LZ Complexity Distance of DNA Sequences and Its Application in Phylogenetic Tree Reconstruction

Bin Li[1]*, Yi-Bing Li[2], and Hong-Bo He[2]

[1] School of Information Science and Engineering, and [2] School of Physics, Central South University, Changsha 410083, China.

DNA sequences can be treated as finite-length symbol strings over a four-letter alphabet (A, C, T, G). As a universal and computable complexity measure, LZ complexity is valid to describe the complexity of DNA sequences. In this study, a concept of conditional LZ complexity between two sequences is proposed according to the principle of LZ complexity measure. An LZ complexity distance metric between two nonnull sequences is defined by utilizing conditional LZ complexity. Based on LZ complexity distance, a phylogenetic tree of 26 species of placental mammals (Eutheria) with three outgroup species was reconstructed from their complete mitochondrial genomes. On the debate that which two of the three main groups of placental mammals, namely Primates, Ferungulates, and Rodents, are more closely related, the phylogenetic tree reconstructed based on LZ complexity distance supports the suggestion that Primates and Ferungulates are more closely related.

Key words: bioinformatics, sequence complexity, conditional LZ complexity, LZ complexity distance, phylogenetic tree reconstruction

Introduction

Approaches of phylogenetic tree reconstruction using biological molecular data, such as DNA, RNA, and protein sequences, can be divided into two groups (*1*). The first group reconstructs phylogenetic trees by evaluating the trees' topology based on certain optimal criteria, among which the two most available ones are maximum parsimony and maximum likelihood. The second group utilizes various distance measures, in which a phylogenetic tree is reconstructed from a distance matrix that is obtained by calculating distances between every two sequences. Traditional sequence distance matrices include p-distance, Jukes-Cantor distance, Kimura distance, Gamma distance, and so on (*2*), all of which require sequence alignment that is strict with the sequence data to be aligned. Generally, before sequence alignment, it is necessary to perform some pretreatments such as extracting related structure or function segments from primary sequences and performing gene prediction (*3*). Furthermore, it is much empirical to select or create a sequence alignment score matrix (*4*), the difference of which may affect alignment results tremendously. To overcome these problems, more and more researchers begin to try alignment-free methods for DNA sequence comparison and analysis (*5*).

Complexity is one of the most basic properties of a symbolic sequence. In respect that DNA sequences can be treated as finite-length symbol strings over a four-letter alphabet (A, C, T, G), DNA sequence complexity is much attractive to many researchers (*5*). Kolmogorov complexity, the first formal theoretical description of sequence complexity, was proposed by Kolmogorov from the view of algorithm information theory (*5*). Li *et al* (*6*) first introduced Kolmogorov complexity to DNA sequence analysis and proposed a DNA sequence distance matrix based on it. Because Kolmogorov complexity is not computable, Chen *et al* (*7*) made use of data compression gain to approximate Kolmogorov complexity. However, the generalization of the approximate method is greatly limited because the data compression gain varies evidently with the object to be compressed and the algorithm that a certain compressor uses (*8*). In contrary, LZ complexity, another significant complexity measure proposed by Lempel and Ziv (*9*), is easily computable and is also a universal depiction of sequence complexity.

Based on the computational principle of LZ com-

*** Corresponding author.**
E-mail: li_bin@126.com

plexity, we propose a concept of conditional LZ complexity between two sequences. An LZ complexity distance metric is defined according to conditional LZ complexity. The LZ complexity distance has been applied to the reconstruction of a phylogenetic tree of 26 species of placental mammals (Eutheria) with three outgroup species.

Model

Sequence LZ complexity and conditional LZ complexity

Preliminaries

Given a symbolic sequence $S = s_1 s_2 \ldots s_n$, the function $l(S) = n$ denotes the length of S. $S(i, j)$ denotes the subsequence $s_i s_{i+1} \ldots s_j$ of S that starts at position i and ends at position j, where if $i > j$ or $j < 1$, then $S(i, j)$ is a null sequence (denoted by φ). The *vocabulary* of S, denoted by $v(S)$, is defined as the set formed by all the subsequences (*words*). The concatenation of S and another sequence Q forms a new sequence $R = SQ$, where S is called a *prefix* of R and R is called an *extension* of S. If there exists an integer i, then $S = R(1, i)$. When the length of S is not specified explicitly, it is convenient to identify the prefix of S by means of a special operator π where $S_{\pi^i} = S\big(1, l(S) - i\big)$, $i = 0, 1, \ldots$ In particular, $S_{\pi^0} = S$, and $S_{\pi^i} = \varphi$ for $i \geq l(S)$.

