteristic sequences of DNA sequences

DNA sequences are composed of four nucleotides. These four nucleotides are arranged linearly on each chain. In research it is usually regarded as a sequence defined over the alphabet {A, C, G, T}, where the letters represent the four nucleotides bases: adenine, guanine, cytosine, and thymine. Biologists generally classify the four bases into groups according to their chemical structure or the strength of hydrogen bonds. They are as follows: 1. purine {A, G} and pyrimidine {C, T}; 2. amino group {A, C } and keto group {G, T}; 3. weak H-bond {A, T} and strong H-bond {C, G}.

Based on these classifications, any DNA sequence can be transformed into three other sequences. The transformation rules are represented by R, M, W, respectively,

Where

$$R(S(i)) = \begin{cases} 1 & S(i) = A, G \\ 0 & \text{else} \end{cases} \quad M(S(i)) = \begin{cases} 1 & S(i) = A, C \\ 0 & \text{else} \end{cases} \quad W(S(i)) = \begin{cases} 1 & S(i) = A, T \\ 0 & \text{else} \end{cases}$$

For example, for DNA sequence *S*=TTACTGAAGCTGAAAGGCGCTGTTCGATCA, it can be transformed into the following three sequences: *R*(*S*)=001001111001111110100100011001, *M*(*S*)= 001100110100111001010000101011, *W*(*S*)=1110101100101110000001011001101.

These sequences are called (0,1) characteristic sequences, which are named characteristic sequences of DNA primary sequences in [6]. The (0,1) characteristic sequences are a coarse-grained description of DNA sequences. They provide another chance for analyzing sequences from different aspects.

The weighted measure

Besides DNA primary sequences, (0,1) characteristic sequences may also be used to complement the existing methods for analyzing DNA sequences because they describe DNA sequences from the aspect of the chemistry of four bases. Motivated by this idea, we introduce a weighted similarity measure for the comparison of DNA sequences.

Given *n* DNA sequences S_1, S_2,….., S_n. We firstly transform them into the corresponding (0,1) characteristic sequences, denoted by R_1, R_2,….., R_n; M_1, M_2,…., M_n; W_1, W_2,……., W_n. Then according to the rule of production process of sequences and our special similarity measure, we can easily deduce three formulae:

$$rd\left(R_i, R_j\right) = \left(c\left(R_iR_j\right) - c(R_i) + c\left(R_jR_i\right) - c\left(R_j\right)\right) / \left(c\left(R_iR_j\right) + c\left(R_jR_i\right)\right)$$
$$rd\left(M_i, M_j\right) = \left(c\left(M_iM_j\right) - c(M_i) + c\left(M_jM_i\right) - c\left(M_j\right)\right) / \left(c\left(M_iM_j\right) + c\left(M_jM_i\right)\right)$$
$$rd\left(W_i, W_j\right) = \left(c\left(W_iW_j\right) - c(W_i) + c\left(W_jW_i\right) - c\left(W_j\right)\right) / \left(c\left(W_iW_j\right) + c\left(W_jW_i\right)\right)$$

Obviously, the above formulae reflect the similarity degree between S_i and S_j from three aspects: 1. from the purine–pyrimidine aspect; 2. from the amino–keto aspect; 3. from the weak–strong H-bond aspect. In other words, *rd* (R_i, R_j), *rd*(M_i, M_j), *rd*(W_i, W_j) reflect the similarity degree between S_i and S_j from the purine–pyrimidine aspect, the amino–keto aspect and weak–strong H-bond aspect, respectively.

Now, we may define the weighted measure as follows:

$$wrd\left(S_i, S_j\right) = \xi rd\left(R_i, R_j\right) + \eta rd\left(M_i, M_j\right) + \zeta rd\left(W_i, W_j\right), \quad i, j = 1,\ 2,\ \ldots\ldots ., n.$$

where ξ, η, ς are real numbers such that $\xi + \eta + \zeta = 1$.

The virtue of this weighted measure is that it is flexible to operate. The reason is that the three weights are variables and hence can be assigned values according to biologists' different requirements and aims if only they meet the restricted condition. For example, when $\xi = 1, \eta = \zeta = 0$, then *wrd*($S_i$, S_j)=*rd*(R_i, R_j) and it reflects the similarity degree between S_i and *Sj* from the purine–pyrimidine aspect; When $\eta = 1, \xi = \zeta = 0$, then *wrd*($S_i$,$S_j$)=*rd*($M_i$,$M_j$) and it reflects the similarity degree between S_i and S_j from the amino–keto aspect; When $\zeta = 1, \xi = \eta = 0$, then *wrd*($S_i$,$S_j$)=*rd*($W_i$,$W_j$) and it reflects the similarity degree between S_i and S_j from the weak–strong H-bond aspect; When $\xi = \eta, = \zeta = 1/3$, then $wrd\left(S_iS_j\right) = \left(rd\left(R_i, R_j\right) + rd\left(M_i, M_j\right) + rd\left(W_i, W_j\right)\right)/3$ and it reflects the full similarity degree between S_i and S_j. As for an optimal assignment of the three weights, it can be perfectly fulfilled by biologists who know well how much a certain base affects the character of DNA sequence.

