BIOINFORMATICS

Vol. 19 no. 16 2003, pages 2122–2130
DOI: 10.1093/bioinformatics/btg295

A new sequence distance measure for phylogenetic tree construction

Hasan H. Otu[1,2,*] and Khalid Sayood[1]

[1]Department of Electrical Engineering, University of Nebraska-Lincoln, 209N WSEC, Lincoln, NE 68503, USA and [2]New England Baptist Bone and Joint Institute, Beth Israel Deaconess Medical Center Genomics Center, Harvard Medical School, Boston, MA 02215, USA

Received on November 18, 2002; revised on March 5, 2003; accepted on April 17, 2003

ABSTRACT

Motivation: Most existing approaches for phylogenetic inference use multiple alignment of sequences and assume some sort of an evolutionary model. The multiple alignment strategy does not work for all types of data, e.g. whole genome phylogeny, and the evolutionary models may not always be correct. We propose a new sequence distance measure based on the relative information between the sequences using Lempel–Ziv complexity. The distance matrix thus obtained can be used to construct phylogenetic trees.

Results: The proposed approach does not require sequence alignment and is totally automatic. The algorithm has successfully constructed consistent phylogenies for real and simulated data sets.

Availability: Available on request from the authors.

Contact: hotu@bidmc.harvard.edu

INTRODUCTION

Phylogenetic analysis using biological sequences can be divided into two groups. The algorithms in the first group calculate a matrix representing the distance between each pair of sequences and then transform this matrix into a tree. In the second type of approach, instead of building a tree, the tree that can best explain the observed sequences under the evolutionary assumption is found by evaluating the fitness of different topologies.

Some of the approaches in the first category utilize various distance measures (Jukes and Cantor, 1969; Kimura, 1980; Barry and Hartigan, 1987; Kishino and Hasegawa, 1989; Lake, 1994) which use different models of nucleotide substitution or amino acid replacement. The second category can further be divided into two groups based on the optimality criterion used in tree evaluation: parsimony (Camin and Sokal, 1965; Eck and Dayhoff, 1966; Cavalli-Sforza and Edwards, 1967; Fitch, 1971) and maximum likelihood methods (Felsenstein, 1973, 1981; Felsenstein and Churchill, 1996). For a detailed comparison of these methods see Yang (1996) and Durbin *et al.* (1999).

*To whom correspondence should be addressed.

All of these methods require a multiple alignment of the sequences and assume some sort of an evolutionary model. In addition to problems in multiple alignment (computational complexity and the inherent ambiguity of the alignment cost criteria) and evolutionary models (they are usually controversial), these methods become insufficient for phylogenies using complete genomes. Multiple alignment becomes misleading due to gene rearrangements, inversion, transposition and translocation at the substring level, unequal length of sequences, etc. and statistical evolutionary models are yet to be suggested for complete genomes. On the other hand, whole genome-based phylogenetic analysis are appealing because single gene sequences generally do not possess enough information to construct an evolutionary history of organisms. Factors such as different rates of evolution and horizontal gene transfer make phylogenetic analysis of species using single gene sequences difficult.

To overcome these problems, Sankoff *et al.* (1992) defined an evolutionary edit distance as the number of inversions, transpositions and deletions or insertions required to change the gene order of one genome into another. Similar distance measures using rearrangement, recombination, breakpoint, comparative mapping and gene order have been extensively studied for applications to genome-based phylogeny (Hannenhalli and Pevzner, 1995; Kececioglu and Sankoff, 1995; Kececioglu and Ravi, 1995, 1998; Boore and Brown, 1998; Sankoff and Blanchette, 1998; Sankoff, 1999a; http://www.agbiotechnet.com/proceedings/jaylush.asp; Sankoff, 1999b; Hannenhalli and Pevzner, 1999; Berman *et al.*, 2001). However, these approaches 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 (as in mammalian mtDNA where genes are in the same order).

Gene content was proposed by Snel *et al.* (1999) as a distance measure in genome phylogeny where 'the similarity between two species is defined as the number of genes they have in common divided by their total number of genes'. The general idea is further extended to identify evolutionary history and protein functionality (Snel *et al.*,

2000, 2002; Huynen *et al*., 2000). A similar approach is taken by Fitz-Gibbon and House (1999). Lin and Gerstein (2000) constructed phylogenetic trees based on the occurrence of particular molecular features: presence or absence of either folds or orthologs throughout the whole genome. Takaia *et al*. (1999) used whole proteome comparisons in deriving genome phylogeny, taking into account the overall similarity and the predicted gene product content of each organism. However, such methods fail to work when the gene content of the organisms are very similar (again as is the case with mammalian mtDNA where the genomes contain exactly the same genes).

In the early 1990s, various data compression approaches were applied to the analysis of genetic sequences (Milosavljevic, 1993; Grumbach and Tahi, 1993, 1994; Rivals *et al*., 1994; Farach *et al*., 1995). Data compression algorithms function by identifying the regularities in the given sequence, and in case of DNA sequences, these regularities would have biological implications.

Grumbach and Tahi (1993, 1994) coded exact repeats and palindromes in DNA along the lines of Lempel–Ziv (LZ) compression scheme (Ziv and Lempel, 1977), and used an arithmetic coder of order 2 when such structures are lacking. Rivals *et al*. (1994, 1996) compressed the repeats which introduced a significant compression gain and introduced a second compressor which made use of approximate tandem repeats. Rivals *et al*. (1997) also introduced a compression algorithm which locates and utilizes approximate tandem repeats of short motifs. Some of the later approaches include (Loewenstern and Yianilos, 1999; Lanctot *et al*., 2000; Apostolico and Lonardi, 2000). Grumbach and Tahi (1994) noted that the compression rate obtained by compressing sequence S using sequence Q would hint at some sort of a distance between the two sequences. Although the proposed distance was not mathematically valid and had some other problems, it applied data compression to phylogeny construction.

