J Math Chem (2009) 46:1137–1148
DOI 10.1007/s10910-008-9504-2

ORIGINAL PAPER

Phylogenetic inference from binary sequences reduced by primary DNA sequences

Xiaoqi Zheng · Yongchao Dou · Jun Wang

Received: 5 January 2008 / Accepted: 31 October 2008 / Published online: 10 December 2008

Abstract Given a bi-classification of nucleotides, we can obtain a reduced binary sequence of a primary DNA sequence. This binary sequence will undoubtedly retain some biological information and lose the rest. Here we want to know what kind of and how much biological information an individual binary sequence carries. Three classifications of nucleotides are explored in the present article. Phylogenetic trees are built from these binary sequences by the Neighbor-Joining (NJ) method, with evolutionary distance evaluated on the basis of a symbolic sequence complexity. We find that, for all data sets studied, binary sequences reduced by the purine/pyrimidine classification give reliable phylogeny (almost the same as that from the primary sequences), while the other two result in discrepancies at different levels. Some possible reasons and a simple model of sequence evolutionary are introduced to interpret this phenomenon.

Keywords DNA · Chemical classification · Characteristic sequence · LZ complexity · Phylogenetic tree · Evolutionary model

X. Zheng · Y. Dou
Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, People's Republic of China

X. Zheng
College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, People's Republic of China

J. Wang (✉)
Department of Mathematics, Shanghai Normal University, Shanghai 200234, People's Republic of China
e-mail: junwang@dlut.edu.cn

J. Wang
Scientific Computing Key Laboratory of Shanghai Universities, Shanghai 200234, People's Republic of China

1 Introduction

Accumulation of protein and DNA sequence data has greatly deepened the understanding of evolution. But meanwhile, these data also raise a fundamental question to biologists: how to process them efficiently? Traditional attempts are focused on proposing new algorithms and revising the ongoing methods to deal with primary sequences. For example, some sequence comparisons based on short words composition and graphical representations have been developed [1–3], while some motif finding (MF) algorithms are studied as revisions of the well known sequence alignment (e.g., [4,5]). In the present work, a new scheme to analyze and compare biological sequences is outlined.

When studying an object, we may be concerned about some specific properties of it and ignore others. Mathematically, we can achieve this aim by a map from one object to another reduced by the original, such as homomorphism in algebra, homotopy and homology in topology, and fiber bundles in differential geometry. From this consideration, we studied a "coarse grained" description of primary DNA sequence as follows [6]. In the first step, nucleotides are classified into different groups according to their chemical structures, e.g., purine $R = \{A,G\}$ and pyrimidine $Y = \{C,T\}$. Based on this classification, the (R,Y)-characteristic sequence (or (R,Y)-sequence, for short) of a DNA sequence is then obtained by replacing elements of R with 1 and Y with 0. Likewise, the (M,K)-sequence and the (S,W)-sequence can be constructed according to the classifications $M = \{A,C\}$, $K = \{G,T\}$ and $S = \{G,C\}$, $W = \{A,T\}$, respectively. It is easy to show that each pair of characteristic sequences can reconstruct the primary sequence. In this sense, no information is lost during these reductions. Moreover, an individual binary sequence may carry a certain kind of information, this provides us a new way to analyze biological information from different aspects. In recent years, some related binary representations of biological molecules have been used broadly in graphical representation and comparison of DNA sequences [7,8], phylogenetic tree construction [9], gene identification [10], etc.

In this article, we aim to address this subject from the evolutionary perspective, i.e., we are concerned how much evolutionary information an individual binary sequence carries. In other words, whether molecular phylogeny can be reconstructed from only one kind of these binary sequences? In Sect. 2, a universal distance measure for symbolic sequences, which makes use of a sequence complexity (Lempel-Ziv complexity), is introduced. In Sect. 3, three widely used datasets are considered. For each dataset, its corresponding (R,Y)-sequences ((M,K)- and (S,W)-sequences, respectively) are got and used to reconstruct phylogeny. Through comparing these phylogenies with that built from primary DNA sequences and the commonly accepted one, we draw a conclusion that (R,Y)-sequences can solely determine a reliable phylogeny, while phylogenies from (M,K)- and (S,W)-sequences have discrepancies at different levels. Then, some possible reasons for this phenomenon and a simple model of sequence evolution are presented. Our result could serve as an alternative method of phylogenetic tree construction and help to understand the mechanism of sequence evolution.