An extension $R = SQ$ is said to be *reproducible* from S, denoted by $S \rightarrow R$, if $Q \in v(R_\pi)$. In sequence reproduction process, since $Q \in v(R_\pi)$ implies the existence of a positive integer $p \leq l(S)$ such that $q_i = r_{p+i-1}$, $i = 1, 2, \ldots, l(Q)$, R can be generated from S by first copying the known symbol $s_p = r_p$ of S to obtain $q_1 = r_{1+l(S)}$; then $q_2 = r_{2+l(S)}$ can be obtained by copying r_{p+1} (which may still be a symbol of S or, if $p = l(S)$, the first and already known symbol of Q), and so on, until the last symbol of Q.

A nonnull sequence S is said to be *producible* from its prefix $S(1, j)$, denoted by $S(1, j) \Rightarrow S$, if $S(1, j) \rightarrow S_\pi$ and $j < l(S)$. The distinction between the production process $S(1, j) \Rightarrow S$ and the reproduction process $S(1, j) \rightarrow S$ lies in the recursive copying process that characterizes the latter. It is required that the extended subsequence $S\big(j+1, l(S)\big)$ belongs to the vocabulary of S_π, namely $S\big(j+1, l(S)\big) \in v(S_\pi)$, in the reproduction process. While in the production process, it is required that the subsequence $S\big(j+1, l(S)-1\big)$ belongs to $v(S_\pi)$. The production process allows for a single-symbol innovation at the end of the copying process.

Sequence LZ complexity

Any nonnull sequence S can be built from a null sequence φ using an m-step production process:

$\varphi \Rightarrow S(1, h_1) \Rightarrow S(1, h_2) \Rightarrow \cdots \Rightarrow S(1, h_i) \Rightarrow \cdots \Rightarrow S(1, h_m)$

Note that $1 \leq m \leq l(S)$ and $h_m = l(S)$. Let $h_0 = 1$, the above m-step production process of sequence S can result in a parsing of S as follows:

$$H(S) = S(h_0, h_1)\, S(h_1 + 1, h_2) \ldots S(h_{i-1} + 1, h_i) \ldots S(h_{m-1} + 1, h_m)$$

where $H(S)$ is called a *production parsing* of sequence S and $H_i(S) = S(h_{i-1} + 1, h_i)$ is called the i^{th} *production component* of $H(S)$. The number of production components in a production parsing is denoted by $c_{H(S)}$.

A production component $H_i(S)$ and the corresponding production step $S(1, h_{i-1}) \Rightarrow S(1, h_i)$ are said to be *maximum* if $S(1, h_{i-1}) \nrightarrow S(1, h_i)$, where $\nrightarrow$ denotes the negation of $\rightarrow$. A production parsing $H(S)$ is said to be *minimum* if each of its production components, with a possible exception of the last one, is maximum. Using $E(S)$ to denote the minimum production parsing, the number of production components in $E(S)$ can be denoted as $c_{E(S)}$. It has been proved by Lempel and Ziv that the minimum production parsing of a given sequence is unique (*9*).

Lempel and Ziv (*9*) defined the complexity of a sequence as the number of production components in the minimum production parsing of this sequence, which is called sequence LZ complexity. Using $c(S)$ to denote the LZ complexity of sequence S, we have $c(S) = c_{E(S)}$. According to the definition of sequence LZ complexity, the minimum production parsing of a certain sequence can be built and then the LZ complexity of this sequence can be easily obtained. Kaspar and Schuster (*10*) presented a detailed algorithm and a flow chart to compute sequence LZ complexity. The following three inequalities have also been proved in previous studies (*9*, *10*):

$$c(S) < c(SQ) \tag{1}$$

$$c(S) < c(QS) \tag{2}$$

$$c(SQ) \leq c(S) + c(Q) \tag{3}$$

For a detailed analysis of many other properties of sequence LZ complexity, see previous studies (*9*, *10*).

