United States Patent Application Publication: SYSTEM AND METHOD FOR SEQUENCE DISTANCE MEASURE FOR PHYLOGENETIC TREE CONSTRUCTION

K Sayood
University of Nebraska-Lincoln, ksayood@unl.edu

Hasan H. Otu
Harvard Medical School, hotu2@unl.edu

Steven H. Hinrichs
University of Nebraska Medical Center, shinrich@unmc.edu

Follow this and additional works at: http://digitalcommons.unl.edu/electricalengineeringfacpub
Part of the Computer Engineering Commons, and the Electrical and Computer Engineering Commons

http://digitalcommons.unl.edu/electricalengineeringfacpub/428

This Article is brought to you for free and open access by the Electrical & Computer Engineering, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Publications from the Department of Electrical and Computer Engineering by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
The present invention permits identification of biological materials following recovery of DNA using standard techniques by comparing a mathematical characterization of the unknown sequence with the mathematical characterization of DNA sequences of known genera and species. The clinical identification of infectious organisms is required for accurate diagnosis and selection of antimicrobial therapeutics. The invention allows an ab initio approach with the potential for rapid identification of biological materials of unknown origin. The approach provides for identification and classification of emergent or new organisms without previous phenotypic identification. The technique may also be used in monitoring situations where the need exists for classification of material into broad categories of bacteria which could have an immediate impact on bio-terrorism prevention.
1. RECEIVE FIRST NUCLEOTIDE OF FIRST SEQUENCE

2. ENTER AS WORD IN DICTIONARY

3. DETERMINE SUM OF WORDS IN DICTIONARY

4. RECEIVE NEXT NUCLEOTIDE SEQUENCE

5. MAKE WORD FROM IN NUCLEOTIDE

6. WORD IN DICTIONARY?

FIG. 1A
FIG. 1B
FIG. 2.

(a)  
A2  
A1  
B'  
A'  
B1  
B2  

(b)  
A2  
A1  
B'  
A'  
B2  
B1  

(c)  
A'  
A2  
A1  
B2  
B1  

FIG. 3.

- opossum
- wallaroo
- platypus
- f whale
- h whale
- cow
- horse
- rhino
- h seal
- g seal
- cat
- human
- macaques
- pig e.
- gorilla
- orangutan
- gibbon
- baboon
- rat
- mouse
FIG. 4.

FIG. 5.

<table>
<thead>
<tr>
<th></th>
<th>NOTHING</th>
<th>A</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>NOTHING</td>
<td>T</td>
<td>A, T</td>
</tr>
<tr>
<td>3</td>
<td>NOTHING</td>
<td>G</td>
<td>A, T, G</td>
</tr>
<tr>
<td>4</td>
<td>TG</td>
<td>A</td>
<td>A, T, G, TGA</td>
</tr>
<tr>
<td>5</td>
<td>ATG</td>
<td>A</td>
<td>A, T, G, TGA, ATG</td>
</tr>
</tbody>
</table>
FIG. 7.
FIG. 8.
FIG. 9.
FIG. 10.
SYSTEM AND METHOD FOR SEQUENCE DISTANCE MEASURE FOR PHYLOGENETIC TREE CONSTRUCTION

BACKGROUND OF THE INVENTION

[0001] 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.

[0002] Some of the approaches in the first category utilize various distance measures which use different models of nucleotide substitution or amino acid replacement. [2][3][8][30][32][34] The second category can further be divided into two groups based on the optimality criterion used in tree evaluation: parsimony [8][11][13][19] and maximum likelihood methods [15][16][18].

[0003] 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 analyses 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.

[0004] To overcome these problems, Sankoff et al. (1992) [51] 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 insertion order have been extensively studied for applications to genome-based phylogeny. [6][7][23][24][29][30][31][48][49][50] 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).

[0005] Gene content was proposed by Snel et al. (1999) [52] 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. [20][27][38][53][54] Li and Gerstein (2000) [38] constructed phylogenetic trees based on the occurrence of particular molecular features: presence or absence of either folds or orthologs throughout the whole genome. Takata et al. (1999) [55] 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).

