Relations between Shannon entropy and genome order index in segmenting DNA sequences

Yi Zhang*

Department of Mathematics, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, People's Republic of China

(Received 13 December 2008; revised manuscript received 14 March 2009; published 21 April 2009)

Shannon entropy H and genome order index S are used in segmenting DNA sequences. Zhang *et al.* [Phys. Rev. E **72**, 041917 (2005)] found that the two schemes are equivalent when a DNA sequence is converted to a binary sequence of S (strong H bond) and W (weak H bond). They left the mathematical proof to mathematicians who are interested in this issue. In this paper, a possible mathematical explanation is given. Moreover, we find that Chargaff parity rule 2 is the necessary condition of the equivalence, and the equivalence disappears when a DNA sequence is regarded as a four-symbol sequence. At last, we propose that $S-2^{-H}$ may be related to species evolution.

DOI: 10.1103/PhysRevE.79.041918 PACS number(s): 87.10.Vg

I. INTRODUCTION

The rapid growth of DNA sequences in the DNA databases has made analysis of DNA sequence very important in biology. Turning points (or segmentation points), through which the base composition undergoes sudden changes, usually have significant biological implications, including replication origins [1] and integration sites of horizontally transferred genes or genomic islands [2,3]. In bioinformatics, the Markov model [4,5], the recursive entropy [6–8], the cumulative GC profile [9], the wavelet multiple scale analysis [10], and the genome order index [11] are usual segmentation algorithms. The genome order index is especially noticeable for its simpleness and speediness. In 2004, Zhang and Zhang [12] proposed the genome order index $S=a^2+c^2+g^2+t^2$, where a,c,g,t are frequencies of bases A,C,G,T in the genome, respectively. Based on the values of S and H ($H=-\log_2(a^a c^c g^g t^t)$, the entropy of distributions of four bases in a genome) for 627 virus genomes, they found that $\frac{1}{4}\le S<\frac{1}{3}$ is valid for almost all genomes, meanwhile the correlation coefficient between S and H is calculated and found to be equal to -1. In 2005, Zhang *et al.* [11] found the equivalence of H and S in segmenting DNA sequences. But they left the mathematical proof to mathematicians who are interested in this issue. Recently, scientists gave some discussions about the two indices [13,14]. The relation between S and H, as well as those among the two indices and $G+C$ content still needs further study.

II. PROOF AND DISCUSSION

According to Zhang *et al.* [11], we describe the genome order index and entropy-based segmentation algorithms as follows. Consider a genome with N bases. Let n be an integer, $2\le n\le N-1$. For a given n, the genome sequence is partitioned into two sequences, one left and the other right. Let $P=(p_1,p_2,\ldots,p_k)$ and $Q=(q_1,q_2,\ldots,q_k)$ be two probability distributions, where $0\le p_i,q_i\le 1$, for $i=1,2,\ldots,k$, and $\Sigma_{i=1}^{k}p_i=1$, $\Sigma_{i=1}^{k}q_i=1$. Define

$$S(P)=\sum_{i=1}^{k}p_i^2,$$

which is the genome order index in the case of $k=4$. Let $P_l=(a_l,g_l,c_l,t_l)$ and $P_r=(a_r,g_r,c_r,t_r)$, where a_l,g_l,c_l,t_l and a_r,g_r,c_r,t_r are the occurrence frequencies of bases A, C, G, and T in the left and right subsequences, respectively. Define

$$\Delta S(P_l,P_r)=(n/N)S(P_l)+[(N-n)/N]S(P_r)-S\{(n/N)P_l+[(N-n)/N]P_r\},\quad n=2,\ldots,N-1. \tag{1}$$

Suppose that n^* is a position at which $\Delta S(P_l,P_r)$ reaches maximum; then, n^* is a compositional segmentation point of the genome first found. The algorithm is recursive. Similarly, the Jensen-Shannon divergence is defined by

$$D(n)=H-\left(\frac{n}{N}H_{\text{left}}+\frac{N-n}{N}H_{\text{right}}\right),\quad n=2,\ldots,N-1, \tag{2}$$

where H_{left} and H_{right} are the Shannon entropy for the left and right subsequences, respectively. Suppose that n^* is calculated by $D(n^*)=\max D(n)$, if $D(n^*)$ is above a given threshold, then n^* is deemed a segmentation point. The algorithm is also recursive.