Sequence conditional LZ complexity

Sequence LZ complexity can significantly describe the complexity of a single sequence. To depict the complexity relationship between two sequences, we propose a notion of conditional LZ complexity according to the principle of sequence LZ complexity.

Given a sequence T, an extension $R = SQ$ of sequence S is said to be *conditional reproducible* from S, denoted by $[T]S \rightarrow R$, if $Q \in v(TR_\pi)$. To extend S into R, the reproduction process only uses the vocabulary of sequence R_π, namely $v(R_\pi)$; whereas by concatenating T before S, the conditional reproduction process also uses the information offered by T, namely the vocabulary $v(TR_\pi)$, where $v(R_\pi) \in v(TR_\pi)$. This is the main difference between the reproduction process and the conditional reproduction process.

Given a sequence T, a nonnull sequence S is said to be *conditional producible* from its prefix $S(1,j)$, denoted by $[T]S(1,j) \Rightarrow S$, if $[T]S(1,j) \rightarrow S_\pi$ and $j < l(S)$. Similar to the production parsing of S, given a conditional sequence T, the *conditional production parsing* of S using an m-step conditional production process can be built as:

$$H(S|T) = S(h_0,h_1)\,S(h_1+1,h_2)\ldots S(h_{i-1}+1,h_i)\ldots S(h_{m-1}+1,h_m)$$

where $H_i(S|T) = S(h_{i-1}+1,h_i)$ is called the i^{th} *conditional production component* of $H(S|T)$. The number of conditional production components in a conditional production parsing is denoted by $c_{H(S|T)}$. A conditional production component $H_i(S|T)$ and the corresponding conditional production step $[T]S(1,h_{i-1}) \Rightarrow S(1,h_i)$ are said to be *maximum* if $[T]S(1,h_{i-1}) \nRightarrow S(1,h_i)$. A conditional production parsing $H(S|T)$ is said to be *minimum* if each of its conditional production components, with a possible exception of the last one, is maximum. Similar to the minimum production parsing, the minimum conditional production parsing is also unique. In respect that, relative to the minimum production parsing, any conditional production component $H_i(S|T)$ is obtained from a larger vocabulary $v(TR_\pi) \supseteq v(R_\pi)$, so the length of each maximum production component will not be longer than that of the corresponding maximum conditional production component. Using $E(S|T)$ to denote the minimum conditional production parsing, the number of conditional production components in $E(S|T)$ can be denoted as $c_{E(S|T)}$.

Definition 1: The conditional LZ complexity of sequence S relative to the conditional sequence T is $c(S|T)$, and $c(S|T) = c_{E(S|T)}$.

Note that the conditional LZ complexity of S relative to T equals the LZ complexity of S when T is null, namely $c(S|T) = c(S)$ if $T = \varphi$.

Given sequences S, Q, and T, the following inequalities can be deduced according to the definition of the minimum conditional production parsing and Inequalities (1) and (2):

$$c(S|TQ) \leq c(S|T), \quad c(S|QT) \leq c(S|T) \tag{4}$$

$$c(S|T) \leq c(SQ|T), \quad c(S|T) \leq c(QS|T) \tag{5}$$

Inequality (4) implies that the conditional LZ complexity of the given sequence will not be increased by concatenating a sequence after or before the conditional sequence. Inequality (5) implies that the conditional LZ complexity of the given sequence will not be decreased by concatenating a sequence after or before the original sequence. We present another inequality as the following:

$$c(SQ|T) \leq c(S|T) + c(Q|TS) \tag{6}$$

Proof: Let sequence $R = SQ$ and $c(SQ|T) = a$. The minimum conditional production parsing of R with given T is $E(R|T) = R(1,h_1)\ldots R(h_{a-1}+1,h_a)$. Assuming that the last symbol of sequence S, $s_{l(S)}$, lies in the k^{th} maximum conditional production component of sequence R with given T, then $E_k(R|T) = R(h_{k-1}+1,h_k)$, we have $(h_{k-1}+1) \leq l(S) \leq h_k$ and $c(S|T) = k$. Let sequence $L = R\big(l(S)+1,h_k\big)$ and sequence $M = R\big(h_k+1,l(R)\big)$, then it is obvious that $Q = LM$. $R(h_k+1,\, h_{k+1})\ldots R(h_{a-1}+1,\, h_a)$, a suffix of $E(R|T)$, happens to be the minimum conditional production parsing of sequence M relative to the conditional sequence TSL, that is, $E(M|TSL) = R(h_k+1,\, h_{k+1})\ldots R(h_{a-1}+1,\, h_a)$. Hence $c(M|TSL) = a - k = c(R|T) - c(S|T)$ and $c(R|T) - c(S|T) = c(M|TSL)$. For $LM = Q$, by Inequality (4), $c(M|TSL) \leq c(M|TS)$, and by Inequality (5), $c(M|TS) \leq c(LM|TS) = c(Q|TS)$. Since $c(R|T) - c(S|T) = c(M|TSL) \leq c(Q|TS)$, so $c(R|T) = c(SQ|T) \leq c(S|T) + c(Q|TS)$.

The following inequality indicates that conditional LZ complexity satisfies the triangle inequality:

$$c(Q|T) \leq c(Q|S) + c(S|T) \tag{7}$$

Proof: By Inequality (4), $c(Q|TS) \leq c(Q|S)$. By Inequality (5), $c(Q|T) \leq c(SQ|T)$. Adding the above two deduced inequalities, we have $c(Q|TS) + c(Q|T) \leq c(Q|S) + c(SQ|T)$, that is, $c(Q|T) \leq c(Q|S) + c(SQ|T) - c(Q|TS)$. By Inequality (6),

$c(SQ|T) - c(Q|TS) \leq c(S|T)$. Hence $c(Q|T) \leq c(Q|S) + c(S|T)$.

Distance metric of sequence LZ complexity

A distance metric defined on a set of objects should satisfy the following four conditions:

1. $d(x, y) > 0, \forall\, x \neq y$ (nonnegative);
2. $d(x, y) = 0, \forall\, x = y$ (identity);
3. $d(x, y) = d(y, x), \forall\, x \neq y$ (symmetry);
4. $d(x, y) \leq d(x, z) + d(z, y), \forall\, x, y, z$ (triangle inequality).

For nonnull sequences, we can measure the similarity between two sequences in quantity by computing their conditional LZ complexity. By Inequality (6), sequence conditional LZ complexity also satisfies the triangle inequality. However, sequence conditional LZ complexity is not in symmetry, thus it cannot be used as a sequence distance metric directly.

Therefore, based on conditional LZ complexity, we propose a distance measure between nonnull sequences as:

$$D(x, y) = \max\{c(x|y), c(y|x)\} \tag{8}$$

For nonnull sequences x, y, and z, by the definition of conditional LZ complexity, $D(x, y) > 0$ is always satisfied if $x \neq y$. The proposed distance also satisfies the identity condition up to an additive $O(1)$ term if $x = y$. It is obvious that $D(x, y)$ is in symmetry for every two sequences x and y. By Inequality (7), we have $c(x|y) \leq c(x|z) + c(z|y)$ and $c(y|x) \leq c(y|z) + c(z|x)$. Hence $\max\{c(x|y), c(y|x)\} \leq \max\{c(x|z), c(z|x)\} + \max\{c(z|y), c(y|z)\}$, which implies that $D(x, y)$ also satisfies the triangle inequality. Thus, the proposed distance is a valid distance metric. We call the proposed distance metric defined on nonnull sequences as *LZ complexity distance*.

Application

The mammalian phylogenetic relationship at the molecular level still remains to be a controversial topic in nowaday molecular genetics (*11*). Researches using different types of molecular data and analysis methods result in different conclusions to the debate about which two of the three main groups of placental mammals, namely Primates, Ferungulates, and Rodents, are more closely related. There are three possible phylogenetic trees, as shown in Figure 1, by introducing an outgroup that has comparatively close relationship to placental mammals into the phylogeny analysis. Alignment analysis using some proteins encoded by mitochondrial genome supports that the evolutional relationship between Primates and Rodents is more closely related (*12*). The reconstructed phylogenetic tree's topology suggested in these studies is [Ferungulates (Primates, Rodents)] (Figure 1B). However, alignment analysis using mitochondrial DNA (mtDNA) sequences (*13*) or some proteins encoded by nuclear genome (*14*) gives the tree's topology of [Rodents (Primates, Ferungulates)], which suggests that Primates and Ferungulates are more closely related (Figure 1A).