[0006] In the early 1990s, various data compression approaches were applied to the analysis of genetic sequences. [14][21][22][41][45] 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) [21][22] coded exact repeats and palindromes in DNA along the lines of Lempel-Ziv (LZ) compression scheme [59] and used an arithmetic coder of order 2 when such structures are lacking. Rivals et al. (1994, 1996) [44][45] 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) [46] 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; Lancet et al., 2000; Apostolico and Lonardi, 2000. [1][35][39] 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. [22] Although the proposed distance was not mathematically valid and had some other problems, it applied data compression to phylogeny construction.

[0007] Varre et al. (1999) [57] 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) [22], is asymmetric. Chen et al. (2000) [12] 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). [36] Ziv and Merhav (1993) [4] and Bennett et al. (1998) [6] provide a detailed analysis of information distance in statistical and algorithmic settings.

[0008] The distance proposed by Chen et al. (2000) and Li et al. (2001) is 1-[K(S)-K(SQ)]/K(SQ), where K(S) is the Kolmogorov complexity of S, K(SQ) is the conditional Kolmogorov complexity of S given Q and K(SQ) is the Kolmogorov complexity of the sequence S concatenated with Q. K(SQ) is the shortest program that outputs S when the input is Q on a universal computer and K(S) is K(S|_), where _ is the empty string. [12][35] Kolmogorov complexity is an algorithmic measure of information in (Li and Vitanyi, 1997) but it is a theoretical limit and generally can only be approximated. [37] In calculating the aforementioned distance, K(MQ) is approximated by the length of the compressed result of S (using the program GenCompress) given Q.

[0009] Benedetto et al. (2002) [3] used a similar idea where relative complexity between sequences S and Q is approximated as it is done by Chen et al. (2000) [12], 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.

[0010] Methods that rely on the compressibility of a sequence using a compression package have an inherent flaw as these 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 measure based on Kolmogorov complexity (implicitly or explicitly) would not necessarily hold for these approximations and the resulting distance may be misleading depending on the performance of the compression algorithms on certain sequences.

[0011] Traditional methods are based on phenotypic identification of organisms following the use of culture techniques. Clinical microbiology is currently undergoing a major transition to the use of molecular approaches. However, molecular approaches require the operator to select from among a list of probes or amplification primers for the identification process to proceed. In other words, the operator must have some predetermined idea as to the name or nature of the organism to be identified.

[0012] What would be beneficial is a system and method for phylogeny construction that does not require multiple alignment and is fully automatic. It would also be beneficial not to have to use approximations and assumptions in calculating the distance between sequences.

SUMMARY OF THE INVENTION

[0013] In one embodiment, the present invention permits identification of biological materials following recovery of DNA using standard techniques by comparing a mathematical characterization of the unknown sequence with the mathematical characterization of DNA sequences of known genera and species. The clinical identification of infectious organisms is required for accurate diagnosis and selection of antimicrobial therapeutics.

[0014] In another embodiment, this invention allows an ab initio approach with the potential for rapid identification of biological materials of unknown origin. The approach provides for identification and classification of emergent or new organisms without previous phenotypic identification. The technique may also be used in monitoring situations where the need exists for classification of material into broad categories of bacteria, such as Bacillus versus Francisella, which could have an immediate impact on bio-terrorism prevention.

[0015] Unequal sequence length or the relatively different positioning of similar regions between sequences (such as different gene order in genomes) is not problematic as the proposed method handles both cases naturally. The proposed metrics utilize the entire information contained in the sequences and require no human intervention.

BRIEF DESCRIPTION OF SEVERAL VIEWS OF THE DRAWINGS

[0016] FIG. 1A is a flow diagram of a method for determining the number of sets of nucleotides in a first sequence in accordance with an embodiment of the present invention;

[0017] FIG. 1B is a flow diagram of a method for determining the number of sets of nucleotides in a second sequence that are different from the pattern of nucleotides in a first sequence in accordance with an embodiment of the present invention;

[0018] FIG. 2 is a diagram of exemplary phylogenetic trees generated using the method of the present invention compared with phylogenetic trees generated using other methods;

[0019] FIG. 3 is an exemplary phylogenetic tree generated using other methods;

[0020] FIG. 4 is an exemplary phylogenetic tree generated using the method of the present invention;

[0021] FIG. 5 is an exemplary table generated using the method of the present invention;

[0022] FIG. 6 is an exemplary phylogenetic tree generated using the method of the present invention;

[0023] FIG. 7 is an exemplary phylogenetic tree generated using the method of the present invention;

[0024] FIG. 8 is an exemplary phylogenetic tree generated using the Jukes-Cantor method;

[0025] FIG. 9 is an exemplary phylogenetic tree generated using the method of the present invention; and

[0026] FIG. 10 includes exemplary phylogenetic trees using the method of the present invention where portions of the genetic sequence have been deleted.