Varre *et al*. (1999) defined a transformation distance where sequence S is built from sequence Q by segment-copy, -reverse-copy and -insertion. The total distance is the Minimum Description Length among all possible operations that convert S into Q. This distance, as the one provided by Grumbach and Tahi (1994), is asymmetric. Chen *et al*. (2000) described a compression algorithm (GenCompress) based on approximate repeats in DNA sequences. The program is then used to approximate the distance proposed therein and the distance proposed by Li *et al*. (2001). For a detailed analysis of information distance in statistical and algorithmic settings, see Ziv and Merhav (1993) and Bennett *et al*. (1998).

The distance proposed by Chen *et al*. (2000) and Li *et al*. (2001) is $1 - [K(S) - K(S|Q)]/K(SQ)$, where $K(S)$ is the Kolmogorov complexity of S, $K(S|Q)$ is the conditional Kolmogorov complexity of S given Q and $K(SQ)$ is the Kolmogorov complexity of the sequence S concatenated with Q. $K(S|Q)$ is the shortest program that outputs S when the input is Q on a universal computer and $K(S)$ is $K(S|\epsilon)$, where ϵ is the empty string. Kolmogorov complexity is an algorithmic measure of information (Li and Vitanyi, 1997) but it is a theoretical limit and generally can only be approximated. In calculating the aforementioned distance, $K(S|Q)$ is approximated by the length of the compressed result of S (using the program GenCompress) given Q.

Benedetto *et al*. (2002) used a similar idea where relative complexity between sequences S and Q is approximated as it is done by Chen *et al*. (2000), this time using *gzip*. However, both *gzip* and GenCompress are complicated programs, composed of multiple complex steps (algorithms to reduce search space, find exact/approximate matches, perform entropy coding, etc.), which would affect the final result on the complexity estimates in an ambiguous way. Therefore the properties of the distance measures based on Kolmogorov complexity (implicitly or explicitly) would not necessarily hold for these approximations depending on the performance of the compression algorithms on certain sequences.

Motivated by the work of Chen *et al*. (2000); Li *et al*. (2001) and Benedetto *et al*. (2002), in this paper, we propose a distance measure between finite sequences based on the LZ complexity (Lempel and Ziv, 1976), which inspired the well-known universal compression schemes (Ziv and Lempel, 1977, 1978) and their numerous variations. LZ complexity of a finite sequence S is related to the number of steps required by a production process that builds S. In the next section, we will give some basic definitions and properties regarding LZ complexity and introduce the proposed distance measure. The following section provides results on phylogenetic tree construction and the last section concludes the paper with some remarks. Finally, we provide the mathematical properties of the proposed distance in the Appendix.

METHODS AND ALGORITHMS

LZ complexity

Let S, Q and R be sequences defined over an alphabet $\mathcal{A}$, $l(S)$ be the length of S, $S(i)$ denote the ith element of S and $S(i, j)$ define the substring of S composed of the elements of S between positions i and j (inclusive). An extension $R = SQ$ of S is *reproducible* from S (denoted $S \rightarrow R$) if there exists an integer $p \leq l(S)$ such that $Q(k) = R(p + k - 1)$ for $k = 1, \ldots, l(Q)$. For example $AACGT \rightarrow AACGTCGTCG$ with $p = 3$ and $AACGT \rightarrow AACGTAC$ with $p = 2$.

Another way of looking at this is to say that R can be obtained from S by copying elements from the pth location in S to the end of S. As each copy extends the length of the new sequence beyond $l(S)$, the number of elements copied can be greater than $l(S) - p + 1$. Thus, this is a simple copying procedure of S starting from position p, which can carry over to the added part, Q.

A sequence S is *producible* from its prefix $S(1, j)$ (denoted $S(1, j) \Rightarrow S$), if $S(1, j) \rightarrow S(1, l(S) - 1)$. For example $AACGT \Rightarrow AACGTAC$ and $AACGT \Rightarrow AACGTACC$

both with pointers $p = 2$. Note that production allows for an extra 'different' symbol at the end of the copying process which is not permitted in reproduction. Therefore, an extension which is reproducible is always producible but the reverse may not always be true.

Any sequence S can be built using a *production process* where at its ith step $S(1, h_{i-1}) \Rightarrow S(1, h_i)$ [note that $\epsilon = S(1, 0) \Rightarrow S(1, 1)$]. An m-step production process of S results in a parsing of S in which $H(S) = S(1, h_1) \cdot S(h_1 + 1, h_2), \ldots, S(h_{m-1} + 1, h_m)$ is called the *history* of S and $H_i(S) = S(h_{i-1}+1, h_i)$ is called the ith component of $H(S)$. For 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.

If $S(1, h_i)$ is not reproducible from $S(1, h_{i-1})$ [denoted $S(1, h_{i-1}) \nrightarrow S(1, h_i)$], then $H_i(S)$ is called *exhaustive*. In other words, for $H_i(S)$ to be exhaustive the ith step in the production process must be a production only, meaning that the copying process cannot be continued and the component should be halted with a single letter innovation. A history is called exhaustive if each of its components (except maybe the last one) is exhaustive. For example the third history given in the preceding paragraph is an exhaustive history of $S = AACGTACC$. Moreover, every sequence S has a unique exhaustive history (Lempel and Ziv, 1976).

Let $c_{\mathrm{H}}(S)$ be the number of components in a history of S. Then the LZ complexity of S is $c(S) = \min\{c_{\mathrm{H}}(S)\}$ (Lempel and Ziv, 1976) over all histories of S. It can be shown that $c(S) = c_{\mathrm{E}}(S)$ where $c_{\mathrm{E}}(S)$ is the number of components in the exhaustive history of S (Lempel and Ziv, 1976). This is quite intuitive as an exhaustive component is the longest possible one at a given step of a production process.