Motivated by the studies of Cao *et al* (*12*) and Reyes *et al* (*11*), we chose the whole mitochondrial genomes of 26 species of placental mammals as molecular data to reconstruct the phylogenetic tree of Eutherian orders. Similar to their studies, opsossum, wallaroo, and platypus were selected as the outgroup. All the 29 data files were obtained from the GenBank database, and the 29 species and their access numbers are listed in Table 1.

Firstly, 29 mtDNA sequences were extracted from the above 29 data files. Then the conditional LZ complexity between every two sequences was computed. The LZ complexity distances were measured according to Equation (8). Using the LZ complexity distances between sequences, a distance matrix was built

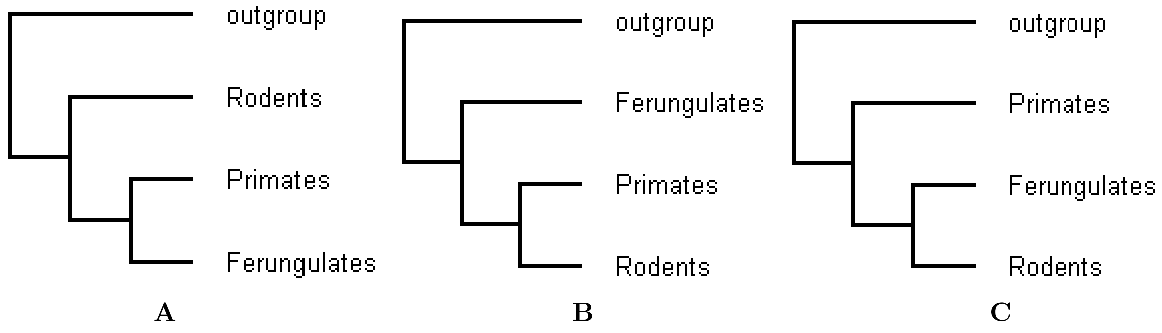

Fig. 1 Three possible trees among Primates, Ferungulates, and Rodents relative to the outgroup.

Table 1 The 29 Mammalian Species and Their GenBank Access Numbers

Group	Species	Access number
Primates	Human (*Homo sapiens*)	V00662
	Common chimpanzee (*Pan troglodytes*)	D38116
	Pigmy chimpanzee (*Pan paniscus*)	D38113
	Gorilla (*Gorilla gorilla*)	D38114
	Orangutan (*Pongo pygmaeus*)	D38115
	Gibbon (*Hylobates lar*)	X99256
	Baboon (*Papio hamadryas*)	Y18001
Ferungulates	White rhinoceros (*Ceratotherium simum*)	Y07726
	Harbor seal (*Phoca vitulina*)	X63726
	Gray seal (*Halichoerus grypus*)	X72004
	Cat (*Felis catus*)	U20753
	Fin whale (*Balenoptera physalus*)	X61145
	Blue whale (*Balenoptera musculus*)	X72204
	Cow (*Bos taurus*)	V00654
	Horse (*Equus caballus*)	X79547
	Donkey (*Equus asinus*)	X97337
	Great rhinoceros (*Rhinoceros unicornis*)	X97336
	Dog (*Canis familiaris*)	U96639
	Sheep (*Ovis aries*)	AF010406
	Pig (*Sus scrofa*)	AJ002189
	Hippopotamus (*Hippopotamus amphibius*)	AJ010957
Rodents	Rat (*Rattus norvegicus*)	X14848
	Mouse (*Mus musculus*)	V00711
	Squirrel (*Sciurus vulgaris*)	AJ238588
	Fat dormouse (*Glis glis*)	AJ001562
	Guinea pig (*Cavia porcellus*)	AJ222767
Outgroup	Opossum (*Didelphis virginiana*)	Z29573
	Wallaroo (*Macropus robustus*)	Y10524
	Platypus (*Ornithorhyncus anatinus*)	X83427

up. To reconstruct the phylogenetic tree, we utilized the neighbor-joining method (*15*) in PHYLIP software package of version 3.63 (*16*) and the TreeView tool of version 1.6.6 (*17*).