DESCRIPTION OF INVENTION

[0027] The present invention provides a system and method to ascertain the provenance of a DNA sequence standard approaches use a base-by-base comparison of the unknown sequence with sequences in a database. In the system and method of the present invention a characterization of the target sequence is obtained using algorithms developed, rather than trying to match the DNA sequence of unknown origin and the DNA sequences in a database, base by base. This characterization is compared with characterizations of DNA sequences belonging to different species. As these characterizations are very compact these comparisons can be done very quickly.

[0028] Furthermore, as characterizations of DNA sequences belonging to closely related species tend to be more similar than characterizations of DNA sequences belonging to more distantly related species, this allows identification of a DNA fragment at the level of genus even if the DNA sequence of the species is not available. Finally, the characterizations can be obtained with relatively short DNA sequences. Results described below were obtained using sequences from a few hundred bases long to a few million bases long. The system and method of the present invention allows for the identification of the genus and species of origin for a DNA sequence the proposed approach is much more effective.

[0029] The characterization of the DNA sequence used is a dictionary of words or database of units that can be used to build that sequence. Given a set of dictionaries corresponding to different genera and an unknown sequence it is possible to identify the sequence as follows: find the distance between the sequence and the dictionaries by counting the number of words in the sequence that are not in a given dictionary. The sequence is then identified with the dictionary closest to it.
The present invention will allow an ab initio approach with the potential for rapid identification of biological materials of unknown origin. An embodiment of the present invention will replace the need for sequencing of DNA for identification of emergent or new organisms without previous phenotypic identification. The technique could also be used in monitoring situations where the need exists for classification of material into broad categories of bacteria, such as Bacillus versus Francisellla, which could have an immediate impact on bio-terrorism prevention.

Let S, Q and R be sequences defined over an alphabet 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→R) if there exists an integer p≥l(S) such that Q(k)→R(p+k−1) for k=1, . . . , l(Q). For example AACGT→AACGTTCTG (SEQ ID NO: 1) with p=3 and AACGT→AACGTA 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 on 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,i) (denoted S(1,i)→S(i)) if S(1,i)→S(i+1)→S(i+2). For example AACGT→AACGTAC and AACGT→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,i)→S(1,i+i) (note that i=l(S)→S(1,1)). An m-step production process of S results in a parsing of S in which H(S)→S(1,1+i)→S(1,1+i+1) . . . S(1,1+i+m) is called the history of S and H(S)=S(i+1) is the ith component of H(S). For example S=AACGTACC, A-AAC-G-T-A-C-C, A-AAC-G-T-A-C-C and A-AAC-G-T-ACC are three different (production) histories of S.

If S(1,h) is not reproducible from S(1,h) [denoted S(1,h)→S(1,h)] then H(i) is called exhaustive. In other words, for H(i) 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.

Let c(S) be the number of components in a history of S. Then the LZ complexity of S is c(S)=min{c(S)} over all histories of S. It can be shown that c(S)=c(S) where c(S) is the number of components in the exhaustive history of S. This is quite intuitive as an exhaustive component is the longest possible at a given step of a production process.

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). In a sense, this is the proportional to the number of words in Q that are not present in 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) uses 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: c(SQ)≤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=AACGT ACC ATG (SEQ ID NO: 2), R=CT AGG ACT AT (SEQ ID NO: 3) and Q=AACGT ACC AA (SEQ ID NO: 4). The exhaustive histories of these sequences would be:

H(S) = A-AAC-G-T-ACC-AT-CT (SEQ ID NO:2)

H(R) = C-T-A-G-GGA-CTT-AT (SEQ ID NO:3)

H(Q) = A-C-G-G TA AA (SEQ ID NO:4)

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

H(SQ) = A AC G T ACC AT TG AGG CG ACCA (SEQ ID NO:5)

H(RQ) = C T A G GGA CTT AT AGC GT CA AA (SEQ ID NO:6)