Proposed distance

Given two sequences Q and S, consider the sequence SQ, and its exhaustive history. By definition, the number of components needed to build Q when appended to S is $c(SQ) - c(S)$. This number will be less than or equal to $c(Q)$ because at any given step of the production process of Q (in building the sequence SQ) we will be using a larger search space due to the existence of S. Therefore the copying process can only be longer which in turn would reduce the number of exhaustive components. This can also be seen from the subadditivity of the LZ complexity (Lempel and Ziv, 1976): $c(SQ) \leq c(S) + c(Q)$. How much $c(SQ) - c(S)$ is less than $c(Q)$ will depend on the degree of similarity between S and Q.

For example, let $S = AACGTACCATTG$, $R = CTAGGGACTTAT$ and $Q = ACGGTCACCAA$. The exhaustive histories of these sequences would be:

$$H_E(S) = A \cdot AC \cdot G \cdot T \cdot ACC \cdot AT \cdot TG$$
$$H_E(R) = C \cdot T \cdot A \cdot G \cdot GGA \cdot CTT \cdot AT$$
$$H_E(Q) = A \cdot C \cdot G \cdot GT \cdot CA \cdot CC \cdot AA$$

yielding $c(S) = c(R) = c(Q) = 7$. The exhaustive histories of the sequences SQ, and RQ would be:

$$A \cdot AC \cdot G \cdot T \cdot ACC \cdot AT \cdot TG \cdot ACGG \cdot TC \cdot ACCAA$$
$$C \cdot T \cdot A \cdot G \cdot GGA \cdot CTT \cdot AT \cdot ACG \cdot GT \cdot CA \cdot CC \cdot AA$$

respectively. Note that it took three steps to build Q in the production process of SQ. On the other hand, we used five steps to generate Q in the production process of RQ. The reason it took more steps in the second case is because Q is 'closer' to S than R. In this example we can observe this by looking at the patterns ACG and ACC which Q and S share. We can formulate the number of steps it takes to generate a sequence Q from a sequence S by $c(SQ) - c(S)$. Thus, if S is closer to Q than R then we would expect $c(SQ) - c(S)$ to be smaller than $c(RQ) - c(R)$ as is the case in the above example. Based on this idea of closeness we define four distance measures.

Distance measure 1 Given two sequences S and Q, define the function $d(S, Q)$ as

$$d(S, Q) = \max\{c(SQ) - c(S), c(QS) - c(Q)\}$$

In order to eliminate the effect of the length on the distance measure, a more satisfying function would be the normalized form of $d(\cdot, \cdot)$:

Distance measure 2 Given two sequences S and Q, define the function $d^*(S, Q)$ as

$$d^*(S, Q) = \frac{\max\{c(SQ) - c(S), c(QS) - c(Q)\}}{\max\{c(S), c(Q)\}}$$

Another distance measure that would naturally follow from the idea of building sequence Q using S is the 'sum distance'. We use this term in the sense that it accounts for the total number of steps it takes to build Q from S and vice versa.

Distance Measure 3 Given two sequences S and Q, define the function $d_1(S, Q)$ as

$$d_1(S, Q) = c(SQ) - c(S) + c(QS) - c(Q)$$

Similarly, the normalized version of $d_1(\cdot, \cdot)$ can be defined as follows.

Distance Measure 4 Given two sequences S and Q, define the function $d_1^*(S, Q)$ as*

$$d_1^*(S, Q) = \frac{c(SQ) - c(S) + c(QS) - c(Q)}{c(SQ)}$$

* An alternative definition would be

$$d_1^{**}(S, Q) = \frac{c(SQ) - c(S) + c(QS) - c(Q)}{\frac{1}{2}[c(SQ) + c(QS)]}.$$

A distance metric, $D(\cdot,\cdot)$, should satisfy the following 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. $D(S,Q) \leq D(S,T) + D(T,Q)$ (triangle inequality).

In order for a metric to be a valid measure of evolutionary change it should also satisfy the following condition:

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

In the Appendix we prove that all four measures defined above satisfy the first three conditions and are, therefore, valid distance metrics. We test the fourth condition by comparing a distance matrix created by the proposed metrics with one reconstructed from the branch lengths of the resulting tree.

In the next section we show that the proposed distance measures, which are based on the relative complexity between sequences imply the evolutionary distance between organisms. The distance between sequences S and Q was obtained using the exhaustive histories of the sequences S, Q, SQ and QS. These exhaustive histories were obtained by parsing the sequences using the production rules described earlier and in (Lempel and Ziv, 1976). The number of components in the exhaustive histories, $c(S), c(Q), c(SQ)$ and $c(QS)$ were then used as described above to compute the various distance measures.

RESULTS AND DISCUSSION

Phylogenetic analysis based on DNA sequences has been intimately connected with multiple alignment. Hence the validity of a new approach is generally examined based on how well the implicit assumptions used for scoring the multiple alignment agree with particular evolutionary theories. As the proposed approach does not depend on multiple alignments we test the validity of the approach in two ways: we use simulated data to show that the proposed distance measures can reasonably be represented by a tree. We also show the superiority of the proposed method on existing techniques using this simulated data. Secondly, we look at how well the results generated by the proposed method agree with existing phylogenies. The trees are generated using the neighbor joining (NJ) program (Saitou and Nei, 1987) in the PHYLIP package (Felsenstein, 1989). The multiple alignments required by parsimony and maximum likelihood methods are calculated using CLUSTAL W (Thompson *et al.*, 1994).

