MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **66** (2011) 425-443

ISSN 0340 - 6253

Conditional LZ Complexity and Its Application in mtDNA Sequence Analysis

Wenwen Wang [*]

School of Sciences, China University of Mining and Technology,
Xuzhou, 221116, P. R. China

Tianming Wang

Department of Applied Mathematics, Dalian University of Technology,
Dalian, 116024, P. R. China

(Received September 9, 2010)

Abstract A DNA sequence is identified with a word over an alphabet $\sum = \{A, C, G, T\}$, where A, C, G, T are four bases of nucleic acids. In terms of classifications of the four bases, $(0, 1)$-characteristic sequences of a DNA sequence can be obtained, which can reveal its different functions. The conditional LZ complexity (CLZ) measure proposed in this paper is an alignment free method, which takes three $(0, 1)$-characteristic sequences of a DNA primary sequence as its side information. This method enables biologists to extract information from biological sequences according to their purpose. Further, based on CLZ complexity we present a new method to construct a phylogeny trees using complete unaligned mtDNA sequences. The proposed method relies on LZ complexity, which takes the $(0, 1)$-characteristic sequences as its side information. The method does not require sequence alignment and is totally automatic. Reasonable phylogeny trees are constructed by the method, which are largely in agreement with previously published trees based on the analysis of identical data sets.

1. Introduction

With the development of sequencing technique, a large number of DNA primary sequences data are collected into various data banks. Analysis of the corresponding evolutionary relationships of the species is becoming more and more important in bioinformatics.

Phylogeny is the study of the evolutionary history among the species. It can also provide information for function prediction and pharmaceutical. Researchers may use

[*] Corresponding author. wangwenwencumt@163.com

phylogenetic methods to determine which species are most closely related to other medicinal species, thus perhaps sharing their medicinal qualities [13].

Nowadays, the two commonly utilized methods for phylogeny analysis using biological molecular data, such as DNA, RNA and protein sequences, have been developed [8]. The first is distance matrix method, which is to produce a matrix by calculating the distances between every two sequences and then transforming this matrix into a tree by virtue of various algorithms. The second is the discrete data method, which is to search the tree based on certain optimal criteria, such as maximal parsimony method and maximum likelihood method.

Most of these methods require a multiple alignment of the sequences and select sequences evolutionary models, and often fail to work when the data sets become large and complex. Moreover, these methods are computationally expensive and do not produce correct results on events such as non-contiguous copies of a gene on the genome or non-decisive gene order. Many researchers are trying to develop efficient methods to overcome these problems. Bayesian methods are used for the phylogeny analysis of the sequences, which are based on maximum likelihood methods but incorporate prior probability [11, 30]. Gene content was proposed by Snel et al.[33] as a distance in genome phylogeny, which did not perform efficiently when the gene content of the organisms are very similar. Recently, a variety of efforts have been made to derive alignment-free methods to overcome this limitation [2, 21, 23, 39, 40]. For example, various graphical representations constitute a separate class of methods which aim to facilitate both numerical and visualization tools for similarity analysis of the sequences [18, 19, 20, 28, 37].

It is well-known that the regulatory regions of biological sequences are highly repetitive. They are rich in direct, symmetric and complemented repeats, and there is no doubt about the functional significance of these repeats [7]. One fundamental characteristic of linear symbolic sequence is sequence complexity, which has been defined by many methods, based on either algorithmic complexity or Shanoon entropy and used in genomic analysis.

Kolmogorov complexity, the first formal theoretical description of sequence complexity, was proposed by Kolmogorov from the view of algorithm information theory [15]. Li et al. [16] first introduced Kolmogorov complexity to DNA sequence analysis. Because Kolmogorov complexity is not computable, Chen et al. [5] 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 [31].

The LZ complexity proposed by Ziv and Lempel [38] is one of the most popular lossless measures. Furthermore, LZ complexity has various applications in the areas of information theory [36]. Many researchers used the LZ complexity method to analyze biological sequences. The method can efficiently extract information on repeated patterns encoded in DNA sequences. Otu and Sayood have proposed a new sequence distance measure based on the relative information between the sequences using LZ complexity. The algorithm they obtained can successfully construct consistent phylogenies for simulated and real date sets [26].

Motivated by the work of LZ complexity and considering the side information (such as the characteristic sequences of DNA sequences), we propose a new method to construct phylogeny trees by CLZ complexity measure [36]. In Section 2, we will give some basic definitions and properties about LZ complexity and CLZ complexity. In order to examine the validity of our method, we analyze the complete unaligned mtDNA sequences of 38 species in Section 3. In addition, we choose 20 sequences used in [26], and obtain two phylogenetic trees by our method. The results are agreement with previous results.

2. Method

DNA sequences can be treated as finite-length symbol strings over a four-letters alphabet $\sum := \{A, G, C, T\}$, where A, G, C, and T denote the four nucleic acid bases: adenine, guanine, cytosine and thymine, respectively. These four nucleotides are arranged linearly on each chain. Comparison of DNA primary sequences should be considered not only the string structures but also their chemical properties. Based on the chemical properties, one can classify the bases of DNA sequences and obtain the corresponding characteristic sequences.

In this section, we present three kinds of characteristic sequences of DNA sequences as well as LZ complexity in order to define the concept of CLZ complexity for the analysis of DNA sequences. This method serves as a basic module for further applications including phylogeny analysis.