2 Computational method

Phylogenetic trees are usually obtained by the following three methods, maximum parsimony, maximum likelihood and distance method. We now use distance method, which seeks to reconstruct the tree topology that best represents a matrix of distances between pairs of taxonomic units. To get an accurate distance tree, a reliable measure of evolutionary distances between sequences is crucial. Here we make use of a certain complexity measure from information theory.

2.1 Lempel-Ziv (LZ) complexity of symbol sequences

LZ complexity, proposed by Lempel and Ziv to measure the randomness of finite sequences, is an easily computable and universal depiction of sequence complexity [11–13]. This complexity measure is related to the number of distinct substrings (i.e., patterns) and the rate of their occurrence along a given sequence [11].

For linear sequences S, T and R defined over a finite alphabet $\mathcal{A}$, let $L(S)$ 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. The sequence S is called an extension of T if S is the concatenation of T and R, i.e. $S = TR$. In the following paragraph, two types of extension are defined.

An extension $S = TR$ of T is called *reproducible* (denoted $T \rightarrow S$), if there exists an integer $m \leq L(T)$ such that $R(k) = S(m + k - 1)$, for $k = 1, 2, \ldots, L(R)$. For example WEST→WESTES with $m = 2$, and AACGT → AACGTCGTCG with $m = 3$. Similarly, an extension $S = TR$ of T is called *producible* (denoted $T \Rightarrow S$) if $T \rightarrow S(1, L(S) - 1)$. That is to say, an extra 'different' symbol at the end of the producible extension is allowed. For example, AACGT ⇒ AACGTCGTCGW. Thus we can say if $T \rightarrow S$ then $T \Rightarrow S$, but the reverse is not always true. If $T \Rightarrow S$ and $T \nrightarrow S$, this extension is called *exhaustive*.

According to the above definitions, any sequence S can be generated from the null sequence using iterative producible processes. For example, $S =$ AACACG can be generated by the following steps: $\phi \Rightarrow$ A⇒ AA ⇒ AAC ⇒ AACA ⇒ AACAC ⇒ AACACG, or $\phi \Rightarrow$ A ⇒ AAC ⇒ AACACG. A generation is called exhaustive if its extension steps are exhaustive (with an exception of the final step) and the number of steps are defined as the LZ complexity ($c(S)$) of S. Accordingly, the later generation process of AACACG is exhaustive, so c(AACACG) = 3. It is clear to say the LZ complexity of any sequence is unique and $c(S)$ is the minimum number of steps which S can generated from a null sequence using producible process.

2.2 Distance measure

The value of $c(ST) - c(S)$ measures the amount of information in T treating information in S as free. So the more similar the sequence S is to sequence T, the smaller $c(ST) - c(S)$ is, that is, $c(ST) - c(S)$ measures the dissimilarity between S and T [14]. But it is not a metric since all three axioms of distance do not hold. To assure

the identity and symmetry axioms, the distance metric between S and T is defined as follows

$$d(S,T) = \begin{cases} \frac{\max\{c(ST)-c(S),c(TS)-c(T)\}}{\max\{c(S),c(T)\}} & S \neq T \\ 0 & S = T. \end{cases} \quad (1)$$

The triangle inequality also holds (up to an additive error term). Known that LZ complexity is an explicitly computable implementation of the sequence entropy, the numerator can be considered as the mutual information in the information theory. The denominator is introduced to eliminate effects of different sequence lengths.

Given n primary DNA sequences under research, we first convert them into characteristic sequences according to the reduced rules. Then, pairwise distances among each kind of characteristic sequences are computed by the Formula (1). Finally, phylogenetic relationships are inferred from pairwise-distance matrices using some mature programs, e.g. Neighbor-Joining or UPGMA.

3 Phylogenetic trees for three data sets

3.1 Experiment no. 1: beta-globin genes of 11 species

In the first experiment, we choose the full *beta*-globin genes of 10 mammals and a nonmammal—gallus from EMBL database: human (*Homo sapiens*, HSHBB), chimpanzee (*Pan troglodytes*, PTGLB1), gorilla (*Gorilla gorilla*, GGBGLOBIN), lemur (*Eulemur macaco*, LMHBB), rat (*Rattus norvegicus*, RNGLB), mouse (*Mus musculus*, MMBGL1), goat (*Capra hircus*, CHHBBAA), bovine (*Bos taurus*, BTGL02), rabbit (*Oryctolagus cuniculus*, OCBGLO), opossum (*Didelphis virginiana*, DVHBBB), gallus (Gallus gallus, GGGL02). Gallus, the only nonmammalian species in this group, is chosen as the outgroup. By the neighbor-joining method in PHYLIP software package [15], phylogenetic trees are established and listed in Fig. 1. To test our method, alignment tree from primary sequences is also constructed (Fig. 1b). After comparing topologies of these phylogenies, we have the following observations:

1. For primary sequences, trees constructed by the LZ complexity-based method and multiple alignment share the same topology except for the position of rabbit (Fig. 1b). Rabbit is more related to Rodents from alignment method, but it is more related to Primates according to the LZ complexity method.
2. Phylogeny constructed from (R,Y)-sequences is exactly the same as that from primary sequences (the consensus phylogeny is shown as Fig. 1a).
3. Trees constructed from (S,W)- and (M,K)-sequences have some discrepancies with that from primary sequences, and two obvious are the overall structure and the position of lemur (Fig. 1c, d).

It is interesting to note that the relationship between rabbits and other mammals is a debatable subject to evolutionary biologists [16]. The main argument for the rabbit-rodent connection is their similar sets of gnawing teeth (though rabbits have an extra pair of incisors). But evidences from fossils and molecular data do not support

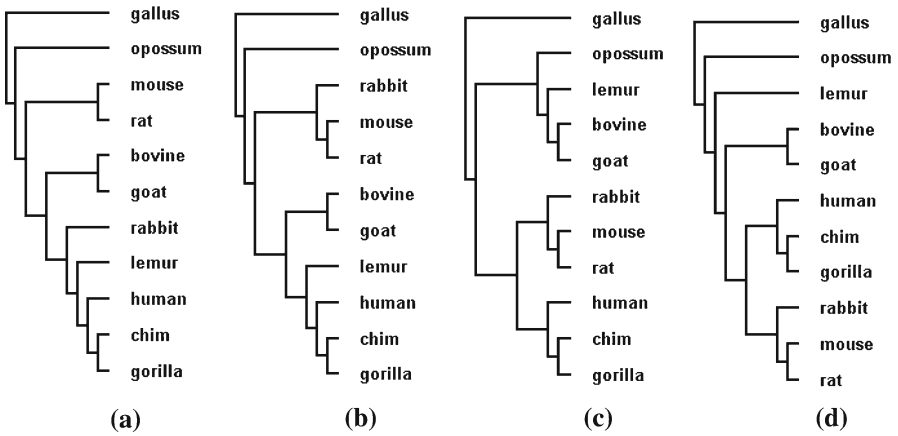

Fig. 1 **a** The consensus tree built from primary sequences and (R,Y)-sequences of 11 globin genes using the present distance metric. **b** Alignment tree built from primary sequences; Phylogenetic trees from (M,K)- and (S,W)-sequences using the LZ complexity method are shown in (**c**) and (**d**), respectively. In this work, branch lengths of all the trees are not scaled according to the distances, and only topologies of the trees are concerned

this viewpoint. So some systematists attributes the dental similarities to similar diets rather than common ancestry.

3.2 Experiment no. 2: Mt genomes of 20 mammals

The mammalian phylogenetic relationships at the molecular level have long been a controversial topic in molecular genetics [17]. The most debatable is the relationship among the three main groups of placental mammals, namely Primates, Ferungulates and Rodents [18–21]. Following Otu and Sayood [14], Cao et al. [20] and Li et al. [22], we choose the complete mtDNA sequences of 20 Eutherian mammals from GenBank as our second group of sequences: 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), horse (*Equus caballus*, X79547), 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), rat (*Rattus norvegicus*, X14848), mouse (*Mus musculus*, V00711), opossum (*Didelphis virginiana*, Z29573), wallaroo (*Macropus robustus*, Y10524) and platypus (*Ornithorhyncus anatinus*, X83427). Here, two marsupials, wallaroo and opossum, and one monotreme, platypus, are used as outgroup. The resulting trees are listed in Fig. 2.

In this experiment, three trees—alignment tree built from primary DNA sequences, LZ complexity-based trees from primary sequences and (R,Y)-sequences, have exactly the same topology. This consensus topology (Fig. 2a) coincides perfectly with the results presented by Cao et al. and Li et al., that is, three main groups of placental