For the simulated data, we started with a 1000 bp sequence and evolved it into two sequences A'' and B'' using point mutations (insertions, deletions, substitutions) and segment-based modifications (inversions, transpositions, translocations, etc.). We then similarly evolved A'' into $A1$ and $A2$ and B'' into $B1$ and $B2$. Point mutations were introduced

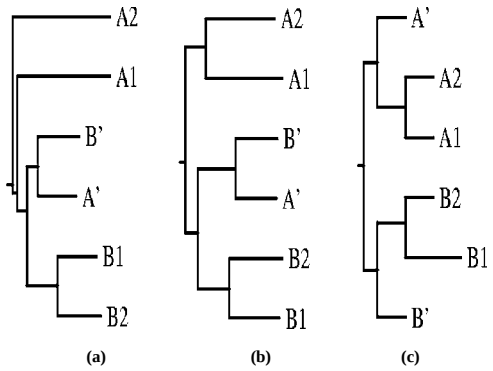

Fig. 1. Phylogenetic trees obtained from the simulated sequences $A', B', A1, A2, B1$ and $B2$ using (**a**) Maximum Likelihood, (**b**) parsimony and (**c**) Proposed methods.

into about 10% of the sequences. Another 10% of the final sequences were a result of sequence rearrangements. These included inversions and translocations. In order to provide length difference and to preserve resemblance to the ancestor sequences, we evolved A'' into A' and B'' into B' using point mutations only. We used the sequences $A', B', A1, A2, B1$ and $B2$ to build phylogenetic trees both using existing methods (maximum likelihood and parsimony) and the proposed method. The results are shown in Figure 1.

The trees obtained by all five of the proposed distance measures resulted in identical topologies. In Figure 1, we show the consensus tree obtained by those five trees along with the trees obtained by maximum likelihood and parsimony methods. The results show that the true evolutionary topology is achieved by the proposed method only. Both the maximum likelihood and the parsimony trees fail to reflect the relation between A' and $A1, A2$. In addition, the maximum likelihood tree fails to group $A1$ and $A2$ together.

We also compared the proposed distance matrix used to build the NJ tree to that reconstructed from the branch lengths of the tree. The purpose was to test additivity of the proposed distance measures. Let M be the distance matrix obtained by the proposed distance measure, T be the tree built using M, and R be the distance matrix reconstructed from the branch lengths of T. In Table 1 we present M, R and $|M-R|$ using d^* for the simulated data set. We omit the corresponding results for the remaining four measures as they are almost identical to the ones presented here. The results in Table 1 show that M and R are very similar to each other, validating the properness of representing the proposed measures with a tree. The maximum percent difference between the corresponding elements of the matrices M and R is 11, with an average percent difference of 2.5. In the remaining part of this section we show that the proposed method agrees with existing phylogenies based on both whole genome and individual gene sequences.

The phylogeny of eutherian orders has been unresolved due to conflicting results obtained from comparison of

Table 1. Fitness of the NJ tree to the distance matrix

Seq.	A2	B1	B2	A′	B′
*A*1	0.7348	0.8093	0.8104	0.7535	0.8056
*A*2		0.8000	0.8046	0.7767	0.8139
*B*1			0.7488	0.8000	0.7674
*B*2				0.8000	0.7952
A′					0.7745
*A*1	0.7348	0.8015	0.8161	0.7630	0.7180
*A*2		0.8056	0.8203	0.7672	0.7221
*B*1			0.7488	0.7878	0.7740
*B*2				0.8024	0.7886
A′					0.7042
*A*1	0.0000	0.0077	0.0057	0.0095	0.0876
*A*2		0.0056	0.0156	0.0095	0.0918
*B*1			0.0000	0.0121	0.0065
*B*2				0.0024	0.0065
A′					0.0702

Distance matrix used to construct the NJ tree, T; distance matrix reconstructed from the branch lengths of T and the difference of the two matrices are shown respectively.

whole mtDNA sequences and individual proteins encoded by mtDNA (see Cao *et al*., 1998 and references therein). Studies using the whole mtDNA sequences suggest the outgroup status of rodents relative to ferungulates and primates [Rodents (Ferungulates, Primates)] while phylogenies using individual proteins confirm the grouping of rodents with primates. There have even been conflicting topologies resulting from the use of different proteins in constructing the evolutionary history.

We chose our first group of sequences from this controversial data set using the following mtDNA sequences from GenBank: 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). Note that we have kept rodent species to murids only and marsupials and monotremes are being used as outgroup.

We applied the proposed distance measures to the complete mitochondrial genomes listed above. All five metrics (d, d^*, d_1, d_1^* and d_1^{**}) resulted in identical trees. In Figure 2, we show the consensus of these five trees. The tree is in complete agreement with Cao *et al*. (1998) confirming the outgroup status of rodents relative to ferungulates and primates.

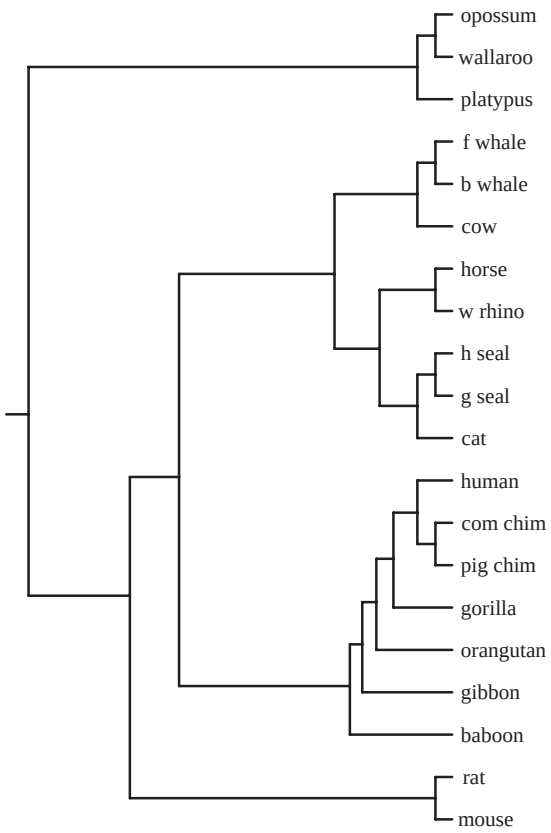

Fig. 2. Topology for eutherians using whole mtDNA where wallaroo, opossum and platypus are used as outgroup.