Note that it took three steps to build Q in the production process of SQ. On the other hand, five steps were used 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 this can be observed looking at the patterns ACG and AGG which Q and S share. The number of steps it takes to generate a sequence Q from a sequence S by c(SQ)−c(S) can be formulated. Thus, if S is closer to Q than R then it would be expected that c(SQ)−c(S) is smaller than c(RQ)−c(R) as is the case in the example above. Based on this idea of closeness the four distance measures are defined as:

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

\[ d(S,Q) = \max(c(SQ) - c(S), c(S) - c(Q)) \]

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

Distance Measure 2: Given two sequences S and Q, define the function d'(S,Q) as:

\[ d'(S, Q) = \frac{\max(c(SQ) - c(S), c(S) - c(Q))}{\max(c(S), c(Q))} \]

Another distance measure that would naturally follow from the idea of building sequence Q using S is the “sum
Distance measure 3: Given two sequences S and Q, define the function \( d_3(S, Q) \) as

\[
d_3(S, Q) = c(S) + c(Q) - c(SQ)
\]

Distance measure 4: Given two sequences S and Q, define the function \( d_4(S, Q) \) as

\[
d_4^*(S, Q) = \frac{c(SQ) - c(S) + c(Q) - c(QS)}{c(SQ) + c(QS)}
\]

An alternative definition would be

\[
d_4^*(S, Q) = \frac{c(SQ) - c(S) + c(Q) - c(QS)}{2(c(SQ) + c(QS))}
\]

A distance metric, \( D(\cdot, \cdot) \), should satisfy the following conditions:

1. \( D(S, Q) = 0 \) 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).
4. \( D(Q, R) + D(T, S) \leq \max(D(S, Q) + D(R, T), D(Q, T) + D(S, R)) \) (additivity).

With reference to FIG. 1A, a method of determining the number of sets of nucleotides in a first sequence \( 100 \) is shown. At step 104 the first nucleotide of a first sequence is received. At block 106, the nucleotide is counted and stored as a word in a dictionary for the first sequence. This process may include storing a nucleotide or set of nucleotides as a unit in a database and/or table. At block 108, it is determined whether there are any nucleotides remaining in the first sequence. If so, at block 112, the next nucleotide in the sequence is received. At step 114, it is determined whether the next nucleotide is a word in the database. If not, the nucleotide is counted and stored as a word in the dictionary or database for the first sequence at step 106.

If the nucleotide is already a word in the dictionary for the first sequence, the next nucleotide of the first sequence is received at step 116. At step 118, the two nucleotides are made into a word or set in the order of being received. In other words, they are a set of nucleotides in sequential order: the first nucleotide is first letter of the word and the second nucleotide is the second letter of the word. At step 120, it is determined whether the word made from two nucleotides is in the dictionary, database or table of words for the first sequence. If not, the word is counted and stored in the dictionary for the first sequence at step 106. If the word is already in the dictionary of the first sequence then at step 120, the next nucleotide of the first sequence is received. At step 118, a new word is made from the three nucleotides. Steps 116, 118 and 120 are continued until a new word of nucleotides is made that is not in the dictionary of the first sequence.

When at step 120 it is determined that the word created is not in the dictionary of the first sequence it is counted and stored as a word in the dictionary at step 106. This process continues until there are no remaining nucleotides in the first sequence at step 108. Then at step 110 the sum of the words in the dictionary for the first sequence is determined. The sum can later be used for calculating the distances between the first sequence and a second sequence. The method also keeps track of all nucleotides of the first sequence in the order they are received. The first nucleotide sequence can be used with the method described in FIG. 1B.

With reference to FIG. 1B, a method for determining the number of words of nucleotides in a second sequence that are not in the sequence of nucleotides in a first sequence \( 101 \) is shown. At step 103, a first nucleic acid sequence is received. The first nucleic acid sequence can be the one generated by the method described in FIG. 1A or may be received from another table or database. At step 105, the next two nucleotides of the second sequence are received. If this is the beginning of the method, the first two nucleotides of the second sequence are the next two nucleotides received. The nucleotides are made into a word or set at step 107. At this point, the word is two nucleotides long. At step 109, the word is compared to the first sequence to determine if the word is in the nucleotide pattern of the first sequence. If it is not in the first sequence, at step 111, the word is counted and stored as a unit in the dictionary for the second sequence.