Zhang *et al.* [11] found the coordinates of segmentation points derived from the genome order index and entropy-based segmentation algorithms are all identical when a DNA sequence is converted into a binary sequence. But they left the mathematical proof for such equivalence to mathematicians who are interested in this issue.

Given a DNA sequence, by $p_{\max}$ and $p_{\min}$ we denote the $\max\{a,c,g,t\}$ and the $\min\{a,c,g,t\}$, respectively, and write $p:=\frac{p_{\min}}{p_{\max}}$. By A_m we denote an arithmetic mean of $\frac{a}{p_{\max}}$, $\frac{g}{p_{\max}}$, $\frac{c}{p_{\max}}$, and $\frac{t}{p_{\max}}$ defined as follows:

$$A_m=\frac{a^2+g^2+c^2+t^2}{p_{\max}}=\frac{S}{p_{\max}}. \tag{3}$$

Similarly, a geometric mean (written as G_m) of them is defined to be

*zhaqi1972@163.com

 ©2009 The American Physical Society

$$\left(\frac{a}{p_{\max}}\right)^a\left(\frac{g}{p_{\max}}\right)^g\left(\frac{c}{p_{\max}}\right)^c\left(\frac{t}{p_{\max}}\right)^t,$$

which is calculated as

$$G_m = \frac{1}{2^H p_{\max}}, \tag{4}$$

where $H=-\log_2(a^a c^c g^g t^t)$, as we mentioned in Sec. III.

When the Chargaff parity rule 2 is considered, the following theorem gives the relation between A_m and G_m.

Theorem 1. Here,

$$A_m = G_m + o\left(\left|p_{\min} - \tfrac{1}{4}\right|\right). \tag{5}$$

Proof. By the Chargaff parity rule 2 [15] which says $a=t$ and $g=c$, there are only two cases: (i) $a=t=p_{\max}$, $g=c=p_{\min}$ or (ii)$a=t=p_{\min}$, $g=c=p_{\max}$. With these conditions we have

$$G_m = \left(\frac{p_{\min}}{p_{\max}}\right)^{2p_{\min}}\left(\frac{p_{\max}}{p_{\max}}\right)^{2p_{\max}} = p^{2p_{\min}}$$

and

$$A_m = 2p_{\max}\frac{p_{\max}}{p_{\max}} + 2p_{\min}\frac{p_{\min}}{p_{\max}} = 1 + 2p_{\min}(p-1).$$

Note that $2p_{\max}+2p_{\min}=a+g+c+t=1$. Setting $p_{\min}=x$, we have $p_{\max}=0.5-x$ and $p=\frac{p_{\min}}{p_{\max}}=\frac{x}{0.5-x}$. Correspondingly,

$$G_m = p^{2p_{\min}} = \left(\frac{x}{0.5-x}\right)^{2x}$$

and

$$A_m = 1 + 2p_{\min}(p-1) = \frac{8x^2-4x+1}{1-2x}.$$

Taylor's mean value theorem says that, when $x_0 \in (a,b)$ and $f(x)$ is $(n+1)$-order differentiable in (a,b), $\forall x \in (a,b)$, we have $f(x)=f(x_0)+\frac{f'(x_0)}{1!}(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+o[(x-x_0)^n]$. Naturally, $\forall x \in (0,1)$, by setting $x_0=\frac{1}{4}\in(0,1)$, one can get Taylor series for function $f(x)=A_m-G_m$ as follows:

$$f(x) = A_m - G_m = \frac{8x^2-4x+1}{1-2x} - \left(\frac{x}{0.5-x}\right)^{2x} = 8\left(x-\tfrac{1}{4}\right)^2 + 32\left(x-\tfrac{1}{4}\right)^3 + \frac{224}{3}\left(x-\tfrac{1}{4}\right)^4 + \frac{896}{3}\left(x-\tfrac{1}{4}\right)^5 + o\left[\left(x-\tfrac{1}{4}\right)^5\right] = o\left(\left|x-\tfrac{1}{4}\right|\right).$$

So,

$$A_m = G_m + o\left(\left|p_{\min} - \tfrac{1}{4}\right|\right).$$

■

Consequently, the negative correlation between the $S=a^2+c^2+g^2+t^2$ and $H=-\log_2(a^a c^c g^g t^t)$ always holds, as shown by Theorem 2.