In order to compare DNA sequences clearly and systematically, we adopt the form of a matrix to show their similarity relationship, in which the entry is *wrd*(S_i, S_j), denoted by *WRD*=(*wrd*(S_i, S_j)), *i*=1, 2, ……., *n* and *j*=1, 2, ……., *n*. This matrix is special for the fact that the entries on the main diagonal are not necessarily zeros, which can be deduced directly from the formula *wrd*(*S*,*Q*). However, they are valuable when we use them to scale the similarity degree between two different sequences as illustrated in the next section. So we won't rectify *WRD* unless we want to construct hierarchical clustering or phylogenetic tree, as described near the end of Section 2.

Results and discussion

In this section, we apply our weighted measure to analyze a set of β-globin genes. At first, we analyze their first exon sequences, whose similarity has been studied by many researchers [4, 7, 8, 12]. The validation of the results is tested by comparing with other results derived by a different method. Table 1 shows the coding sequences of the first exon of these species.

Tables 2, 3, 4 and 5 are obtained by assigning special values to ξ, η and ζ. They show the corresponding similarity results when 1. $\xi = 1,\ \eta = \zeta = 0$; 2. $\eta = 1,\ \xi = \zeta = 0$;

Table 1 The coding sequences of the first exon of β-globin from different species

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

3. $\zeta = 1,\ \xi = \eta = 0$; 4. $\xi = \eta = \zeta = 1/3$, respectively. On the whole, Tables 2, 3 and 4 show the similarity information of the nine sequences from three aspects on the basis of the three classifications of four bases, respectively, and Table 5 shows the full similarity information by synthesizing the information from the three aspects. From the first three Tables, we observe that: 1. from each of the three aspects, the first exon sequences of human and chimpanzee, opossum and human, goat and bovine are always the most similar; those of gallus and human, lemur and human are always the most dissimilar among these sequences; 2. the first exon sequences of rat and human are more dissimilar from the purine–pyrimidine and weak–strong H-bond aspects but are similar from the amino–keto aspect, relatively; 3. the first exon sequences of mouse and rat are more similar from the purine–pyrimidine aspect but more dissimilar from the amino–keto and weak–strong H-bond aspects. From Table 5, we obtain the full similarity information of these nine sequences derive from the average value of *rd*(*Ri*,*Rj*), *rd*(*Mi*,*Mj*) and *rd*(*Wi*,*Wj*). This result basically conforms to the result obtained by Randic by means of the Euclidean distance between the end points

Table 2 The WRD of the first exon coding sequences of β-globin genes from 9 species listed in Table 1 by our new method, when ξ=1, η=ζ=0

Species	Human	Goat	Gallus	Opossum	Lemur	Mouse	Rat	Bovine	Chimpanzee
Human	0.037	0.236	0.208	0.248	0.223	0.223	0.240	0.192	0.102
Goat		0.042	0.254	0.224	0.249	0.258	0.222	0.109	0.219
Opossum			0.037	0.271	0.264	0.256	0.264	0.242	0.229
Gallus				0.035	0.259	0.267	0.252	0.237	0.248
Lemur					0.035	0.236	0.252	0.219	0.207
Mouse						0.035	0.201	0.227	0.232
Rat							0.035	0.245	0.232
Bovine								0.039	0.189
Chimpanzee									0.034

Table 3 The WRD of the first exon coding sequences of β-globin genes from 9 species listed in Table 1 by our new method, when η=1, ξ=ζ=0

Species	Human	Goat	Gallus	Opossum	Lemur	Mouse	Rat	Bovine	Chimpanzee
Human	0.042	0.200	0.250	0.251	0.264	0.241	0.231	0.196	0.112
Goat		0.044	0.272	0.248	0.269	0.236	0.264	0.081	0.199
Opossum			0.039	0.272	0.268	0.254	0.278	0.267	0.258
Gallus				0.044	0.260	0.264	0.280	0.254	0.262
Lemur					0.042	0.283	0.283	0.264	0.263
Mouse						0.039	0.236	0.231	0.224
Rat							0.039	0.259	0.258
Bovine								0.042	0.194
Chimpanzee									0.037