2.1 $(0,1)$-Characteristic sequences of DNA sequences

In biology, analysis of DNA sequences is a very important task. One method, which is popular, relies on the characteristic sequences of given sequences according to the different classifications of the four bases $\{A, G, C, T\}$. Biologists generally classify the four bases into two groups: purine $\{A, G\}$ and pyrimidine $\{C, T\}$ according to their chemical structures. Another classification is based on the weak H-bond $\{A, T\}$ and strong H-bond $\{C, G\}$, which reflects the difference of the strength of hydrogen bonds. Furthermore, the four bases can be divided into amino group $\{A, C\}$ and keto group $\{T, G\}$. For convenience, these classifications are generally denoted by $\boldsymbol{R} = \{A, G\}$, $\boldsymbol{Y} = \{C, T\}$, and $\boldsymbol{W} = \{A, T\}$, $\boldsymbol{S} = \{C, G\}$, and $\boldsymbol{M} = \{A, C\}$, $\boldsymbol{K} = \{T, G\}$. In terms of the above three classifications, any DNA sequence can be transformed into three $(0,1)$-characteristic sequences [9], and the transformation rules are represented by $\boldsymbol{RY}$, $\boldsymbol{WS}$ and $\boldsymbol{MK}$, respectively. For sequence S, we define that

$$\boldsymbol{RY}(S(i)) = \begin{cases} 0, & \text{for } S(i) = A, G, \\ 1, & \text{for } S(i) = C, T; \end{cases} \tag{1}$$

$$\boldsymbol{WS}(S(i)) = \begin{cases} 0, & \text{for } S(i) = A, T, \\ 1, & \text{for } S(i) = C, G. \end{cases} \tag{2}$$

$$\boldsymbol{MK}(S(i)) = \begin{cases} 0, & \text{for } S(i) = A, C, \\ 1, & \text{for } S(i) = G, T; \end{cases} \tag{3}$$

Thus, we obtain three $(0,1)$-characteristic sequences for a DNA primary sequence, which are called $\boldsymbol{RY}(S)$, $\boldsymbol{WS}(S)$, and $\boldsymbol{MK}(S)$ characteristic sequences of the given sequence S. As an example, for the sequence $S := GTGGCAATGAT$, it can be transformed into the following three characteristic sequences: $\boldsymbol{RY}(S) = 01001001001$, $\boldsymbol{WS}(S) = 10111000100$ and $\boldsymbol{MK}(S) = 11110001101$.

On the one hand, the three characteristic sequences reveal the different functions about the given primary sequence. On the other hand, each characteristic sequence is a coarse-grained description for the DNA sequence, i.e., some information of the DNA primary sequence may be lost in a characteristic sequence so that different DNA primary sequences may have certain similar characteristic sequences [9]. However, the characteristic sequences do make it easier to compare sequences, and they reflect the functions of the classifications. Moreover, they also provide another chance for analyzing sequences from different aspects. Therefore, comparing the characteristic sequences has special significance to a extent, and in this paper, we will take these characteristic sequences as the side information to analyze DNA sequences.

2.2 LZ complexity

As a universal complexity measure, LZ complexity is valid to analyze of DNA sequences [26, 22]. Here we introduce some basic concepts about LZ complexity.

Give symbolic sequences S, Q and R defined over a finite alphabet Σ, the length of S is denoted by $l(S)$. Let $S(i)$ be the ith element of S and $S(i,j)$ be the substring of S that starts at position i and ends at position j. The concatenation of S and 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 such that $S = R(1,i)$.

An extension $R = SQ$ of S is reproducible from S (denoted by $S \to R$), if there exists an integer $p \leq l(S)$ such that $Q(k) = R(p+k-1)$ for $k = 1, \cdots, l(Q)$. A sequence S is producible from its prefix $S(1,j)$ (denoted by $S(1,j) \Rightarrow S$), if $S(1,j) \to S(1, l(S)-1)$. Note that production allows for an extra different symbol at the end of the copying process which is not permitted in reproduction.

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

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

where $1 \leq m \leq l(S)$, and $h_m = l(S)$. Based on the above process we obtain a parsing of S:

$$H(S) = S(1,h_1)S(h_1+1,h_2)\cdots S(h_{m-1}+1,h_m),$$

which is called the history of S. Additionally, we call $H_i(S) = S(h_i+1,h_i)$ the ith component of $H(S)$. As an example, for $S := AACGTACC$, $A \cdot A \cdot C \cdot G \cdot T \cdot A \cdot C \cdot C$, $A \cdot AC \cdot G \cdot T \cdot A \cdot C \cdot C$ and $A \cdot AC \cdot G \cdot T \cdot ACC$ are three different production histories of S.

A component $H_i(S)$ is called exhaustive, if $S(1,h_{i-1}) \to S(1,h_i)$ is not true. A history is called exhaustive if each of its components (except maybe the last one) is exhaustive. It has been proved by Lempel and Ziv [38] that the exhaustive history of any sequence is unique and the number of components in the exhaustive production history of S is the least possible number of steps that generate S according to the rules of production process. This number ia called the LZ complexity of the sequence S. For more details of the LZ complexity, the reader is referred to [5, 10, 35].

2.3 Conditional LZ complexity

Let X and Y be two finite alphabets. For $\boldsymbol{x} = x_1 x_2 \cdots x_n \in X^n$ and $\boldsymbol{y} = y_1 y_2 \cdots y_n \in Y^n$, the sequence

$$(\boldsymbol{xy}) = (x_1 y_1)(x_2 y_2) \cdots (x_n y_n) \in (XY)^n$$

of pairs of symbols is called a joint sequence.