The phylogenetic tree reconstructed through the proposed LZ complexity distance method is shown in Figure 2. It indicates the topology of [Rodents (Primates, Ferungulates)] about the Eutherian orders' phylogeny, which is in accordance with the overall structure of the phylogeny presented in the studies of Cao *et al* (*12*) and Reyes *et al* (*11*). Furthermore, all branches in the tree completely agree with the result in Cao *et al* (*12*) and most of the clades conform to the result in Reyes *et al* (*11*) except for the position of guinea pig. As a species of nonmurid rodents, guinea pig is grouped into neither nonmurid rodents nor murid rodents, but shows an outgroup status relative to Primates, Ferungulates, and Rodents in Figure 2. Such an unexpected disagreement may suggest some deep biological implications, for the phylogenetic position of guinea pig stays as one of the most controversial topics in system biology (*18–20*).

In this study, we also reconstructed a phylogenetic tree using sequences of coding regions (data not shown). A total of 12 mitochondrial genes that encode 12 mitochondrial proteins were extracted from each of the 29 species' mitochondrial genomes. Then the 12 gene sequences corresponding to one species were concatenated to form a new mtDNA sequence. We computed the LZ complexity distance between every two of these 29 concatenated sequences and then built up a distance matrix from these data. Using the distance matrix, another phylogenetic tree was reconstructed and it was completely in accordance with the

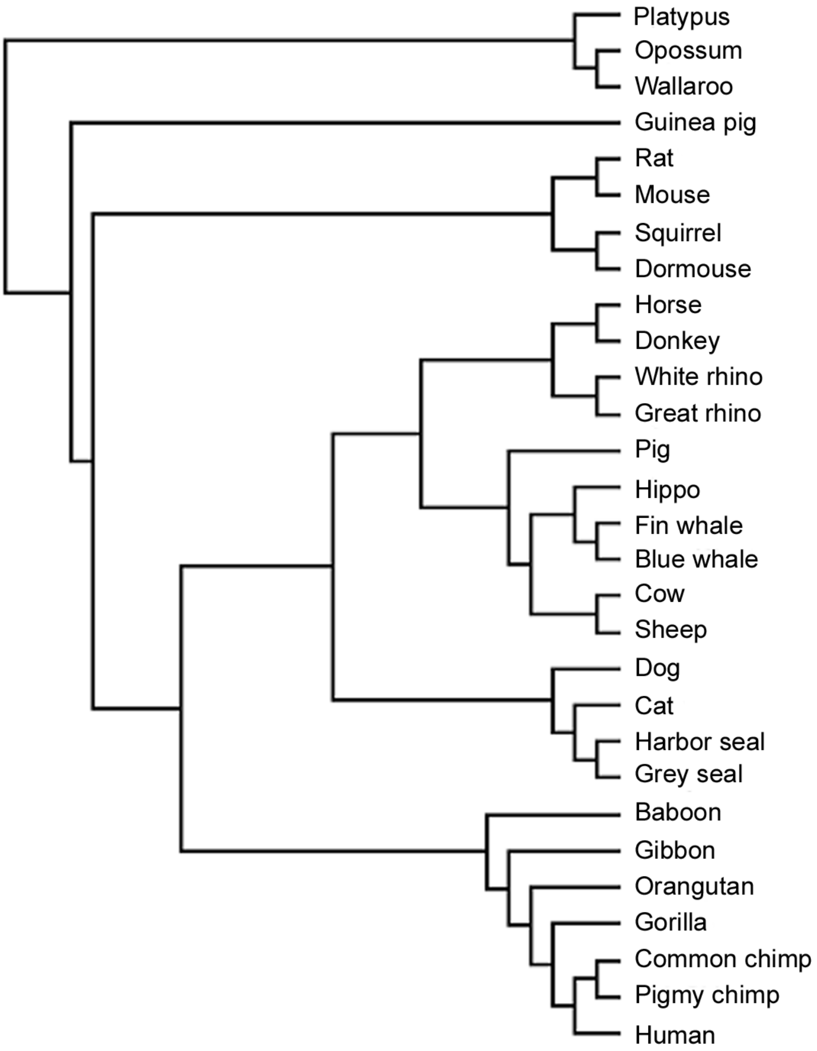

Fig. 2 The phylogenetic tree reconstructed from the mtDNA sequences of 26 species of placental mammals using LZ complexity distance, where opossum, wallaroo, and platypus were used as the outgroup.

tree shown in Figure 2. Phylogeny inferred through the above approach also implied that Primates and Ferungulates are more closely related.