Table 4 The WRD of the first exon coding sequences of β-globin genes from 9 species listed in Table 1 by our new method, when ζ=1, η=ξ=0

Species	Human	Goat	Gallus	Opossum	Lemur	Mouse	Rat	Bovine	Chimpanzee
Human	0.042	0.220	0.259	0.280	0.262	0.221	0.260	0.186	0.083
Goat		0.051	0.267	0.280	0.283	0.248	0.268	0.157	0.229
Opossum			0.039	0.233	0.265	0.254	0.256	0.256	0.258
Gallus				0.042	0.263	0.284	0.279	0.279	0.286
Lemur					0.037	0.265	0.250	0.268	0.263
Mouse						0.039	0.256	0.217	0.224
Rat							0.044	0.256	0.268
Bovine								0.044	0.205
Chimpanzee									0.037

Table 5 The WRD of the first exon coding sequences of β-globin genes from 9 species listed in Table 1 by our new method, when ξ=η=ζ=1/3

Species	Human	Goat	Gallus	Opossum	Lemur	Mouse	Rat	Bovine	Chimpanzee
Human	0.040	0.219	0.239	0.260	0.250	0.228	0.244	0.191	0.099
Goat		0.045	0.264	0.251	0.267	0.247	0.251	0.116	0.216
Opossum			0.039	0.259	0.266	0.255	0.266	0.255	0.249
Gallus				0.040	0.261	0.272	0.270	0.257	0.265
Lemur					0.038	0.261	0.261	0.250	0.245
Mouse						0.038	0.231	0.225	0.226
Rat							0.040	0.253	0.253
Bovine								0.042	0.196
Chimpanzee									0.036

of the 12-component vectors of the normalized leading eigenvalues of the **L**/**L** matrices, listed in Table 6 [4]. By further study of the values in these Tables, we can gain more information about their similarity.

From the four Tables, one may notice that the entries on the main diagonal are not zeros. This can be theoretically explained by the formula for *rd*(*S*,*S*), where $rd(S,S) = (c(SS) - c(S) + c(SS) - c(S))c(SS) + c\,(SS) = (c(SS) - c(S))/c(SS)$. According to the production rule introduced by Lempel and Ziv [10], *c*(*SS*)−*c*(*S*)=0 if and only if *S* is composed of a repeated substring. That is because, on this condition, its last component must be non-exhaustive and so *SS* will have the same number of components as *S* does. Otherwise, *c*(*SS*)−*c*(*S*)=1. This phenomenon does not affect the similarity analysis at all. On the contrary, the entries *wrd*(*S*,*S*) on the main diagonal represent the self-similarity and can be regarded as a reference value for the analysis. That is to say, if we know *wrd*(*S*,*S*), then the similarity degree between sequence *S* and any other sequence *Q* can be estimated well by comparing *wrd*(*S*,*Q*) with *wrd*(*S*, *S*). Take the first row in Table 5 for example. We observe that the value on the diagonal is 0.037, which represents the *wrd*(*S*,*S*) value, and among the other values in the same row, the value that is closer to 0.037 is 0.102, so we say chimpanzee is the most similar to human in terms of the coding sequences of the first exon of β-globin genes. In contrast, gallus, rat and lemur are the most remote from human.

WRD offers biologists a chance to take a close look at the similarity degree between any two sequences. Another usage of *WRD* is that it can be used to construct hierarchical clustering or phylogenetic trees. Here we construct four kinds of hierarchical clustering of these nine species based on their *WRD*s using their complete genes. The four Figures (Figs. 1, 2, 3 and 4) show basically consistent results, although they are not identical. There is little difference among these four Figures. Obviously, it is the classification of the four bases that results in the difference. This shows that different groups emphasize different aspects. Biologists may obtain further information by further study of the difference.

Table 6 The similarity/dissimilarity matrix for the coding sequences of Table 1 based on the Euclidean distances between the end points of the 12-component vectors of the normalized leading Eigenvalues of the L/L matrices

Species	Human	Goat	Gallus	Opossum	Lemur	Mouse	Rat	Bovine	Chimpanzee
Human	0	0.061	0.148	0.109	0.087	0.083	0.043	0.084	0.017
Goat		0	0.155	0.084	0.097	0.090	0.079	0.072	0.068
Opossum			0	0.129	0.093	0.130	0.143	0.207	0.152
Gallus				0	0.115	0.127	0.109	0.133	0.121
Lemur					0	0.050	0.078	0.155	0.089
Mouse						0	0.085	0.147	0.083
Rat							0	0.108	0.055
Bovine								0	0.088
Chimpanzee									0