The second data set is an extension of the first one obtained by the addition of non-murid rodents (squirrel, dormouse and guinea pig) and more ferungulate sequences. The GenBank accession codes for these additional mtDNA sequences are as follows: squirrel (*Sciurus vulgaris*, AJ238588), fat dormouse (*Glis glis*, AJ001562), guinea pig (*Cavia porcellus*, AJ222767), donkey (*Equus asinus*, X97337), Indian rhinoceros (*Rhinoceros unicornis*, X97336), dog (*Canis familiaris*, U96639), sheep (*Ovis aries*, AF010406), pig (*Sus scrofa*, AJ002189), and hippopotamus (*Hippopotamus amphibius*, AJ010957). This more controversial data set deals with the relative positions of two rodent clades, murids and non-murids (whether there is rodent monophyly, paraphyly or polyphyly) and the phylogenetic position of guinea pigs (Cao *et al*., 1997; Reyes *et al*., 2000).

The resulting trees from the five distance metrics were in agreement for the most part. All of the metrics confirmed rodent paraphyly (except for d which suggested rodent monophyly) and guinea pig was not grouped with either rodent clade in each case (except for d^* which suggested grouping of guinea pig with nonmurids). We present the consensus tree of the five trees obtained using the proposed metrics in Figure 3. The consensus phylogeny is in agreement with

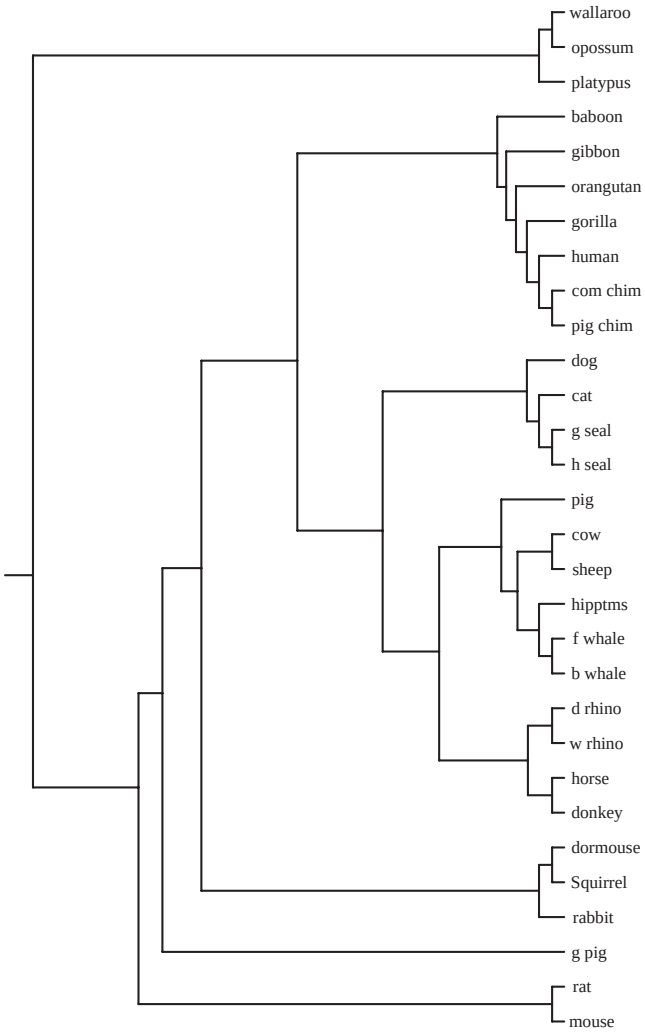

Fig. 3. The consensus tree for the proposed distance metrics using complete mtDNA.

Reyes *et al*. (2000) except for the position of guinea pig which remains an open question (Cao *et al*., 1997). Figure 3 groups squirrel with dormouse, which has shown to be based on strong molecular, palaeontological and morphological evidence [see Reyes *et al*. (2000) and references therein]. On the other hand, nonmurid rodents are placed at the base of primates and ferungulates with murids being an early branch of the tree (suggesting rodent paraphyly), which is presented as the most likely hypotheses by Reyes *et al*. (2000).

The results presented in Figures 2 and 3 are in accord with (Li *et al*., 2001), which also applied an information theoretic distance measure to these data sets. However, mammalian phylogeny still remains to be a controversial topic. Two recent studies suggest a monophyletic clade of rodents and primates (Madsen *et al*., 2001; Murphy *et al*., 2001). As noted earlier, conflicting results have been reported regarding the phylogeny of eutherian orders based on whole genome sequences or individual genes. The sequences used in Madsen *et al*. (2001) and Murphy *et al*. (2001) use individual genes to build the trees. We also note that the data sets used in these two studies differ from each other and the data sets used in this paper, which could result in varying topologies.

Note that in both of these analyses we have used the whole mitochondrial genomes of the species. The results are not based on the phylogenies inferred using coding regions or individual proteins. Instead we use the complete sequences as opposed to partial genome data. The phylogenies inferred using the proposed distances confirm that our method can successfully construct evolutionary histories using whole genome sequences.

We also successfully applied the proposed method to constructing phylogenies using individual genes (data not shown). In most phylogenetic analysis using single genes, the sequences are preprocessed before inferring phylogenies. A multiple alignment is constructed and the ambiguous parts are disregarded in the tree construction phase. The proposed method does not go through any such trimming of the sequences and processes them as given. Even with such an input of sequences (rendering a more noisy data) the proposed method is able to construct successful phylogenies both using whole genomes and single genes.