Conclusion

The proposed sequence LZ complexity distance satisfies all the four conditions of distance metric theoretically and has been applied successfully to the phylogenetic tree reconstruction of 26 species of placental mammals. The phylogeny inferred through the LZ complexity distance measure is in agreement with the overall structure of some previous studies, which indicates the validity of using the proposed sequence LZ complexity distance to analyze the evolutionary relationship of DNA sequences in quantity. The computation of the proposed distance is totally automatic and alignment-free. Unlike most existing methods of phylogenetic tree reconstruction, the proposed method does not require gene identification nor any prior biology knowledge such as an accurate alignment score matrix.

Among the debate that which two of the three main groups of placental mammals, namely Primates, Ferungulates, and Rodents, are more closely related, the phylogenetic tree reconstructed based on the proposed sequence LZ complexity distance using whole mitochondrial genome supports the suggestion that Primates and Ferungulates are more closely related. In the reconstruction of the phylogenetic tree of 26 species of placental mammals, results obtained respectively from the complete mitochondrial genomes and some coding regions in mitochondrial genomes are both significant in biological sense. Thus we see that the proposed method works well without the limitations of coding sequences. The proposed sequence LZ complexity distance provides a new available choice to compare and analyze noncoding sequences abounded in genomes.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 60371046).

References

1. Hao, B.L. and Zhang, S.Y. 2002. *Handbook of Bioinformatics* (second edition). Shanghai Scientific and Technical Publishers, Shanghai, China.
2. Nei, M. and Kumar, S. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York, USA.
3. Misener, S. and Krawetz, S.A. (eds.) 2000. *Bioinformatics: Methods and Protocols*. Humana Press, Totowa, USA.
4. Vinga, S. and Almeida, J. 2003. Alignment-free sequence comparison—a review. *Bioinformatics* 19: 513-523.
5. Li, M. and Vitanyi, P. 1997. *An Introduction to Kolmogorov Complexity and Its Applications* (second edition). Springer-Verlag, New York, USA.
6. Li, M., *et al.* 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* 17: 149-154.
7. Chen, X., *et al.* 1999. A compression algorithm for DNA sequences and its applications in genome comparison. *Genome Inform. Ser. Workshop Genome Inform.* 10: 51-61.
8. Sato, H., *et al.* 2001. DNA data compression in the post genome era. *Genome Informatics* 12: 512-514.
9. Lempel, A. and Ziv, J. 1976. On the complexity of finite sequences. *IEEE Trans. Inf. Theory* 22: 75-81.
10. Kaspar, F. and Schuster, H.G. 1987. Easily calculable measure for the complexity of spatiotemporal patterns. *Phys. Rev. A* 36: 842-848.
11. Reyes, A., *et al.* 2000. Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. *Mol. Biol. Evol.* 17: 979-983.
12. Cao, Y., *et al.* 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J. Mol. Evol.* 47: 307-322.
13. Janke, A., *et al.* 1997. The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. *Proc. Natl. Acad. Sci. USA* 94: 1276-1281.
14. Kuma, K. and Miyata, T. 1994. Mammalian phylogeny inferred from multiple protein data. *Jpn. J. Genet.* 69: 555-566.
15. Satton, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic tress. *Mol. Biol. Evol.* 4: 406-425.
16. Felsenstein, J. 1989. PHYLIP—phylogeny inference package (version 3.2). *Cladistics* 5: 164-166.
17. Page, R.D. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* 12: 357-358.
18. Cao, Y., *et al.* 1997. Phylogenetic position of guinea pigs revisited. *Mol. Biol. Evol.* 14: 461-464.
19. Sullivan, J. and Swofford, D.L. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J. Mammal. Evol.* 4: 77-86.
20. Reyes, A., *et al.* 1998. Complete mitochondrial DNA sequence of the fat dormouse, *Glis glis*: further evidence of rodent paraphyly. *Mol. Biol. Evol.* 15: 499-505.