Suppose that the joint sequence $(\boldsymbol{xy})$ is parsed into $c = c(\boldsymbol{x}, \boldsymbol{y})$ distinct words as follows:

$$(\boldsymbol{xy}) = (xy)_{n_0+1}^{n_1} (xy)_{n_1+1}^{n_2} \cdots (xy)_{n_{c-1}+1}^{n_c}, \tag{4}$$

where $(xy)_i^j$ denotes a subsequence $(x_i y_i)(x_{i+1} y_{i+1}) \cdots (x_j y_j)$ of the sequence $(\boldsymbol{xy})$ and $n_0 = 0, n_c = n$.

Let us apply the incremental parsing procedure of the LZ complexity algorithm to the sequence of pairs $(x_1 y_1) \cdots (x_n y_n)$. According to this procedure, $(\boldsymbol{xy})$ is parsed sequentially into phrases, where each new phrase is the shortest sequence that has not appeared earlier as a parsed phrase. Thus all phrases are distinct with a possible exception of the last phrase, which might be incomplete. Let $c(\boldsymbol{y})$ be the number of distinct phrases in the parsing of $\boldsymbol{y}$ induced by the parsing (4), and $y(l)$ be the lth distinct phrase in the induced parsing on $\boldsymbol{y}$.

For example, given a DNA sequence $S = AGTAACG\ TAATGTCCCAT$. Its exhaustive history is $x = H(S) = A \cdot G \cdot T \cdot AA \cdot C \cdot GTAAT \cdot GTC \cdot CCA \cdot T$. Consider the characteristic sequence $\boldsymbol{RY}(S)$, then the corresponding parsing is $\boldsymbol{y} = \boldsymbol{RY}(S) = 0 \cdot 0 \cdot 1 \cdot 00 \cdot 1 \cdot 01001 \cdot 011 \cdot 110 \cdot 1$, and $c(\boldsymbol{y}) = 6, y(1) = 0, y(2) = 1, y(3) = 00, y(4) = 01001, y(5) = 011, y(6) = 110$.

The CLZ complexity of $\boldsymbol{x}$ with side information $\boldsymbol{y}$ induced by the parsing (4) is defined by

$$C(\boldsymbol{x}) = \sum_{l=1}^{c(\boldsymbol{y})} c_l(\boldsymbol{x}|\boldsymbol{y}) \log c_l(\boldsymbol{x}|\boldsymbol{y}), \tag{5}$$

where $c_l(\boldsymbol{x}|\boldsymbol{y})$ is the number of distinct $\boldsymbol{x}$ phrases that appear jointly with $y(l)$ for all $l = 1, \cdots c(\boldsymbol{y})$, which is a minor modification of the definition proposed by Uyematsu and Kuzuoka [36]. Now, the influence of the length of $\boldsymbol{x}$ is canceled. In the above example, $c_1(\boldsymbol{x}|\boldsymbol{y}) = 2$, $c_2(\boldsymbol{x}|\boldsymbol{y}) = 3$, $c_3(\boldsymbol{x}|\boldsymbol{y}) = c_4(\boldsymbol{x}|\boldsymbol{y}) = c_5(\boldsymbol{x}|\boldsymbol{y}) = c_6(\boldsymbol{x}|\boldsymbol{y}) = 1$. Applying the formula (5), we have

$$C(\boldsymbol{x}) = 2 \cdot \log 2 + 3 \cdot \log 3 + 4 \cdot \log 1 = 4.6821.$$

The CLZ complexity $C(\boldsymbol{x})$ is an important complexity indicator which is associated with our distance measure.

2.4 Distance metric of CLZ complexity

According to LZ complexity, for any given sequences S and Q, we consider the sequence SQ and its exhaustive history. Otu and Sayood [26] have pointed out that the more similar the sequence S is to Q, the smaller $c(SQ) - c(S)$ is. On the basis of the LZ complexity, we get the CLZ complexity measure by taking the characteristic sequences as their side information. In a similar way, we also believe that the number $C(SQ) - C(S)$ is smaller, which shows that the sequences S and Q are more similar. That is to say, $C(SQ) - C(S)$ depends on how much Q is similar to S.

Here we give a simple example. Consider three DNA sequences S, Q and R:

$$S = AAGGGGTGAAGCTT,$$

$$Q = AAGGCGTGAATCCT,$$

$$R = CCGCAATGTGACTT.$$

We analyze the similarity between them by computing the CLZ complexity with the characteristic sequence $\boldsymbol{RY}$. According to the formula (5), $C(S) = 1.3863, C(SQ) = 2.7726$ and $C(SR) = 4.6821$. After some computation, it yields $C(SQ) - C(S) = 1.3863, C(SR) - C(S) = 3.2958$. From the results, we know sequence S is closer Q than R which is identical with the fact.

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

$$d(S,Q) = \begin{cases} max\{C(SQ) - C(S), C(QS) - C(Q)\}, & if\ S \neq Q; \\ 0, & if\ S = Q. \end{cases}$$

as the relative distant measure between S and Q, where $C(S)$, $C(Q)$, $C(SQ)$ and $C(QS)$ denote the CLZ complexities associated with one characteristic sequence of the corresponding sequences, respectively. According to Otu and Sayood [26] presented the Lemma 1 and Theorem 1, we can know $d(S,Q)$ also satisfies the following four conditions:

(1) $d(S,Q) \geq 0$, where the equality is satisfied iff $S = Q$ (identity);

(2) $d(S,Q) = d(Q,S)$ (symmetry);

(3) $d(S,Q) \leq d(S,R) + d(R,Q)$ (triangle inequality);

(4) $d(S,Q) + d(R,P) \leq max\{d(S,R) + d(Q,P), d(S,P) + d(Q,R)\}$ $(additivity)$.