CONCLUSIONS

In this paper, we propose a new sequence distance measure and its variations. The proposed metric uses LZ complexity which relates the number of steps in a production process of a sequence to its complexity. We extend this idea to generate a sequence from a different sequence linking the resulting number of steps to the 'closeness' between two sequences.

Unlike most existing phylogeny construction methods, the proposed method does not require multiple alignment and is fully automatic. Therefore, we are able to perform comparisons at the whole genome level where multiple alignment based strategies fail. Unequal sequence length or the relatively different positioning of similar regions between sequences (such as different gene order in genomes) are not problematic as the proposed method handles both cases naturally. Moreover, we use no approximations and assumptions in calculating the distance between sequences. The proposed metrics utilize the entire information contained in the sequences and require no human intervention.

The results show that the proposed method can successfully construct phylogenies using either whole genomes or single genes. This is quite promising as the genome level phylogeny construction becomes important with the arrival of such data. 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.

REFERENCES

Apostolico,A. and Lonardi,S. (2000) Compression of biological sequences by greedy off-line textual substitution. In Storer,J.A. and Cohn,M. (eds.), *IEEE Data Compression Conference, DCC*, IEEE Computer Society TCC, Snowbird, Utah, pp. 143–152.

Barry,D. and Hartigan,J.A. (1987) Statistical analysis of hominoid molecular evolution. *Stat. Sci.*, **2**, 191–210.

Benedetto,D., Caglioti,E. and Loreto,V (2002) Language trees and zipping. *Phys. Rev. Lett.*, **88**, 048702.

Bennett,C.H., Gacs,P., Li,M., Vitanyi,P. and Zurek,W. (1998) Information distance. *IEEE T. Inform. Theory*, **44**, 1407–1423.

Berman,P., Hannenhalli,S. and Karpinski,M. (2001). Approximation algorithm for sorting by reversals. Technical Report TR01-047, ECCC.

Boore,J.L. and Brown,W.M. (1998) Big trees from little genomes: mithochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev.*, **8**, 668–674.

Camin,J. and Sokal,R. (1965) A method for deducing branching sequences in phylogeny. *Evolution*, **19**, 311–326.

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

Cao,Y., Okada,N. and Hasegawa,M. (1997) Phylogenetic position of guinea pigs revisited. *Mol. Biol. Evol.*, **14**, 461–464.

Cavalli-Sforza,L.L. and Edwards,A.W.F. (1967) Phylogenetic analysisis: models and estimation procedures. *Evolution*, **21**, 550–570.

Chen,X., Kwong,S. and Li,M (2000) A compression algorithm for DNA sequences and its applications in genome comparison. In Shamir,R., Miyano,S., Istrail,S., Pevzner,P. and Waterman,M. (eds) *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB)*, ACM Press, Tokyo, Japan, pp. 107–117.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1999) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

Eck,R.V. and Dayhoff,M.O. (1966) *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Silver Spring, MD, pp. 161–202.

Farach,M., Noordewier,M.O., Savari,S.A., Shepp,L.A., Wyner,A.D. and Ziv,J. (1995) On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In *Symposium on Discrete Algorithms*, pp. 48–57.

Felsenstein,J. (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.*, **22**, 240–249.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Felsenstein,J. (1989) PHYLIP (Phylogeny Inference Package). *Cladistics*, **5**, 164–166.

Felsenstein,J. and Churchill,G.A. (1996) A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Bio. Evol.*, **13**, 93–104.

Fitch,W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **35**, 406–416.

Fitz-Gibbon,S.T. and House,C.H. (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.*, **27**, 4218–4222.

Grumbach,S. and Tahi,F. (1993) Compression of DNA sequences. In *Data Compression Conference*, IEEE Computer Society Press, Sbowbird, Utah, USA.

Grumbach,S. and Tahi,F. (1994) A new challenge for compression algorithms: genetic sequences. *J. Info. Proc. Man.*, **30**, 875–866.

Hannenhalli,S. and Pevzner,P.A. (1995) Towards a computational theory of genome rearrangements. *Lect. Notes Comput. Sci.*, **1000**, 184–202.

Hannenhalli,S. and Pevzner,P.A. (1999) Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *JACM*, **46**, 1–27.

Huynen,M.A., Snel,B.,III, W.L. and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.

Jukes,T.H. and Cantor,C.R. (1969) *Mammalian Protein Metabolism*, Academic Press, New York, pp. 21–132.

Kececioglu,J. and Ravi,R. (1995) Of mice and men. Evolutionary distances. In *Proceedings of the 6th ACM-SIAM Symposium on Discrete Algorithms*, pp. 604–613.

Kececioglu,J. and Ravi,R. (1998) Reconstructing a history of recombinations from a set of sequences. *Discrete Appl. Math.*, **88**, 239–260.

Kececioglu,J. and Sankoff,D. (1995) Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica*, **13**, 180–210.

Kimura,M. (1980) A simple model for estimating evolutionary rates of base substitiutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.

Kishino,H. and Hasegawa,M. (1989) Evolution of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoida. *J. Mol. Evol.*, **29**, 170–179.

Lake,J.A. (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proc. Natl Acad. Sci. USA*, **91**, 1455–1459.

Lanctot,J.K., Li,M. and Yang,E.-H. (2000) Estimating DNA sequence entropy. In *Symposium on Discrete Algorithms* pp. 409–418.

Lempel,A. and Ziv,J. (1976) On the complexity of finite sequences. *IEEE T. Inform. Theory*, **22**, 75–81.

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

Li,M. and Vitanyi,P.M.B. (1997) *An Introduction to Kolmogorov complexity and its Approximations*, 2nd edn. Springer-Verlag, New York.

Lin,J. and Gerstein,M. (2000) Whole-genome trees based on the occurence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res.*, **10**, 808–818.