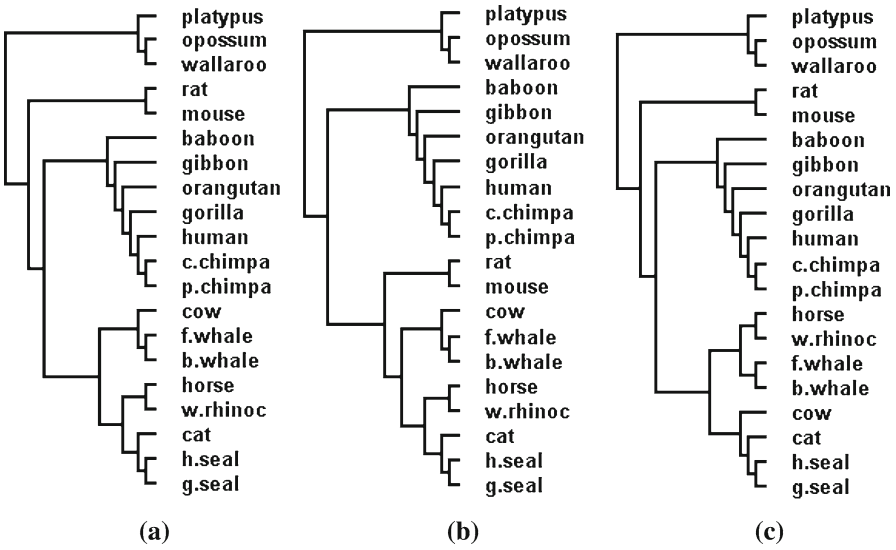

Fig. 2 Dendrograms built from **a** primary sequences and (R,Y)-sequences, **b** (S,W)- sequences, **c** (M,K)-sequences of 20 mammalian mt genomes. Alignment tree from whole mt genomes is in agreement with (**a**)

mammals cluster, and this topology confirms the outgroup status of Rodents relative to Ferungulates and Primates. Compared with primary and (R,Y)-sequences, (S,W)-sequences give a somewhat different phylogeny (Fig. 2b), which prefers Ferungulates and Rodents as the closest pair. But species cluster within each main clade. The tree constructed from (M,K)-sequences has some discrepancies in Ferungulates clade (Fig. 2c).

3.3 Experiment no. 3: whole genomes of 24 viruses

Sequences in above two experiments are all from vertebrates. In order to further assure our results, we turn to some viruses in the third experiment. Full genome sequences of 24 coronaviruses including SARS-CoVs and a torovirus are downloaded from GenBank (data are shown in Table 1). Coronaviruses are members of a family of enveloped viruses that replicate in the cytoplasm of animal host cell. According to the type of the host, coronaviruses isolated previously can be classified into three groups, groups I and II contain mammalian viruses, whereas group III contains only avian viruses. In order to infer the evolutionary relationships between SARS-CoV and other coronaviruses, Marra et al. [23] and Rota et al. [24] first chose data from above three groups and some SARS-CoV strains to construct phylogenetic tree. Their results indicated that SARS-CoVs are not closely related to any of the previously characterized coronaviruses and form a novel, fourth group of coronaviruses. Using similar data set, Zheng et al. [25] applied a geometric approach. They transformed each of the coronavirus genomes into "Z-Curve"—an equivalent graphical representation of DNA sequence, and then used the geometric center and three associated eigenvectors of the "Z-Curves" as

Table 1 Coronaviruses and a torovirus used to constructed phylogenetic tree

No.	Accession	Abbreviation	Genome	Group	Length (nt)
1	NC_002654	HCoV-229E	Human coronavirus 229E	I	27317
2	NC_002306	TGEV	Transmissible gastroenteritis virus	I	28586
3	NC_003436	PEDV	Porcine epidemic diarrhea virus	I	28033
4	U00735	BCoVM	Bovine coronavirus strain Mebuus	II	31032
5	AF391542	BCoVL	Bovine coronavirus isolate BCoV-LUN	II	31028
6	AF220295	BCoVQ	Bovin coronavirus strain Quebec	II	31100
7	NC_003045	BCoV	Bovine coronavirus	II	31028
8	AF208067	MHVM	Murine hepatituis virus strain ML-10	II	31233
9	AF201929	MHV2	Murine hepatitis virus stain 2	II	31276
10	AF208066	MHVP	Murine hepatitis virus stain Penn 97-1	II	31112
11	NC_001846	MHV	Murine hepatitis virus	II	31357
12	NC_001451	IBV	Avian infectiouis bronchitis virus	III	27608
13	AY278488	BJ01	SARS coronavirus BJ01	IV	29725
14	AY278741	Urbani	SARS coronavirus Urbani	IV	29727
15	AY278491	HKU-39849	SARS coronavirus HKU-39849	IV	29742
16	AY278554	CUHK-W1	SARS coronavirus CUHK-W1	IV	29736
17	AY282752	CUHK-Su10	SARS coronavirus CUHK-Su10	IV	29736
18	AY283794	SIN2500	SARS coronavirus SIN2500	IV	29711
19	AY283795	SIN2677	SARS coronavirus SIN2677	IV	29705
20	AY283796	SIN2679	SARS coronavirus SIN2679	IV	29711
21	AY283797	SIN2748	SARS coronavirus SIN2748	IV	29706
22	AY283798	SIN2774	SARS coronavirus SIN2774	IV	29711
23	AY291451	TW1	SARS coronavirus TW1	IV	29729
24	NC_004718	TOR2	SARS coronavirus	IV	29751
25	X52374	EToV	Equine torovirus	–	7920