From the formula (6), the distance between the sequences S and Q can therefore be calculated, i.e., $d(S,Q) = 1.3863$. Moreover, let us consider $C(SS)$, which denotes the sequence obtained by catenation of sequence S to itself. In general, $C(SS) \neq C(S)$. However, $C(SS) - C(S)$ is always smaller than the other $C(SQ) - C(S)$ if $S \neq Q$. So we may define $d(S,Q) = 0$ if $S = Q$ and take $d(S,Q)$ as a special distance measure.

3. Application

Today in molecular genetics the mammalian phylogenetic relationship at the molecular level still is a controversial topic [29]. 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 [17]. There are three possible phylogeny trees by introducing an outgroup, which are shown in Figure 1. Recently, many efforts has been done on the phylogenetic relationships among major groups of Eutherian. The evolutional relationship between Primates and Ferungulates is more closely by analysis of complete mtDNA sequences [4, 12] and is in agreement with the published results of several proteins encoded by nuclear DNA [3, 14, 16]. However, Stanhope et al. [32] and Porter et al. [27] give the tree's topology of [Ferungulates (Rodents, Primates)] from the analysis of IRBP, which suggests that Primates and Rodents are more closely related (Figure 1A).

Motivated by the studies of Cao et al. [4], Otu and Sayood [26] and Reyes et al. [29], we apply the proposed distance measure (6) to the complete mitochondrial genomes of 20 species of placental mammals to reconstruct the phylogeny tree of Eutherian orders. Note that wallaroo, opossum and platypus are used as outgroup. All the 38 data files are obtained from the GenBank database (http://www.ncbi.nlm.nih.gov), and the 38 species

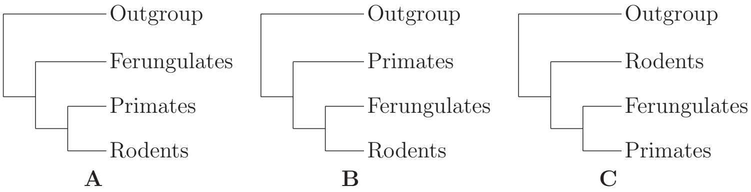


Figure 1: The possible trees among Primates, Ferungulates and Rodents relative to the Outgroup.

and their access numbers are listed in Table 1.

In this paper, we don't eliminate the effect of the length on the distance measure $d(S,Q)$, that is to say, $d(S,Q)$ are unnormalized in terms of the length of the sequences. However, we choose the complete mitochondrial genomes of the 38 species to verify the validity of our method, and Table 1 shows the length of each species with a very small difference, so that we can apply the $d(S,Q)$ in this paper. Of course, we can also give the normalized forms of $d(S,Q)$, just as $d^*(S,Q)$ and $d_1^{**}(S,Q)$ given in [26]. Here, our main concern is the new similarity measure

$$d(S,Q) = \begin{cases} max\{C(SQ) - C(S), C(QS) - C(Q)\}, & if\ S \neq Q; \\ 0, & if\ S = Q. \end{cases}$$

Table 1. The 38 Mammalian Species and their GenBank Access Numbers

Group	Species	Access number	Lengths
Primates	human (Homo sapiens)	V00662	16569
	common chimpanzee (Pan troglodytes)	D38116	16563
	pigmy chimpanzee (Pan paniscus)	D38113	16554
	gorilla (Gorilla gorilla)	D38114	16364
	orangutan (Pongo pygmaeus)	D38115	16389
	gibbon (Hylobates lar)	X99256	16472
	baboon (Papio hamadryas)	Y18001	16521
	capuchin (Cebus albifrons)	AJ309866	16554
	tarsier (Tarsius bancanus)	AF348159	16972
	slow Loris (Nycticebus coucang)	AJ309867	16764
Ferungulates	pig (Sus scrofa)	AJ002189	16680
	alpaca (Lama pacos)	Y19184	16652
	cow (Bos taurus)	V00654	16338
	sheep (Ovis aries)	AF010406	16616
	hippo (Hippopotamus amphibius)	AJ010957	16407
	blue whale (Balenoptera musculus)	X72204	16402
	fin whale (Balenoptera physalus)	X61145	16398
	sperm whale (Physeter macrocephalus)	AJ277029	16428
	donkey (Equus asinus)	X97337	16670

	horse (Equus caballus)	X79547	16660
	India rhino (Ceratotherium simum)	X97336	16829
	white rhinoceros (Ceratotherium simum)	Y07726	16832
	cat (Felis catus)	U20753	17009
	dog (Canis familiaris)	U96639	16727
	black bear (Ursus americanus)	AF303109	16841
	polar bear (Ursus maritimus)	AF303111	17017
	gray seal (Halichoerus grypus)	X72004	16797
	harbor seal (Phoca vitulina)	X63726	16826
Rodents	rat (Rattus norvegicus)	X14848	16300
	mouse (Mus musculus)	V00711	16295
	vole (Volemys kikuchi)	AF348082	16312
	squirrel (Sciurus vulgaris)	AJ238588	16507
	dormouse (Myoxus glis)	AJ001562	16602
	guinea pig (Cavia porcellus)	AJ222767	16801
	cane rat (Thryonomys swinderianus)	AJ301644	16626
Outgroup	opossum (Didelphis virginiana)	Z29573	17084
	wallaroo (Macropus robustus)	Y10524	16896
	platypus (Ornithorhyncus anatinus)	X83427	17019