Loewenstern,D. and Yianilos,P.N. (1999) Significantly lower entropy estimates for natural dna sequences. *J. Comput. Biol.*, **6**, 125–142.

Madsen,O., Scally,M., Douady,C.J., Kao,D.J., DeBry,R.W., Adkins,R., Amrine,H.M., Stanhope,M.J. de Jong,W.W. and Springer,M.S. (2001) Parallel adaptive radiations

in two major clades of placental mammals. *Nature*, **409**, 610–618.

Milosavljevic,A. (1993) Discovering sequence similarity by the algorithmic significance. In *Intelligent Systems for Molecular Biology*, AAAI Press, Vienna, pp. 284–291.

Murphy,W.J., Eizirik,E., O'brein,S.J., Madsen,O., Scally,M., Douady,C.J., Teeling,E., Ryder,O.A., Stanhope,M.J., de Jong,W.W. and Springer,M.S. (2001) Resolution of the early placental mammal radiation using bayesian phylogenetics. *Science*, **294**, 2348–2351.

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

Rivals,E., Dauchet,M., Delahaye,J.-P. and Delgrange,O. (1996) Compression and genetic sequences analysis. *Biochimie*, **78**, 315–322.

Rivals,E., Delgrange,O., Dauchet,M. and Delahaye,J. (1994) Compression and sequence comparison. In Apostolico,A. (ed.) *DIMACS Workshop on Sequence Comparison*.

Rivals,E., Delgrange,O., Delahaye,J.P., Dauchet,M., Delorme,M.O., Henaut,A. and Ollivier,E. (1997) Detection of significant patterns by compression algorithms: the case of approximate tandem repeats in DNA sequences. *Comput. Appl. Biosci.*, **13**, 131–136.

Rowe,D.L. and Honeycutt,R.L. (2002) Phylogenetic relationships, ecological correlates, and molecular evolution within the Cavioidea (Mammalia, Rodentia). *Mol. Biol. Evol.*, **19**, 263–277.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Sankoff,D. (1999a) Geneome rearrangement with gene families. *Bioinformatics*, **15**, 909–917.

Sankoff,D. (1999b) Comparative mapping and genome rearrangement. In *From Jay Lush to Genomics: Visions For Animal Breeding and Genetics*, pp. 124–134.

Sankoff,D. and Blanchette,M. (1998) Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.*, **5**, 555–570.

Sankoff,D., Leduc,G., Antoine,N., Paquin,B., Lang,B.F. and Cedergen,R. (1992) Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl Acad. Sci. USA*, **89**, 6575–6579.

Snel,B., Bork,P. and Huynen,M.A. (1999) Genome phylogeny based on gene content. *Nat. Genet.*, **21**, 108–110.

Snel,B., Bork,P. and Huynen,M.A. (2000) Genome evolution: gene fusion versus gene fission. *Trends Genet.*, **16**, 9–11.

Snel,B., Bork,P. and Huynen,M.A. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.*, **12**, 17–25.

Tekaia,F., Lazcano,A. and Dujon,B. (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Res.*, **9**, 550–557.

Thompson,J.D., Higgins,D. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Varre,J.-S., Delahaye,J.-P. and Rivals,E. (1999) Transformation distances: a family of dissimilarity measures based on movements of segments. *Bioinformatics*, **15**, 194–202.

Yang,Z. (1996) Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.*, **42**, 294–307.

Ziv,J. and Lempel,A. (1977) A universal algorithm for sequential data compression. *IEEE T. Inform. Theory*, **23**, 337–343.

Ziv,J. and Lempel,A. (1978) Compression of indiviual sequences via variable-rate coding. *IEEE T. Inform. Theory*, **24**, 530–536.

Ziv,J. and Merhav,N. (1993) A measure of relative entropy between individual sequences with application to universal classification. *IEEE T. Inform. Theory*, **39**, 1270–1279.

APPENDIX

LEMMA 1. $c(SQ)-c(S) \leq c(ST)-c(S)+c(TQ)-c(T)$

PROOF. First we note that

$$c(STQ) - c(ST) \leq c(TQ) - c(T). \qquad (1)$$

The LHS of (1) is the number of components Q would have when parsed using ST and the RHS is the number of components Q would have when parsed using T. Having ST instead of T cannot increase the number of components in parsing of Q. Since $c(SQ)-c(S) \leq c(STQ)-c(S)$, using (1) we have $c(SQ) - c(S) \leq c(ST) - c(S) + c(TQ) - c(T)$.

COROLLARY 1. $c(Q) \leq c(TQ)$

PROOF. Let $S = \epsilon$, the empty string, in Lemma 1.

Let $S = A^{(n)}$ denote the sequence obtained by $n-1$ concatenations of the sequence A to itself. For the remainder of this appendix, if $S = A^{(n)}$, we consider S to be equal to A.

THEOREM 1. *The function $d(S, Q)$ is a distance metric.*

PROOF. By definition $d(\cdot,\cdot)$ satisfies the symmetry condition. The identity condition is satisfied up to an additive error term of $O(1)$ depending on whether the last component of the sequence is exhaustive or not. In order to prove the triangle inequality, we need to show:

$$\begin{aligned}\max\{c(SQ) - c(S), c(QS) - c(Q)\} \\ \leq \max\{c(ST) - c(S), c(TS) - c(T)\} \\ + \max\{c(TQ) - c(T), c(QT) - c(Q)\}\end{aligned}$$

From Lemma 1, we have the following two symmetric inequalities:

$$c(SQ) - c(S) \leq c(ST) - c(S) + c(TQ) - c(T)$$
$$c(QS) - c(Q) \leq c(QT) - c(Q) + c(TS) - c(T)$$

which proves the triangle inequality. Hence, the function $d(S, Q)$ is a distance metric.