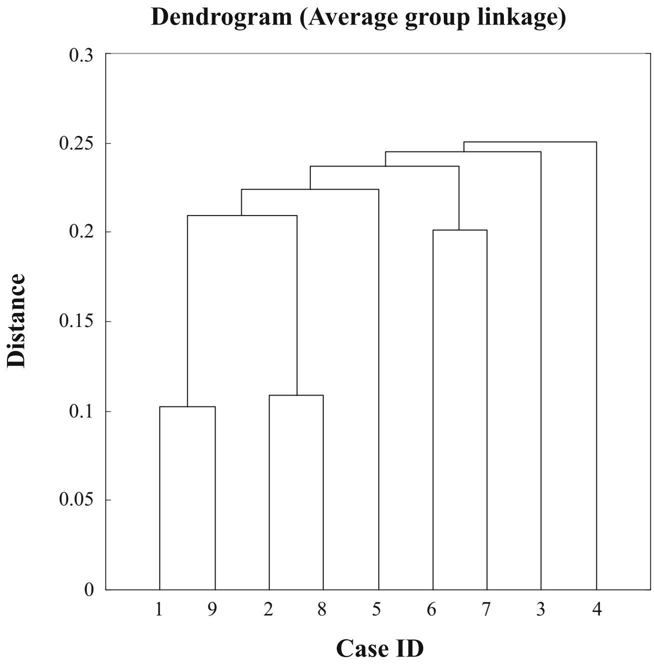

Fig. 1 The dendrogram of the hierarchical clustering of these nine species from the purine–pyrimidine aspect. 1-Human, 2-goat, 3-Opossum, 4-Gallus, 5-Lemur, 6-Mouse, 7-Rat, 8-Bovine, 9-Chimpanzee

Conclusion

The expanding list of genome sequences in various databases has provided a fertile ground for mathematics and computer science. It provides biologists with methods and algorithms to extract information from biological se-

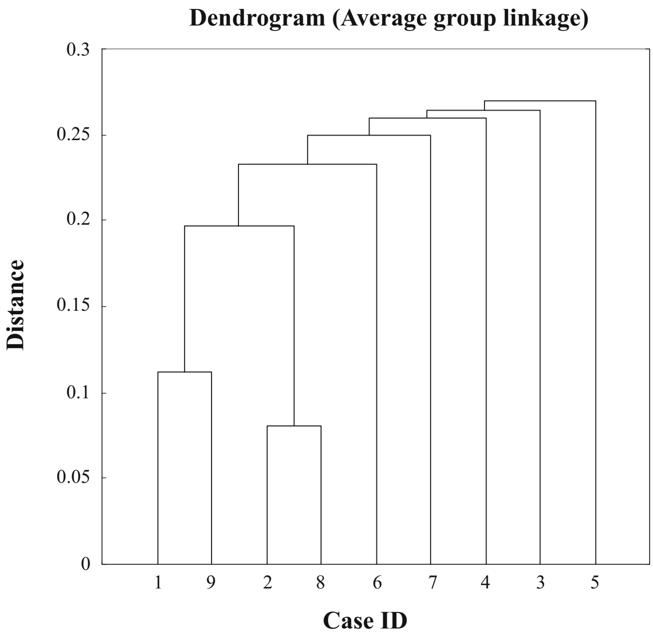

Fig. 2 The dendrogram of the hierarchical clustering of these nine species from the amino–keto aspect. 1-Human, 2-goat, 3-Opossum, 4-Gallus, 5-Lemur, 6-Mouse, 7-Rat, 8-Bovine, 9-Chimpanzee