We obtain three distance matrices D_1, D_2 and D_3 from the three transformation rules (1)-(3). Finally, we put the pairwise distance matrix D_1, D_2 and D_3 into the Neighbor program in the PHYLIP package [6, 24]. For making these larger trees, the rapid Unweighted Pair Group Method with Arithmetic Mean (UPGMA) option is chosen over the NJ option as a balance between speed and rigor. We obtain three phylogeny trees drawn by TreeView program (Page 1996) when choosing the UPGMA option. See Figures 2, 3 and 4.

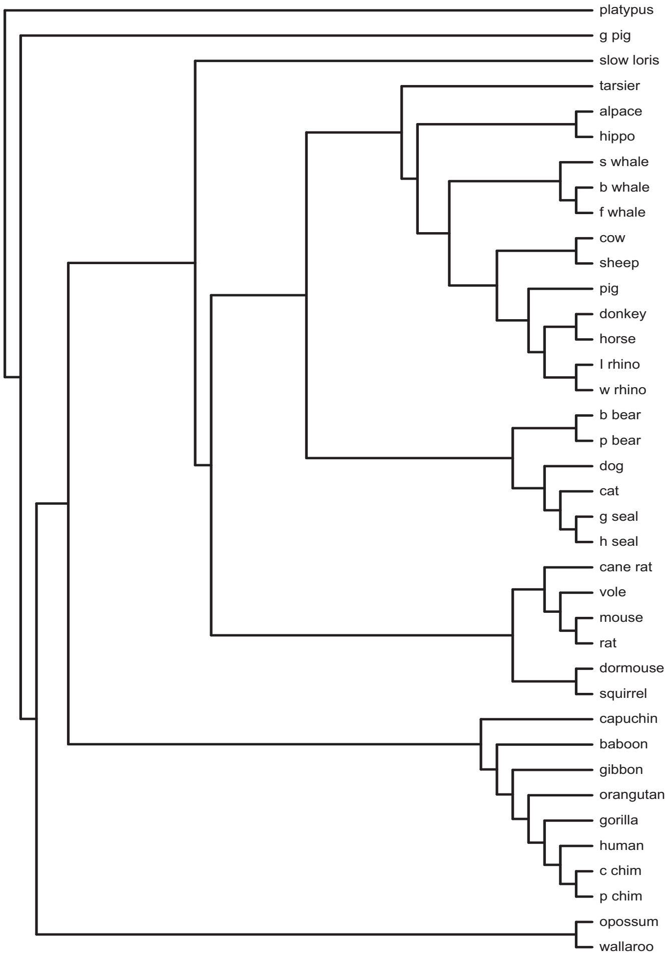


Figure 2: The phylogeny tree for the 38 cmDNA sequences taking $\boldsymbol{RY}$ characteristic sequence as side information by our method.

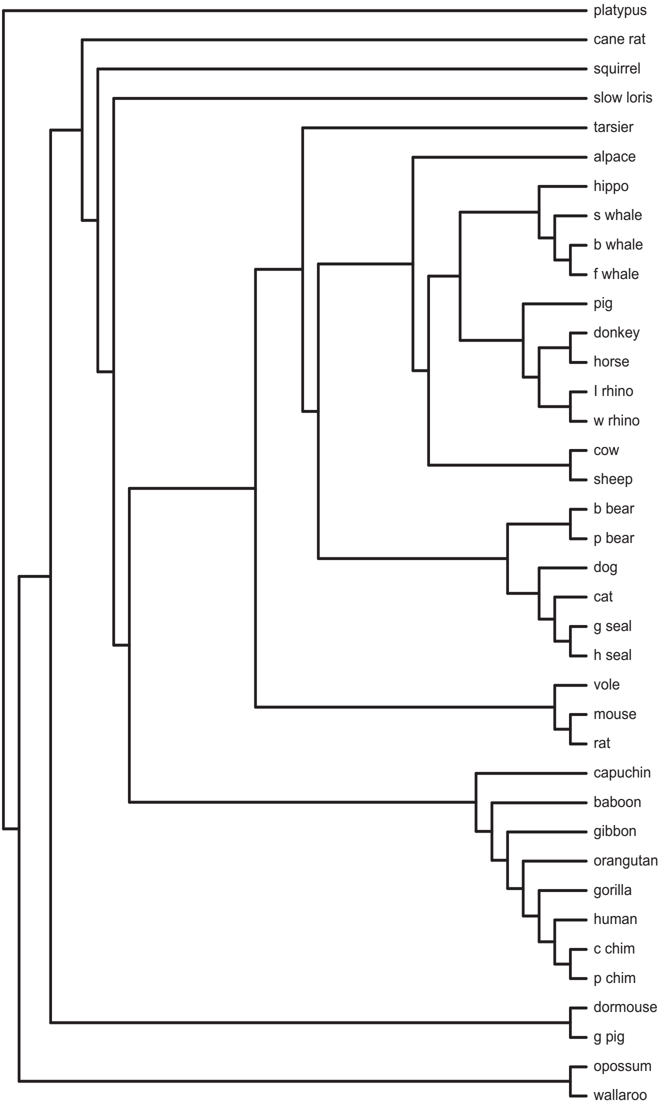


Figure 3: The phylogeny tree for the 38 cmDNA sequences taking ***WS*** characteristic sequence as side information by our method.