Theorem 2. H is negatively correlated with S.

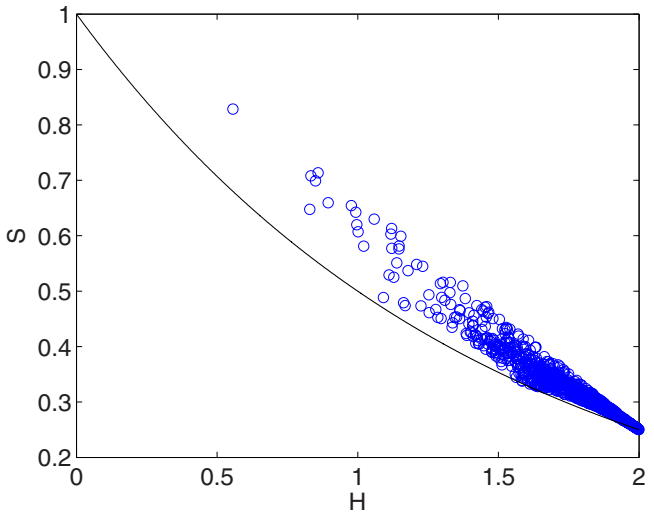

FIG. 1. (Color online) The circle plot demonstrates the relationship between S and H of 1000 groups of random values for a, c, g, and t. The curve is the plot of function $S=2^{-H}$.

Proof. Based on Eqs. (3) and (4) and Theorem 1 we come to the conclusion immediately:

$$S = 2^{-H} + p_{\max}(A_m - G_m) = 2^{-H} + 2\left(x-\tfrac{1}{4}\right)^2 - \frac{40}{3}\left(x-\tfrac{1}{4}\right)^4 + o\left[\left(x-\tfrac{1}{4}\right)^4\right] = 2^{-H} + o\left(\left|x-\tfrac{1}{4}\right|\right).$$

Obviously, H is negatively correlated with S. ■

Based on Theorem 2, it is necessary that correlation coefficient between S and H is -1, which was calculated based on 627 virus genomes in [12]. Moreover, in Fig. 1 of [11], Zhang *et al.* pointed out that H and S are linearly correlated. However, according to Theorem 2, $S \approx 2^{-H}$. When $H \in [1.86,2]$, the straight line (in Fig. 1) and curve $S=2^{-H}$ are approximately overlapped. In order to show the general rule independent of data, a general graph is plotted in Fig. 1. We pick 1000 groups of random values for a,c,g,t and plot S against H. The result is quite striking, an odd cusped bounded area near the curve of function $S=2^{-H}$. It may verify $S-2^{-H}=o(|x-\frac{1}{4}|)$. Moreover, the density is highly concentrated in the region [1.86, 2], where S seems to be linearly related to H.

In Fig. 3 of [12], Zhang and Zhang presented the correlation between "$G+C$ content" and Shannon Entropy H, as well as the correlation between $G+C$ content and genome order index S. In fact, based on the conclusion shown above, such correlations are also necessary.

Theorem 3. With Chargaff parity rule 2,

$$H = 1 - \log_2 x^x - \log_2 (1-x)^{(1-x)}, \tag{6}$$

where $x=g+c$.

Proof. By the Chargaff parity rule 2, $a=t$ and $g=c$, there are only two cases: (i) $a=t=p_{\max}$, $g=c=p_{\min}$ or (ii) $a=t=p_{\min}$, $g=c=p_{\max}$. With these conditions, using