descriptors for each genome. Their phylogeny is mainly consistent with the results of Marra et al. and Rota et al.

The closest known outgroup for coronaviruses are the toroviruses, which form a separate genus in the same virus family [26]. In this experiment, we choose the equine torovirus as our outgroup. The resulting phylogenetic topologies are shown in Fig. 3. As with the previous results, trees constructed from (R,Y)-sequences and primary sequences are mainly consistent except for some positions of SARS-CoVs. However, these discrepancies are acceptable since all SARS-CoV strains are almost completely identical in sequence ($\sim$99% aligned sequence identity). In such case, chance plays a more important role, and therefore it is not possible to get any meaningful phylogeny within the SARS-CoV group. Just like the above two experiments, trees from (M,Y)- and (S,W)-sequences have obvious discrepancies with Fig. 3a at different levels.

Notably, our phylogenies (Fig. 3a, b) are slightly different from the results given by Marra et al. and Rota et al., who prefer the outgroup status of SARS-CoVs relative

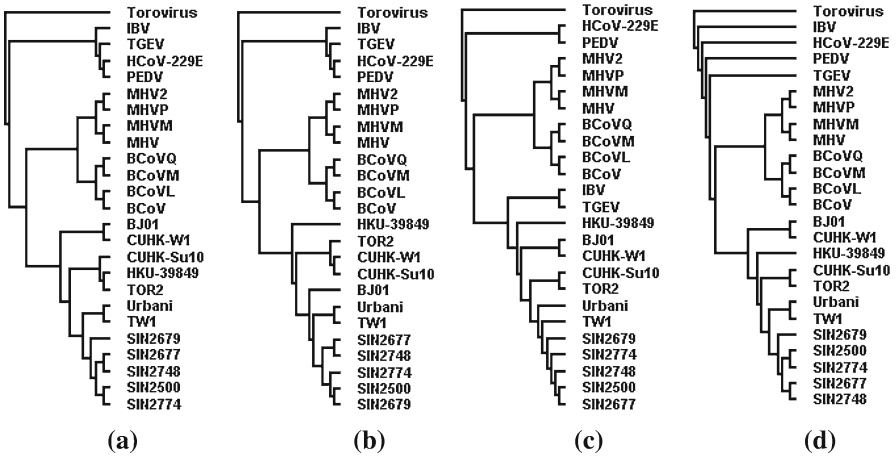

Fig. 3 Dendrograms obtained from **a** primary sequences, **b** (R,Y)-sequences, **c** (S,W)-sequences, **d** (M,K)-sequences of 25 virus genomes using proposed method, with torovirus chosen as the outgroup. We find that the topologies of (**a**, **b**) are in accordance except for some positions of SARS-CoVs, while (**c**, **d**) have discrepancies with (**a**) at different levels

to other coronaviruses. According to our phylogenies, SARS-CoVs are more closely related to group II of the coronavirus genus (bovine coronavirus and murine hepatitis virus). Similar conclusion was drawn by Liò and Goldman [27] using a fragment of the spike protein, and Snijder et al. [28] who aligned replicase ORF1b amino acid sequences of some related viruses. Here our method makes use of the whole genomes of these viruses, so it exhibits less bias.

4 An evolutionary model

Undoubtedly, this interesting phenomenon is attributed to the mechanisms of sequence evolution. It is widely detected that misincorporation errors during DNA duplication or repair are facilitated if a base is replaced by a similar one, and thus transitions ($T \leftrightarrow C, A \leftrightarrow G$) occur at higher frequencies than transversions ($T \leftrightarrow C, T \leftrightarrow G, C \leftrightarrow A, C \leftrightarrow T$): often twice as frequently, but the ratio can be much higher [29–32]. Therefore, the purine/pyrimidine distribution of nucleotides is relatively conservative through evolution compared with other distributions. In other words, the (R,Y)-sequence is more conservative than the other two sequences, and phylogenies from (R,Y)-sequences should be more reliable.