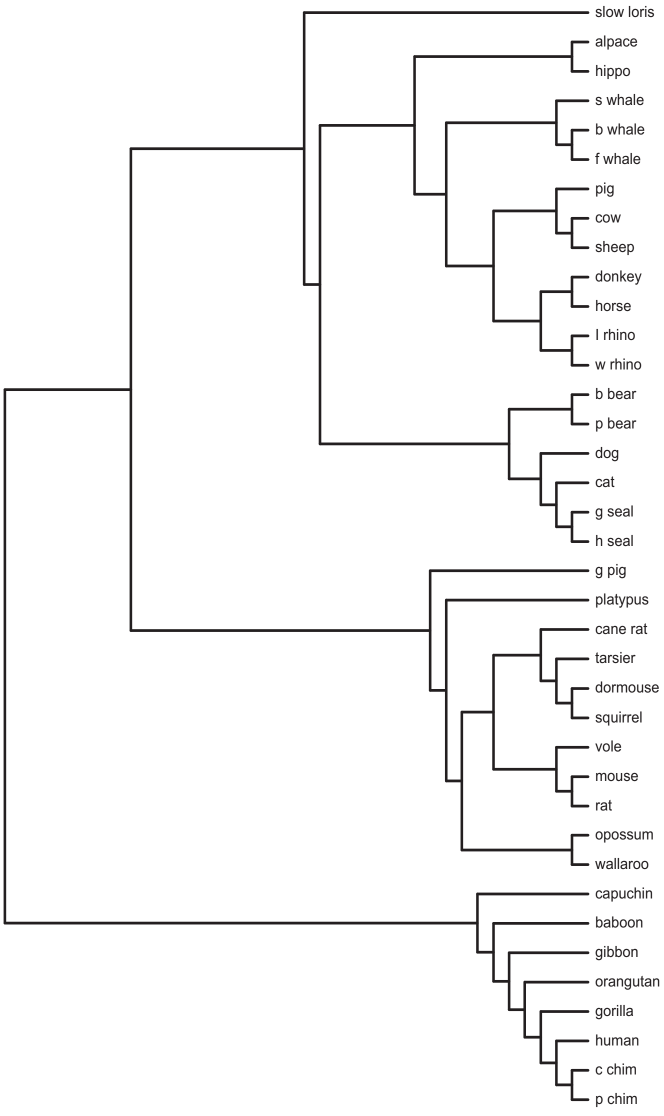


Figure 4: The phylogeny tree for the 38 cmDNA sequences taking ***MK*** characteristic sequence as side information by our method.

Comparing the trees of above Figures, we can find that the results of Figure 2 is more consistent to reality. Such as:

(1) Baboon, capuchin, gibbon, orangutan, gorilla, human, pigmy chimpanzee and common chimpanzee are grouped closely (they belong to Primates), where human are most closely related to chimpanzee [1, 25].

(2) Cat, dog, black bear, polar bear, harbor seal and gray seal belong to Carnivora. Horse, donkey, India rhinoceros and white rhinoceros belong to Perissodactyla. Cow, pig, sheep, hippo, alpaca, sperm whale, fin whale and blue whale are in Cetartiodactyla. All of them are grouped closely (they all belong to Ferungulates).

(3) Rat, mouse, vole, cane rat, dormouse and squirrel are in the same group of Rodents.

(4) Platypus is the only non-mammal, and wallaroo, opossum are two most remote species from the remaining mammals. The three species are separated from others.

It shows that, by our method, all 38 species have been separated well and almost grouped into corresponding structural classes. However, there exists some minor disappointed results. For example, tarsier should belong to the branch of Primates, but from Figure 2 we can see that it is much closer to the species of Ferungulates. And about the position of guinea pig still remains to be a controversial topic [4, 26]. Particularly, our method confirms the outgroup status of Primates relative to Ferungulates and Rodents [Primates (Ferungulates, Rodents)], which is largely in accordance with the results given in (Stuart et al. Figures 5 and 7 [34]).

Because we take three characteristic sequences as side information, there exists minor differences between the three trees. But from the whole level, we can find that the $\boldsymbol{RY}$ characteristic sequence reflect the relationship of each species much better than another. It is known that $\boldsymbol{RY}$ characteristic sequences mainly shows the chemical structures of DNA sequences, so that we think the chemical property of sequences may be a key factor to analyze the relationships of mammals species.

Moreover, we choose 20 species used in [26] from Table 1 to verify the validity of our method. Based on the same analysis, the three phylogentic trees using UPGMA algorithm are identical with each other, just as shown in the following Figure (a), which is agreement with the results given in [16, 17, 26, 34]. Additionally, we construct a similar phylogeny tree by virtue of the above distance measure using the Kitsch algorithm within PHYLIP package [6]. It is shown in the following Figure (b). The result is also in agreement with the tree's topology of [Primates (Ferungulates, Rodents)].

From above Figures, it can be found that our method is reasonable and valid. Comparing $(0,1)$-characteristic sequences, we can get some information that can not be obtained from direct comparisons of DNA sequences, and find some characteristics of given species from different aspects. Although some information may be lost in the process of the transformation into different characteristic sequences, we can focus our attention on the information we are interested in. To a great extent, this is an advantage of our method.