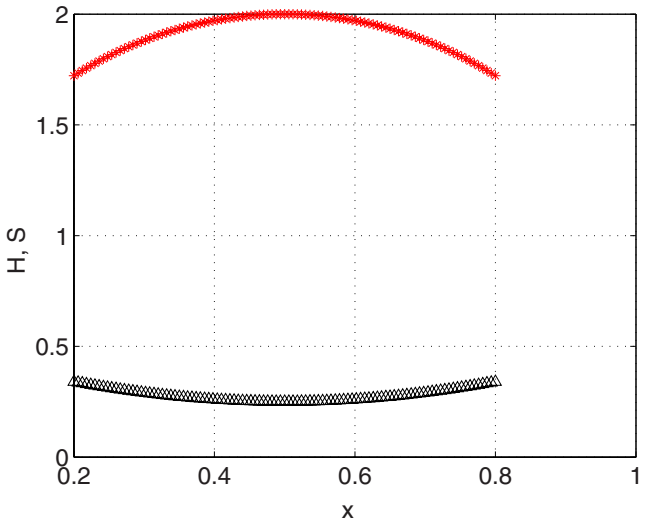

FIG. 2. (Color online) The upper is the plot of $H=1-\log_2 x^x-\log_2(1-x)^{(1-x)}$; the lower is the plot of $S=x^2-x+\frac{1}{2}$.

$$G_m = \frac{1}{2^H p_{\max}}$$

and

$$G_m = \left(\frac{p_{\min}}{p_{\max}}\right)^{2p_{\min}}\left(\frac{p_{\max}}{p_{\max}}\right)^{2p_{\max}} = p^{2p_{\min}},$$

we come to the conclusion immediately: $H=1-\log_2 x^x-\log_2(1-x)^{(1-x)}$, where $x=g+c$. ■

Theorem 4. With Chargaff parity rule 2,

$$S = x^2 - x + \tfrac{1}{2}, \tag{7}$$

where $x=g+c$.

Proof. By the Chargaff parity rule 2, $a=t$ and $g=c$, there are only two cases: (i) $a=t=p_{\max}$, $g=c=p_{\min}$, or (ii)$a=t=p_{\min}$, $g=c=p_{\max}$. With these conditions, using

$$A_m = \frac{a^2+g^2+c^2+t^2}{p_{\max}} = \frac{S}{p_{\max}}$$

and

$$A_m = 2p_{\max}\frac{p_{\max}}{p_{\max}} + 2p_{\min}\frac{p_{\min}}{p_{\max}} = 1 + 2p_{\min}(p-1),$$

we come to the conclusion immediately: $S=x^2-x+\frac{1}{2}$, where $x=g+c$. ■

Then, we plot $H=1-\log_2 x^x-\log_2(1-x)^{(1-x)}$ and $S=x^2-x+\frac{1}{2}$ for $x\in(0.2,0.8)$ in Fig. 2, respectively. The plots are consistent with those shown in Fig. 3 of [12]. It means the correlations among S,H and $G+C$ content appeared in Fig. 3 of [12] are necessary. On the other hand, the consistency also verifies the correctness of above theorems.

From a physics point of view, the entropy generally is maximized. Based on the negative correlation of S and H, most genomes are likely to maximize their entropy and minimize their S, which makes $S<\frac{1}{3}$ valid to most genomes except those appeared above the line $S=\frac{1}{3}$ in Fig. 3 of [12].

TABLE I. Absolute and relative difference between S and 2^{-H} of some DNA segments of the MHC, where the start and end points represent the base positions of the MHC.

Start point	End point	$S-2^{-H}$	$\frac{S-2^{-H}}{S+2^{-H}}$
1	2483966	0.0006888	0.0006874
2483966	3673777	0.0043608	0.0043039
2483966	3384906	0.0079023	0.0077161
3384906	3673777	0.0000377	0.0000377
3384906	3444779	0.0033878	0.0033535
3444779	3673777	0.0000564	0.0000564
2483966	3054365	0.0094747	0.0092072
3054365	3384906	0.0055121	0.0054214
1	1841871	0.0017523	0.0017431
1841871	2483966	0.0003481	0.0003478
1	833239	0.0032821	0.0032498
833239	1841871	0.0008454	0.0008432
1841871	2276710	0.0001611	0.0001610
2276710	2483966	0.0009711	0.0009683

Elhaik *et al.* [13] claimed that "$S<\frac{1}{3}$ is a mathematical property of S that should be always valid regardless of specific data." This claim is wrong in the case where the frequencies $a=0.6$, $c=0$, $g=0.4$, $t=0$, so $S=0.52>\frac{1}{3}$. Note that in [14] Zhang gave a similar example.