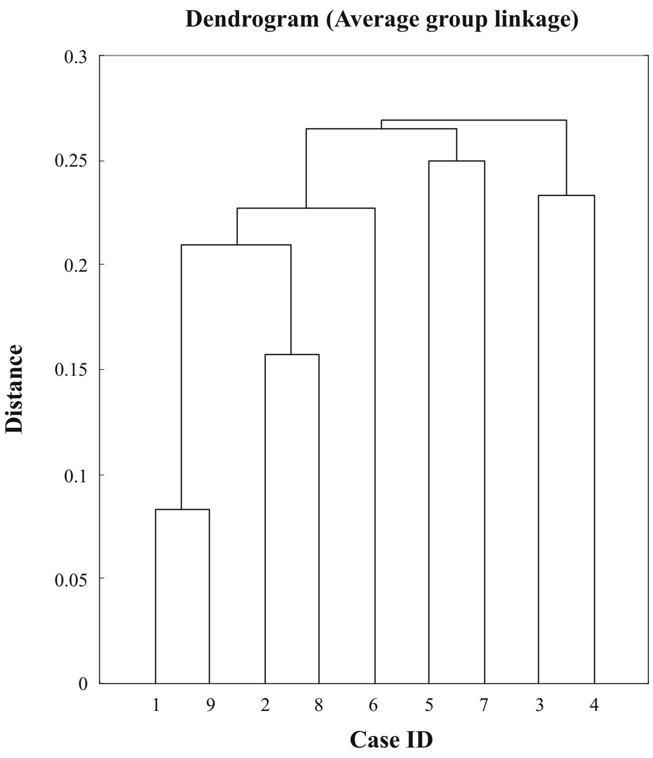

Fig. 3 The dendrogram of the hierarchical clustering of these nine species from the weak H-bond–strong H-bond aspect. 1-Human, 2-goat, 3-Opossum, 4-Gallus, 5-Lemur, 6-Mouse, 7-Rat, 8-Bovine, 9-Chimpanzee

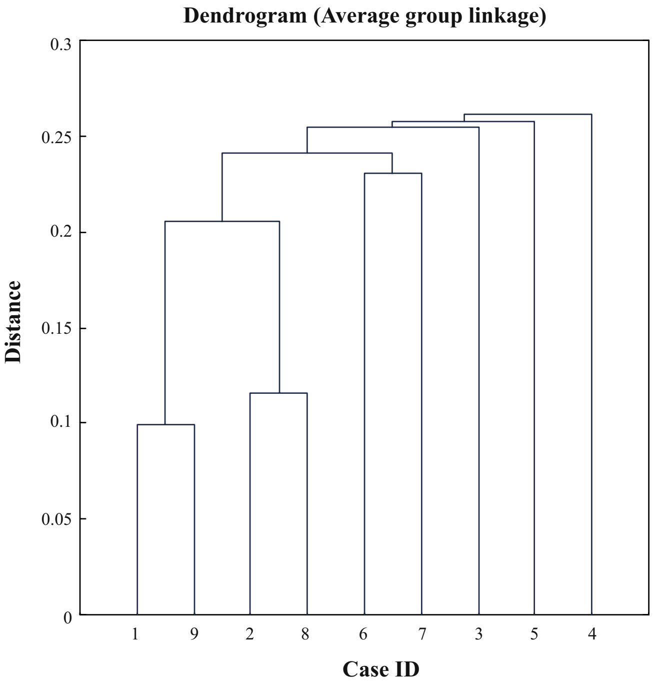

Fig. 4 The dendrogram of the hierarchical clustering of these nine species on the whole. 1-Human, 2-goat, 3-Opossum, 4-Gallus, 5-Lemur, 6-Mouse, 7-Rat, 8-Bovine, 9-Chimpanzee

quences, to analyze biological sequences and to evaluate the results that they obtain. They serve for biological research. In this paper, we present a weighted measure for biologists to analyze the similarity of DNA sequences. This weighted measure brings great flexibility to biologists' research by endowing these weights with variability. By this measure, biologists may assign optimal values to weights to obtain the corresponding optimal analysis result according to their own purpose. For instance, as shown at the end of this paper, this weighted similarity measure may be used to analyze the DNA sequences from three aspects based on different classifications of the four bases according to their chemical structure. It plays a complementary role in the analysis of DNA sequences.

Acknowledgements The author thanks all the anonymous referees for their valuable suggestions and support.

References

1. Mount DW (2002) Bioinformatics: sequence and genome analysis, 2nd edn. Cold Spring Harbor Laboratory, Island, pp 16–128
2. Raychaudhury C, Nandy A (1999) J Chem Inf Comput Sci 39:243–247
3. Randic M, Vracko M, Lers N, Plavsic D (2003) Chem Phys Lett 368:1–6
4. Randic M, Vracko M, Lers N, Plavsic D (2003) Chem Phys Lett 371:202–207
5. Nandy A (1996) Curr Sci 70:661–668
6. He PA, Wang J (2002) J Chem Inf Comput Sci 42:1080–1085
7. Randic M (2000) Chem Phys Lett 317:29–34
8. Randic M (2000) J Chem Inf Comput Sci 40:50–56
9. Randic M, Guo XF, Basak SC (2001) J Chem Inf Comput Sci 41:619–626
10. Lempel A, Ziv J (1976) IEEE T Inform Theory 22:75–81
11. Otu HH, Sayood K (2003) Bioinformatics 19:2122–2130
12. Liu N, Wang TM (2005) Chem Phys Lett 408:307–311