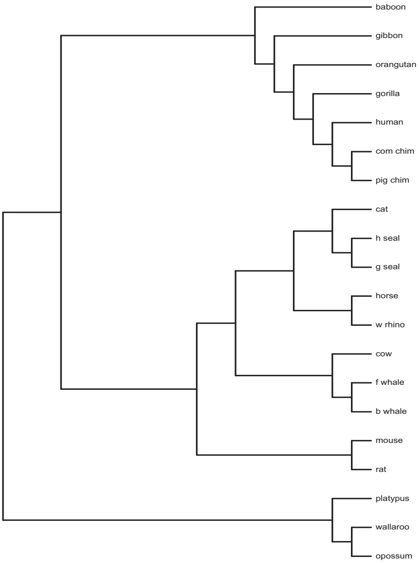
4. Conclusions

Recently, LZ algorithm has been introduced into bioinformatics. The main advantage of this algorithm is that it can extract repeated patterns from biological sequences. LZ complexity can be applied to the sequence with its side information. In this paper, we propose a new CLZ complexity associated with $(0,1)$-characteristic sequences as its side information to analyze DNA sequences. The phylogeny trees inferred by CLZ complexity is almost in accordance with published results, which indicates that our method is a reliable computational approach to the construction of DNA phylogeny trees.

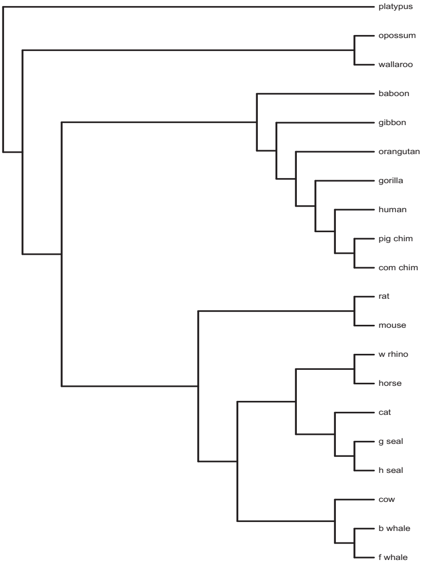
The computation for whole genome comparison and phylogeny using the proposed distance dose not require multiple alignment and is totally automatic. Moreover, it also shows that for the phylogeny analysis of DNA sequences, our method is a fast one. The shortage of our method is that some information may be lost when the DNA sequences are transformed into their corresponding characteristic sequences. However, the applications indicate that, under such circumstances, phylogeny trees using complete unaligned mtDNA sequences can still be successfully constructed.

Finally, it is worth noting that our distance measures do not use any evolutionary model and seem to be more fitting for whole genome phylogenies where current evolutionary models do not apply directly.

In the future, we shall apply our approach to comparison-based biological sequence research, such as multiple sequence alignment. We think that the CLZ complexity measure may be a keystone of analyzing biological sequences, and it will be used to most places, where LZ complexity measure can be used to.



(a) The phylogeny tree for the 20 cmDNA sequences using UPGMA algorithm.



(b) The phylogeny tree for the 20 cmDNA sequences using Kitsch algorithm.

Acknowledgments

We are grateful to the anonymous referees for their various valuable comments and thank Dr. Weiping Wang for his kind help. This work is supported financially by National Natural Science Foundation of China (10871219), the Special Foundation for Talents of China University of Mining and Technology, the Youth Foundation of China University of Mining and Technology (NO. OK091261) and the Fundamental Research Funds for the Central Universities (2010LKSX03).

References

[1] W. J. Bailey, K. Hayasaka, C. G. Skinner, S. Kehoe, L. C. Sieu, J. L. Slightom, M. Goodman, Reexamination of the African hominoid trichotomy with additional sequences from the primate beta-globin gene cluster, *Mol. Phylog. Evol.* **1** (1992) 97–135.

[2] D. Bielinska–Waz, S. Subramaniam, Classification studies based on a spectral representation of DNA, *J. Theor. Bio.* **266** (2010) 667–674.

[3] M. Bulmer, K. Wolfe, P. Sharp, Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders, *Proc. Natl. Acad. Sci. USA.* **88** (1991) 5974–5978.

[4] Y. Cao, A. Janke, P. J. Waddell, M. Westerman, O. Takenaka, S. Murata, N. Okada, S. Paabo, M. Hasegawa, Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders, *J. Mol. Evol.* **47** (1998) 307–322.

[5] X. Chen, S. Kwong, M. Li, A compression algorithm for DNA sequences and its applications in genome comparison, *Genome Inform. Ser. Workshop Genome Inform.* **10** (1999) 51–61.

[6] J. Felsenstein, PHYLIP–phylogeny inference package (version 3.2), *Cladistics* **5** (1989) 164–166.

[7] V. D. Gusev, L. A. Nemytikova, N. A. Chuzhanova, On the complexity measures of genetic sequences, *Bioinformatics* **15** (1999) 994–999.

[8] B. L. Hao, S. Y. Zhang, *Handbook of Bioinformatics*, Shanghai Sci. Tech. Publishers, Shanghai, China, 2002.

[9] P. A. He, J. Wang, Characteristic Sequences for DNA Primary Sequence, *J. Chem. Inf. Comput. Sci.* **42** (2002) 1080–1085.