When a DNA sequence is converted into a binary sequence of S (strong H bond) and W (weak H bond), set $x=\min\{a+t,g+c\}$, we have

$$S - 2^{-H} = [x^2+(1-x)^2] - [(x^x)(1-x)^{(1-x)}]. \tag{8}$$

Setting $x_0=\frac{1}{2}$, we can calculate the Taylor series of $(S-2^{-H})$ as follows. $S-2^{-H}=[x^2+(1-x)^2]-[(x^x)(1-x)^{(1-x)}]=[\frac{1}{2}+\log 2-\frac{1}{2}\times(\log 2)^2-\frac{1}{2}\times(1-\log 2)]\times(-1+\log 2)\times(x-\frac{1}{2})^2+o(x-\frac{1}{2})^2=o(|x-\frac{1}{2}|)$. Obviously, when $o(|x-\frac{1}{2}|)$ is small enough, S and 2^{-H} can be regarded as equal.

Generally speaking, for a DNA sequence, the difference between S and 2^{-H} is very small. Following Zhang *et al.* [11], we take the complete sequence of human major histocompatibility complex (MHC) as an example. The MHC is 3673 777 bp long and can be seen from [16]. Recently, due to the extensive study of its isochore structure, the MHC sequence becomes a touchstone for testing any segmentation algorithm. We list 14 pairs of $S-2^{-H}$ and $\frac{S-2^{-H}}{S+2^{-H}}$ in Table I for 14 segments of the MHC, respectively. Obviously, $S-2^{-H}$ is very small. It verifies the equivalence of S and 2^{-H} for DNA segmentation. Additionally, when 2^{-H} increases, H will decrease. Consequently, when H reaches the maximum, S will reach the minimum. In this sense, the closeness of S and 2^{-H} may underlie the equivalence of S and H in segmenting DNA. That is to say, the equivalence of S and 2^{-H} may make $\Delta S(P_l,P_r)$ [in Eq. (1)] and $D(n)$ [in Eq. (2)] reach their maximum at the same point.

Note that article [11] only pointed out that the equivalence exists in the MHC and other human chromosomes; according to the mathematical analysis above, such equivalence should

TABLE II. Absolute and relative difference between S and 2^{-H} of some DNA segments of the *E. coli*, where the start and end points represent the base positions of the *E. coli*.

Start point	End point	$S-2^{-H}$	$\frac{S-2^{-H}}{S+2^{-H}}$
1	526200	0.0003766925089	0.0003762669374
526200	4639675	0.0000412604561	0.0000425526665
526200	2113505	0.0000004425977	0.0000004425941
2113505	4639675	0.0001183103201	0.0001182683322
526200	2100883	0.0000000355739	0.0000000355735
2100883	2113505	0.0112656857554	0.0108079149083
2113505	2126034	0.0024776783140	0.0024592963188
2126034	4639675	0.0001141155947	0.0001140765310

exist in all genomes. This conclusion should be regarded as an extension of the result in [11]. In order to verify the extension, we take the *E. coli* K12 (NC_000913, 4639 675 bp) as an example. As expected, S and H present complete equivalence in segmenting the *E. coli* genome (data not shown). Merely, in Table II, we present the differences of S and 2^{-H} for some *E. coli* DNA segments to show that, on the segment points, $S-2^{-H}$ values are always small enough to make S and H equivalent.

Analysis shown above indicates that the equivalence of H and S in segmenting DNA sequences should derive from the equivalence of S and 2^{-H}, and it also shows that the equivalence of H and S is mathematical, neither biological nor chemical.