In this section, we shall propose a simple model for sequence evolution to interpret the above phenomenon. The model employs only point substitutions (transition and transversion), and transition/transversion rate bias is constant to 2. Explicitly, in our model, the following three steps are implemented:

1. A random sequence S of length L is generated.
2. Two sequences S_1 and S_2 are evolved from S by point mutations, during which every nucleotide has the same mutation rate μ, and transition rate is twice as large

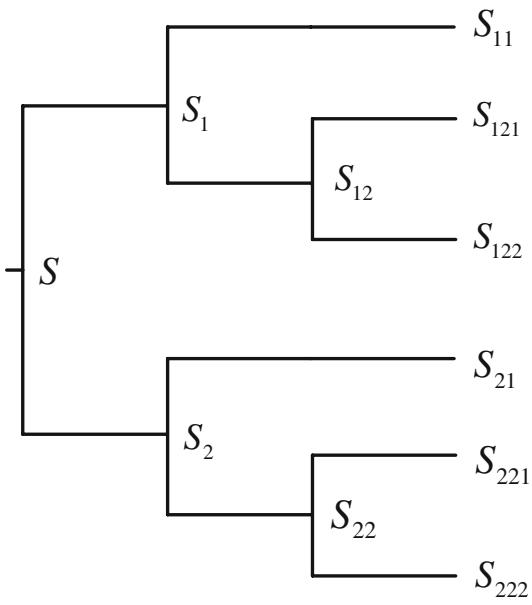

Fig. 4 Model tree in our simulations. The tree is generated using only point mutations from a random sequence S

as transversion. Accordingly, we evolve S_1 to S_{11} and S_{12}, S_2 to S_{21} and S_{22}. Then S_{12} is evolved to S_{121} and S_{122}, S_{22} to S_{221} and S_{222}. Evolutionary relationships among these six OTUs (S_{11}, S_{121}, S_{122}, S_{21}, S_{221} and S_{222}) are shown in Fig. 4.

3. Using our reduced rules, characteristic sequences for each primary sequence is got. Then phylogenetic trees for each kind of characteristic sequences are constructed to check their consistency with the real one (Fig. 4).

Sequences in the first dataset (11 *beta*-globin genes) have the approximate length of 1,500. We therefore set $L = 1,500$ in our first simulation. For the test of the bootstrap error rate, the above three steps are performed for 1,000 times.

Probabilities to get the correct topology (accuracies) at different mutation rates are shown in Fig. 5. As can be seen in this figure, accuracies for all four sequences increase first to a maximum value, then decrease, with peak values $\geq$99%. Remember that an individual reduction will lose some information, so compared to characteristic sequences, primary sequences give relatively higher values. If we define the *Reliable Interval* of one kind of sequences to be the range of mutation rate at which the probability to get the correct topology is $\geq$95%. Then, in this simulation, the reliable interval of primary sequences is (1/100, 1/8), while (R,Y)-, (M,K)- and (S,W)-sequences have reliable intervals of (1/50, 1/9), (1/50, 1/15) and (1/50, 1/15), respectively. From Fig. 5 we can see that (M,K)- and (S,W)-sequences perform similarly in getting the correct tree, this is not surprising since the (M,K)- and (S,W)-sequences are symmetry in our model.

Sequence lengths in the second and third datasets are ranged from 10,000 to 30,000. For simplicity, we set $L = 10,000$ in the second simulation. The above three steps are

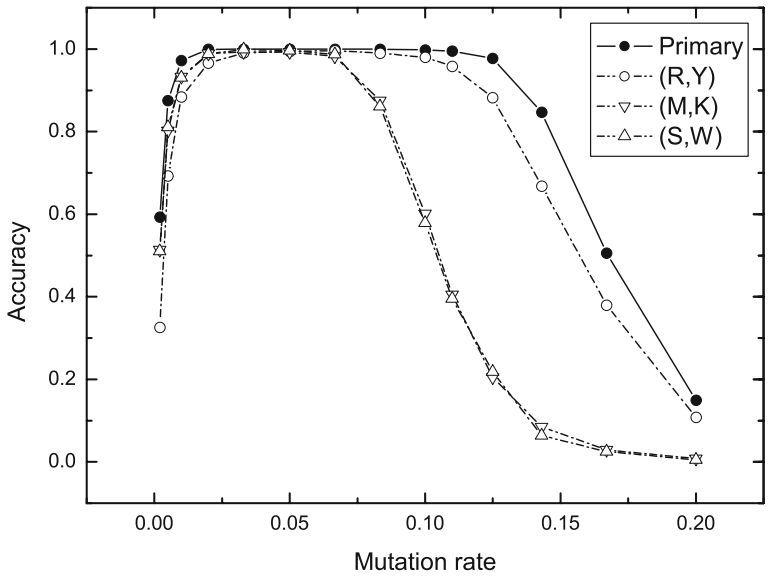

Fig. 5 Accuracies to give the correct topology at different mutation rates. In this simulation, sequence length $L = 1500$, 1000 experiments are performed