[10] S. Hisahiko, Y. Takashi, DNA date compression in the post genome era, *Genome Informatics* **12** (2001) 512–514.

[11] J. P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees, *Bioinformatics* **17** (2001) 754–755.

[12] A. Janke, X. Xu, U. Arnason, The complete mitochondrial genome of the wallaroo (Macropus robustus) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria, *Proc. Natl. Acad. Sci.* **94** (1997) 1276–1281.

[13] K. Komatsu, S. Zhu, H. Fushimi, T. K. Qui, S. Cai, S. Kadota, Phylogenetic analysis based on 18s rRNA gene and matk gene sequences of Panax vietnamensis and five related species, *Planta Med.* **67** (2001) 461–465.

[14] K. Kuma, T. Miyata, Mammalian phylogeny inferred from multiple protein data, *Jpn. J. Genet.* **69** (1994) 555–566.

[15] M. Li, P. M. B. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Approximations*, Springer–Verlag, New York, 1997.

[16] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, H. Y. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics* **17** (2001) 149–154.

[17] B. Li, Y. B. Li, H. B. He, LZ complexity distance of DNA sequences and its application in phylogenetic tree reconstruction, *Geno. Prot. Bioinfo.* **3** (2005) 206–212.

[18] B. Liao, X. Y. Xiang, W. Zhu, Coronavirus phylogeny based on 2D graphical representation of DNA sequence, *J. Comput. Chem.* **27** (2006) 1196–1202.

[19] B. Liao, W. Zhu, Y. Liu, 3D graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree, *MATCH Commun. Math. Comput. Chem.* **56** (2006) 209–216.

[20] B. Liao, X. Z. Shan, W. Zhu, R. F. Li, Phylogenetic tree construction based on 2D graphical representation, *Chem. Phys. Lett.* **422** (2006) 282–288.

[21] B. Liao, L. J. Liao, G. X. Yue, R. H. Wu, W. Zhu, A vertical and horizontal method for constructing phylogenetic tree, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 691–700.

[22] N. Liu, T. M. Wang, A relative similarity measure for the similarity analysis of DNA sequences, *Chem. Phys. Lett.* **498** (2005) 307–311.

[23] Z. Liu, B. Liao, W. Zhu, A new method to analyze the similarity based on dual nucleotides of the DNA sequence, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 541–552.

[24] C. D. Michener, R. R. Sokal, A quantitative approach to a problem in classification, *Evolution* **11** (1957) 130–162.

[25] M. M. Miyamoto, J. L. Slightom, M. Goodman, Phylogenetic relations of humans and African apes from DNA sequences in the psi eta–globin region, *Science* **238** (1987) 369–373.

[26] H. H. Otu, K. Sayood, A new sequence distance measure for phylogenetic tree con-

struction, *Bioinformatics* **19** (2003) 2122–2130.

[27] C. Porter, M. Goodman, M. Stanhope, Evidence on mammalian phylogeny from sequences of exon 28 of the von Willebrand factor gene, *Mol. Phylogenet. Evol.* **5** (1996) 89–101.

[28] M. Randić, M. Vračko, N. Lers, D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, *Chem. Phys. Lett.* **371** (2003) 202–207.

[29] A. Reyes, C. Gissi, G. Pesole, F. M. Catzeflis, C. Saccone, Where do rodents fit? Evidence from the complete mitochondrialgenome of Sciurus vulgaris, *Mol. Biol. Evol.* **17** (2000) 979–983.

[30] F. Ronquist, J. P. Huelsenbeck, MrBayes 3: Bayesian phylogenetic inference under mixed models, *Bioinformatics* **19** (2003) 108–110.

[31] H. Sato, T. Yoshioka, A. Konagaya, T. Toyoda, DNA data compression in the post genome era, *Genome Informatics* **12** (2001) 512–514.

[32] M.J. Stanhope, M. R. Smith, V. G. Waddell, C. A. Porter, M. S. Shivji, M. Goodman, Mammalian evolution and the interphotoreceptor retinoid binding protein (IRBP) gene: convincing evidence for several superordinal clades, *J. Mol. Evol.* **43** (1996) 83–92.

[33] B. Snel, P. Bork, M. A. Huynen, Genome phylogeny based on gene content, *Nature Genet.***19** (1999) 1572–1574.

[34] G. W. Stuart, K. Moffet, S. Baker, Integrated gene and species phylogenies from unligned whole genome protein sequences, *Bioinformatics* **18** (2002) 100–108.

[35] T. Uyematsu, Lempel–Ziv coding as a tool for information theory, *IEICE Trans. Fundamentals (Japanese Edition)* **84** (2001) 681–690.

[36] T. Uyematsu, S. Kuzuoka, Conditional Lempel–Ziv complexity and its application to source coding theorem with side information, *IEICE Trans. Fundamentals Yokohama Japan* **86** (2003) 2615–2618.

[37] W. P. Wang, B. Liao, T. M. Wang, A graphical method to construct phylogenetic tree, *Int. J. Quantum Chem.* **106** (2006) 1998–2006.

[38] J. Ziv, A. Lempel, Compression of individual sequences by variable rate coding, *IEEE Trans. Inform. Theory* **24** (1978) 530–536.

[39] S. Zhang, T. Wang, Phylogenetic analysis of protein sequences based on conditional LZ complexity, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 701–716.

[40] W. Zhu, B. Liao, R. Li, A novel method for constructing phylogenetic tree based on a dissimilarity matrix, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 483–492.