From the proof of Theorem 1, one can see that a necessary condition of the equivalence of H and S is Chargaff parity rule 2. That is to say when Chargaff parity rule 2 is not obeyed, the equivalence should be broken. For example, when a DNA sequence is regarded as a four-symbol sequence, i.e., when deviations from Chargaff parity rule 2 (saying $|g-c|$ and $|a-t|$) is considered, the equivalence will not be perfect anymore. According to the result of Zhang *et al.* [11], in the segment between 3444 780 and 3673 777 of the MHC, for S and H schemes, there are three common segmentation points: 3491 519, 3552 176, and 3638 110 when two-symbol (S and W) sequence is used. But for four-symbol (A, G, C, and T) sequences, S and H result in different segmentation points, which are 3637 779 and 3632 040, respectively, i.e., the equivalence is broken. In fact, only 1396 980, 1490 846, 1742 437, 1841 871, 2483 966 and 3444 780 are common segmentation points for two-symbol and four-symbol sequences of the MHC. This example shows that Chargaff parity rule 2 is really the necessary condition of the equivalence of H and S in segmenting DNA sequences. Additionally, in article [13], Elhaik *et al.* claimed that "S and H functions are strictly equivalent to and derivable from each other;" from our discussion, their conclusion is not correct.

III. APPLICATION OF $S-2^{-H}$ IN ANALYZING SEQUENCE SIMILARITY

Through studying the $S-2^{-H}$ values of DNA sequences from different species, we found that the invariant is species specific, and it is related to species evolution to a certain extent. Historically, many researchers have studied the evolutionary similarity between different organisms by examining the similarity of biological sequences, which has been reviewed by Nandy *et al.* in [17]. In this paper, we use the DNA sequences presented in Table III to study the similarity

TABLE III. The coding sequences of the exon 1 of beta-globin gene of ten different species.

Species	Coding sequence	Length
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAA-GTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG	86
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG-GTGAACGTGGATGAAGTTGGTGGTGAG-GCCCTGGGCAGGTTGGTATCAAGG	105
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAA-GTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG	86
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG-GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG	93
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG-GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG	92
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAG-GTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG	92
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAA-GGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG	94
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAG-GTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG	92
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAG-GTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC	90
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAG-GTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG	92

TABLE IV. The $S-2^{-H}$ of the ten DNA sequences presented in Table III.

Species	$S-2^{-H}$
Bovine	0.0177
Chimpanzee	0.0145
Goat	0.0176
Gorilla	0.0158
Human	0.0145
Lemur	0.0135
Mouse	0.0096
Opossum	0.0031
Rabbit	0.0190
Rat	0.0086

between ten species.

First, the $S-2^{-H}$ values of the ten DNA sequences are calculated and listed in Table IV. Apparently, opossum has the least $S-2^{-H}$ value for it is the only pouched animal among the ten species. Four primates have similar $S-2^{-H}$ values, showing their close evolution relations. Especially, the values from chimpanzee and human are identical.

Naturally, the similarity between the ten species can be measured by the "differences" between their $S-2^{-H}$ values, as listed in Table V.

From Table V, we see that the three kinds of primate (human, chimpanzee, and gorilla) DNA sequences are strongly similar to each other for the entries associated with them are all very small. Opossum shows great dissimilarity with others for it is the only pouched animal listed here. This is coincident with the results reported in [18–24]. Table V also shows that the small entry is associated with the pair bovine goat. So, one may expect $S-2^{-H}$ is related to species evolution.

To further show the $S-2^{-H}$ being related to species evolution, we apply it to the other data set, which includes 19 mitochondrial genomes (shown in Table VI).

By our method, a relationship tree for the 19 DNA sequences is obtained, as shown in Fig. 3. For concision, corresponding $S-2^{-H}$ values and the similarity matrix are not shown.

TABLE VI. The accession number and length of 19 mitochondrial genomes.

Species	Accession No.	Length (bp)
Blue whale	NC001601	16402
Fin whale	NC001321	16398
Opossum	NC003039	17191
Baboon	NC001992	16521
Bornean orangutan	NC001646	16389
Chimpanzee	NC001643	16554
Common gibbon	NC002082	16472
Cow	NC006853	16338
Donkey	NC001788	16670
Gray seal	NC001602	16797
Harbor seal	NC001325	16826
Hippopotamus	NC000889	16407
Horse	NC001640	16660
Human	NC001807	16571
White rhinoceros	NC001808	16832
Indian rhinoceros	NC001779	16829
Platypus	NC000891	17091
Sheep	NC001941	16616
Sumatran orangutan	NC002083	16499

In Fig. 3, the relationship tree of 19 species is reasonable. Two pouched out groups, i.e., platypus and opossum, are grouped together and located far away from others. All primates, including human, chimpanzee, baboon, common gibbon, Sumatran orangutan, and Bornean orangutan are close in the relationship tree. Additionally, the close relationships between other species, including gray-seal and harbor-seal, fin whale and blue whale, Indian rhinoceros and white rhinoceros, horse and donkey, and cow and sheep are also reasonable. These relationships are consistent with those shown in [25]. Based on the discussion above, one may come to a

TABLE V. The similarity/dissimilarity matrix for the ten coding sequences of Table III based on the differences between $S-2^{-H}$ values shown in Table IV.