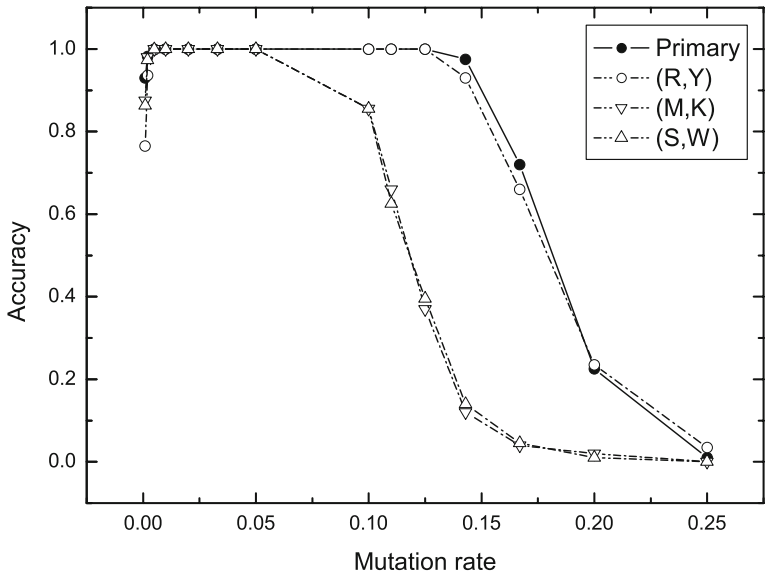

Fig. 6 Accuracies to give the correct topology at different mutation rates. In this simulation, sequence length $L = 10,000$, 1,000 experiments are performed

performed for 1,000 times at each mutation rate, and the final accuracies are shown in Fig. 6. Similar features are seen from this simulation. The only difference is some slight increase of accuracies at some mutation rates. Our explanation is the finite size effect,

which plays an important role at the short sequence length (1,500) and diminishes gradually as sequence length increase to 10,000.

It is worth noting that accuracies of (R,Y)-sequence is not always larger than the other two characteristic sequences—when mutation rates are very small ($\leq$1/50 in the first simulation and $\leq$1/200 in the second simulation), both (M,Y)- and (S,W)-sequences get better results. This can be interpreted from an informational perspective. Since point mutations are less likely to alter the purine/pyrimidine distribution, (R,Y)-sequence is relatively stable, while (M,Y)- and (S,W)-sequences are more sensitive to mutations. Therefore the sensitive one is expected to get a more reliable result when mutation rates are very small. While in the case of high mutation rates, the sensitive one may be disturbed by random mutation and the stable one will dominant.

5 Conclusion

What kind of and how much biological information an individual characteristic sequence carries? Our experiment and simulation present an answer from the phylogenetic perspective: (R,Y)-sequence of a primary sequence plays a pivotal role in determining its phylogenetic position. When sequence distance lies in a certain range, we can reconstruct a reliable phylogenetic tree using only (R,Y)-sequences.

It is exciting that only binary sequences reduced from primary DNA sequences carry enough information to infer a phylogeny. Compared to the primary sequence, these binary sequences have much higher compression ratio. This will, on one hand, reduce the storage space and execution time (Table 2), and on the other hand, facilitate the use of some signal processing techniques in biological data analysis. However, drawing the above conclusion from only one method (LZ complexity) is not so convincing. We have also tried some other sequence comparison methods, e.g., k-mers composition and traditional alignment. But k-mers based methods got poor results for all datasets even using primary sequences (representing sequences as vectors of k-mer frequencies may lose too much information), and alignments also fail to compare binary sequences for the lack of reliable score matrix.

In our future experiments, some other sequence analyses (e.g., gene identification and structure prediction) will be tried on these three kinds of binary sequences, to check whether there exists a dominant one. Known that distribution of weak/strong H-bonds affects the structure of this molecular seriously, we conjecture that the (S,W)-sequences play a key role in the configuration of the molecular.