THEOREM 2. *The function $d^*(S, Q)$ is a distance metric.*

PROOF. Again, by definition $d^*(\cdot,\cdot)$ satisfies the symmetry condition. The identity condition is satisfied up to an additive error term of $O(1/c(S))$ depending on whether the last

component of the sequence S is exhaustive or not. We now need to show that $d^*(\cdot,\cdot)$ satisfies the triangle inequality:

$$\begin{aligned}&\frac{\max\{c(SQ)-c(S),c(QS)-c(Q)\}}{\max\{c(S),c(Q)\}}\\&\leq\frac{\max\{c(ST)-c(S),c(TS)-c(T)\}}{\max\{c(S),c(T)\}}\\&\quad+\frac{\max\{c(TQ)-c(T),c(QT)-c(Q)\}}{\max\{c(T),c(Q)\}}\end{aligned}$$

Without loss of generality, assume $c(Q)\leq c(S)$.

Case 1: Assume $c(T)\leq c(S)$. In this case we have:

$$\begin{aligned}&\frac{\max\{c(SQ)-c(S),c(QS)-c(Q)\}}{\max\{c(S),c(Q)\}}\\&=\frac{\max\{c(SQ)-c(S),c(QS)-c(Q)\}}{c(S)}\\&\leq\frac{\max\{c(ST)-c(S),c(TS)-c(T)\}}{c(S)}\\&\quad+\frac{\max\{c(TQ)-c(T),c(QT)-c(Q)\}}{c(S)}\\&\leq\frac{\max\{c(ST)-c(S),c(TS)-c(T)\}}{\max\{c(S),c(T)\}}\\&\quad+\frac{\max\{c(TQ)-c(T),c(QT)-c(Q)\}}{\max\{c(T),c(Q)\}}\end{aligned}$$

where the first inequality follows from Theorem 1 and the second inequality follows from the assumptions.

Case 2: Assume $c(S)\leq c(T)$. Since $c(SQ)=c(QS)$ up to a logarithmic factor (Lempel and Ziv, 1976), due to the assumptions we have:

$$\begin{aligned}\max\{c(SQ)-c(S),c(QS)-c(Q)\}&=c(QS)-c(Q)\\\max\{c(ST)-c(S),c(TS)-c(T)\}&=c(ST)-c(S)\\\max\{c(TQ)-c(T),c(QT)-c(Q)\}&=c(QT)-c(Q)\end{aligned}$$

Therefore, we need to show:

$$\frac{c(QS)-c(Q)}{c(S)}\leq\frac{c(QT)-c(Q)+c(ST)-c(S)}{c(T)}$$

Since the LHS of the above inequality is ≤ 1, we can start by adding the non-negative quantity $c(T)-c(S)$ to both the numerator and denominator of the LHS:

$$\begin{aligned}\frac{c(QS)-c(Q)}{c(S)}&\leq\frac{c(QS)-c(Q)+c(T)-c(S)}{c(T)}\\&\underset{\leq}{\log}\frac{c(QT)-c(Q)+c(ST)-c(S)}{c(T)}\end{aligned}$$

where the last inequality follows from Lemma 1 using it in the form $c(T)\leq c(QT)+c(TS)-c(QS)$ and log means the inequality holds up to a logarithmic factor. $\underset{\leq}{}$

THEOREM 3. *The function $d_1(S,Q)$ is a distance metric.*

PROOF. By definition $d_1(\cdot,\cdot)$ satisfies the symmetry condition. The identity condition is satisfied up to an additive error term of $O(2)$ depending on whether the last component of the sequence is exhaustive or not. The triangle inequality follows directly from Lemma 1.

THEOREM 4. *The function $d_1^*(S,Q)$ is a distance metric (The corresponding theorem and proof for d_1^{**} is almost identical and therefore will not be included here.).*

PROOF. Again, by definition $d_1^*(\cdot,\cdot)$ satisfies the symmetry condition (up to a logarithmic factor). The identity condition is satisfied up to an additive error term of $O(2/c(S))$ depending on whether the last component of the sequence S is exhaustive or not. Next, we prove the triangle inequality for $d_1^*(\cdot,\cdot)$. It suffices to show the two inequalities

$$\begin{aligned}\frac{c(SQ)-c(S)}{c(SQ)}&\leq\frac{c(ST)-c(S)}{c(ST)}+\frac{c(TQ)-c(T)}{c(TQ)}\\\frac{c(QS)-c(Q)}{c(SQ)}&\leq\frac{c(TS)-c(T)}{c(ST)}+\frac{c(QT)-c(Q)}{c(TQ)}\end{aligned}$$

Since these two inequalities are symmetric we will prove the first one only. Let $\delta=c(TQ)-c(T)+c(ST)-c(S)-[c(SQ)-c(S)]$. From Lemma 1, $0\leq\delta$. As $[c(SQ)-c(S)]/[c(SQ)]\leq 1$ it follows that

$$\begin{aligned}\frac{c(SQ)-c(S)}{c(SQ)}&\leq\frac{c(SQ)-c(S)+\delta}{c(SQ)+\delta}\\&=\frac{c(ST)-c(S)+c(TQ)-c(T)}{c(ST)+c(TQ)-c(T)}\\&\leq\frac{c(ST)-c(S)}{c(ST)}+\frac{c(TQ)-c(T)}{c(TQ)}\end{aligned}$$

since $C(ST)+C(TQ)-C(T)\geq C(ST)$ and $C(ST)+C(TQ)-C(T)\geq C(TQ)$ (from Corollary 1). As the second inequality is proved symmetrically we have

$$\begin{aligned}&\frac{c(SQ)-c(S)+c(QS)-c(Q)}{c(SQ)}\\&\leq\frac{c(ST)-c(S)+c(TS)-c(T)}{c(ST)}\\&\quad+\frac{c(TQ)-c(T)+c(QT)-c(Q)}{c(TQ)}\end{aligned}$$

which is what we needed to show.