	Bovine	Chimpanzee	Goat	Gorilla	Human	Lemur	Mouse	Opossum	Rabbit	Rat
Bovine	0	0.0033	0.0001	0.0019	0.0033	0.0042	0.0081	0.0147	0.0012	0.0091
Chimpanzee		0	0.0032	0.0014	0	0.001	0.0048	0.0114	0.0045	0.0059
Goat			0	0.0018	0.0032	0.0042	0.0008	0.0146	0.0013	0.0091
Gorilla				0	0.0014	0.0023	0.0062	0.0127	0.0031	0.0072
Human					0	0.001	0.0048	0.0114	0.0045	0.0059
Lemur						0	0.0038	0.0104	0.0055	0.0049
Mouse							0	0.0066	0.0093	0.0011
Opossum								0	0.0159	0.0055
Rabbit									0	0.0104
Rat										0

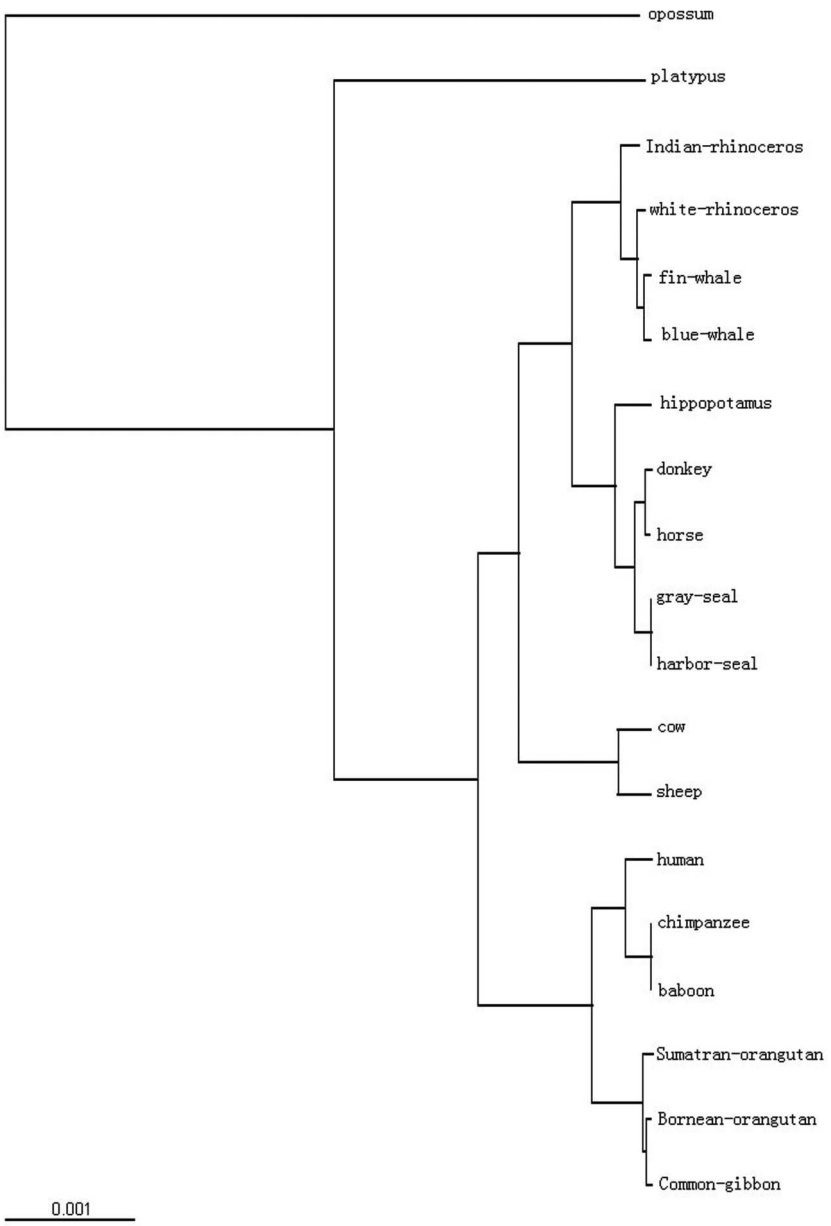

FIG. 3. The relationship tree for the 19 mitochondrial genomes by KITSCH (or UPGMA) method based on our scheme.

conclusion that $S-2^{-H}$ is related to species evolution. To give reasons for it, one may expect the $|x-\frac{1}{4}|$ (appeared in the proof of Theorem 2, $p_{\min}=x$) might be related to species evolution, so, based on $S-2^{-H}=o(|x-\frac{1}{4}|)$, $S-2^{-H}$ should also be related to species evolution. Perhaps, being modified, $S-2^{-H}$ can deal with similar questions appeared in other fields, such as analyzing similarity for amino acid or codon sequences.

ACKNOWLEDGMENTS

The author thanks Professor Jun Wang (Shanghai Normal University) for his helpful discussion. This work was partially supported by Hebei Province Education Foundation under Grant No. Z2008111 and Dalian University of Technology Foundation under Grant No. 2MXDUT073002.

[1] J. R. Lobry, Biochimie **78**, 323 (1996).
[2] R. Zhang and C. T. Zhang, Archaea **1**, 335 (2004).
[3] R. Zhang and C. T. Zhang, Bioinformatics **20**, 612 (2004).