Table 2 Time consumption to compute distance matrices from primary and three characteristic sequences

	Primary	(R,Y)	(M,K)	(S,W)
Beta-globin genes	3.593 s	2.703 s	2.913 s	2.906 s
Mt genomes	23 m 4 s	17 m 2 s	18 m 44 s	17 m 25 s
Virus genomes	1 h 29 m 31 s	1 h 8 m 33 s	1 h 8 m 49 s	1 h 10 m 0 s

Experiments are performed on a PC with Pentium IV CPU (1.8 GHZ) and 512 MB RAM

Acknowledgements This work was supported in part by Leading Academic Discipline Project of Shanghai Normal University (No. DZL803) and Shanghai Leading Academic Discipline Project (No. S30405).

References

1. C. Burge, A.M. Campbell, S. Karlin, Proc. Natl. Acad. Sci. **89**, 1358 (1992)
2. S. Vinga, J. Almeida, Bioinformatics **19**, 512 (2003)
3. H.J. Jeffrey, Nucleic Acids Res. **18**, 2163 (1990)
4. G.D. Shuler, S.F. Altschul, D.J. Lipman, Proteins: Struct. Funct. Genet. **9**, 180 (1991)
5. C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, J.C. Wootton, Science **262**, 208 (1993)
6. P.A. He, J. Wang, J. Chem. Inf. Comput. Sci. **42**, 1080 (2002)
7. M. Randić, M. Vračko, J. Zupan, M. Novič, Chem. Phys. Lett. **373**, 558 (2003)
8. Y.H. Yao, X.Y. Nan, T.M. Wang, J. Mol. Struct. Theochem. **764**, 101 (2006)
9. N. Liu, T.M. Wang, J. Mol. Model. **12**, 897 (2006)
10. P.A. He, C. Li, J. Wang, Internet Electron. J. Mol. Des. **4**, 613 (2005)
11. A. Lempel, J. Ziv, IEEE Trans. Inf. Theory **22**, 75 (1976)
12. J. Ziv, A. Lempel, IEEE Trans. Inf. Theory **23**, 337 (1977)
13. J. Ziv, A. Lempel, IEEE Trans. Inf. Theory **24**, 530 (1978)
14. H.H. Otu, K. Sayood, Bioinformatics **19**, 2122 (2003)
15. J. Felsenstein, *PHYLIP (Phylogenetic Inference Package) ver. 3.57*. (Department of Genetics, University of Washington, Seattle, WA, 1995)
16. D. Graur, L. Duret, M. Gouy, Nature **379**, 333 (1996)
17. A. Reyes, C. Gissi, G. Pesole, F.M. Catzeflis, C. Saccone, Mol. Biol. Evol. **17**, 979 (2000)
18. A. Janke, G. Feldmaier-Fuchs, W. K. Thomas, A. von Haeseler, S. Pääbo, Genetics **137**, 243 (1994)
19. A. Janke, X. Xu, U. Arnason, Proc. Natl. Acad. Sci. **94**, 1276 (1997)
20. Y. Cao, J. Adachi, A. Janke, S. Pääbo, M. Hasegawa, J. Mol. Evol. **39**, 519 (1994)
21. D. Penny, M. Hasegawa, Nature **387**, 549 (1997)
22. M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, H.Y. Zhang, Bioinformatics **17**, 149 (2001)
23. M.A. Marra, S.J. Jones, C.R. Astell et al., Science **300**, 1399 (2003)
24. P.A. Rota, M.S. Oberste, S.S. Monroe et al., Science **300**, 1394 (2003)
25. W.C. Zheng, L.L. Chen, H.Y. Ou, F. Gao, C.T. Zhang, Mol. Phylogenet. Evol. **36**, 224 (2005)
26. E.J. Snijder, M.C. Horzinek, J. Gen. Virol. **74**, 2305 (1993)
27. P. Liò, N.Goldman, Trends Microbiol. **12**, 106 (2004)
28. E.J. Snijder, P.J. Bredenbeek, J.C. Dobbe, V. Thiel, J. Ziebuhr, L.L. Poon, Y. Guan, M. Rozanov, W.J. Spaan, A.E. Gorbalenya, J. Mol. Biol. **331**, 991 (2003)
29. T. Gojobori, W.H. Li, J. Mol. Evol. **18**, 360 (1982)
30. J. Wakeley, Mol. Biol. Evol. **11**, 436 (1994)
31. J. Wakeley, Tree **11**, 158 (1996)
32. Z. Yang, A.D. Yoder, J. Mol. Evol. **48**, 274 (1999)