[4] J. W. Fickett, D. C. Torney, and D. R. Wolf, Genomics **13**, 1056 (1992).
[5] G. A. Churchill, Comput. Chem. (Oxford) **16**, 107 (1992).
[6] J. L. Oliver, P. Bernaola-Galvan, P. Carpena, and R. Roman-Roldan, Genetics **276**, 47 (2001).
[7] W. Li, P. Bernaola-Galvan, F. Haghighi, and I. Grosse, Comput. Chem. (Oxford) **26**, 491 (2002).

[8] W. Li, Genetics **276**, 57 (2001).
[9] C. T. Zhang and R. Zhang, Genomics **83**, 384 (2004).
[10] S. Y. Wen and C. T. Zhang, Biochem. Biophys. Res. Commun. **311**, 215 (2003).
[11] C. T. Zhang, F. Gao, and R. Zhang, Phys. Rev. E **72**, 041917 (2005).
[12] C. T. Zhang and R. Zhang, Comput. Biol. Chem. **28**, 149 (2004).
[13] E. Elhaik, D. Graur, and K. Josic, Comput. Biol. Chem. **32**, 147 (2008).
[14] R. Zhang, Comput. Biol. Chem. (to be published).
[15] D. R. Forsdyke and J. R. Mortimer, Genetics **261**, 127 (2000).
[16] http://www.sanger.ac.uk/HGP/Chr6/
[17] A. Nandy, M. Harle, and S. C. Basak, Arkivoc **9**, 211 (2006).
[18] M. Randic, X. F. Guo, and S. C. Basak, J. Chem. Inf. Comput. Sci. **41**, 619 (2001).
[19] P. He and J. Wang, Internet Electron. J. Mol. Des. **1**, 668 (2002).
[20] P. He and J. Wang, J. Chem. Inf. Comput. Sci. **42**, 1080 (2002).
[21] Y. Liu, Internet Electron. J. Mol. Des. **1**, 675 (2002).
[22] J. Wang and Y. Zhang, Chem. Phys. Lett. **423**, 50 (2006).
[23] J. Wang and Y. Zhang, Chem. Phys. Lett. **425**, 324 (2006).
[24] Y. Zhang and J. Wang, J. Math. Chem. **43**, 864 (2008).
[25] J. Barral, P. L. Cantinib, A. Hasmy, J. Jimenezc, and A. Marcano, J. Theor. Biol. **236**, 422 (2005).