If at step 109, it is determined that the word is not in the first sequence, the next nucleotide from the second sequence is received at step 113. At step 107 the nucleotides are made into word or set. In this instance, the word made is three nucleotides long and the nucleotides are in the order received. At step 109, it is determined whether the new word is within the first nucleic acid sequence. Steps 107, 109 and 113 are continued until a new word is made that is not within the pattern of the first nucleic acid sequence. The new word that is not within the pattern of the first nucleic acid sequence is counted and stored as a word in the dictionary for the second nucleic acid sequence at step 111.

The method 101 is continued until there are no more nucleotides in the second nucleic acid sequence. Following, the sum of words in the dictionary for the second sequence is determined. The sum of the words in the dictionary for the first sequence is subtracted from the sum of the words in the dictionary for the second sequence to determine the difference. This difference is used in a number of ways, as described in detail above and in the following examples, to determine the distance of the two sequences from one another.

Example 1

In this example it is shown how all four distance measures defined above satisfy the first three conditions and are, therefore, valid distance metrics. The fourth condition was tested by comparing a distance matrix created by the proposed metrics with one reconstructed from the branch lengths of the resulting tree.
LEMMA 1. $c(SQ) - c(S) - c(ST) + c(TQ) - c(T)$

PROOF. First note:

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

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 $c(SQ) - c(S) = c(STQ) - c(S)$, using (1) we have $c(SQ) - c(S) = c(ST) - c(TQ) - c(T)$.

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

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

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 $S$ is exhaustive or not. To prove the triangle inequality, show:

$$\max\{c(ST) - c(S), c(TS) - c(T)\} \leq \max\{c(S), c(T)\} \max\{c(TO) - c(T), c(OT) - c(O)\} \max\{c(T), c(Q)\}$$

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

Case 1. Assume $c(S) \leq c(T)$. Since $c(SQ) = c(QS)$ up to a logarithmic factor due to the assumptions we have:

$$\max\{c(ST), c(TS) - c(Q)\} = c(QS) - c(Q)$$
$$\max\{c(ST) - c(S), c(TS) - c(T)\} = c(ST) - c(T)$$
$$\max\{c(TQ) - c(T), c(QT) - c(Q)\} = c(QT) - c(Q)$$

Therefore, the following should be shown:

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

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

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 $S$ is exhaustive or not. The triangle inequality follows directly from Lemma 1.

THEOREM 4. The function $d^*(S, Q)$ 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 $d^*(\cdot, \cdot)$ 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, prove the triangle inequality for $d^*(\cdot, \cdot)$. It suffices to show the two inequalities...
Since these two inequalities are symmetric the first one is provided. Let \( \delta = c(T) - c(S)\). From Lemma 1, \( 0 \leq \delta \). As \( C(SQ) - c(S) \geq C(ST) - c(S)\), it follows that

\[
\frac{c(SQ) - c(S)}{c(SQ)} \leq \frac{c(ST) - c(S)}{c(ST)} + \frac{c(TQ) - c(T)}{c(TQ)}
\]

since \( C(ST) + c(TQ) = c(ST) + c(TQ) = C(TQ) \leq C(TQ) \) (from Corollary 1). As the second inequality is proved symmetrically we have

\[
\frac{c(SQ) - c(S) + c(QS) - c(Q)}{c(SQ)} \leq \frac{c(ST) - c(S) + c(TS) - c(T)}{c(ST)} + \frac{c(TQ) - c(T) + c(QT) - c(Q)}{c(TQ)}
\]

which is what needed to be shown.

**EXAMPLE 2**

In this example it is shown 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. 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.

**Phylogenetic analysis based on DNA sequences has been intimately connected with multiple alignment.** Hence this embodiment of the present invention can be generally examined based on how well the implicit assumptions used for scoring the multiple alignment agree with particular evolutionary theories. As this embodiment does not depend on multiple alignments the validity of the embodiment is tested in two ways: using simulated data to show that the proposed distance measures can reasonably be represented by a tree. The superiority of the proposed method on existing techniques is shown using this simulated data. Secondly, it is determined whether the results generated by the proposed method agree with existing phylogenies. The trees are generated using the neighbor joining (NJ) program [47] in the PHYLIP package. [17] The multiple alignments required by parsimony and maximum likelihood methods are calculated using CLUSTAL W. [56]

For the simulated data, a 1000 bp sequence was used and evolved into two sequences \( A^* \) and \( B^* \) using point mutations (insertions, deletions, substitutions) and segment based modifications (inversions, transpositions, translocations, etc.). Similarly evolved \( A^* \) evolved into \( A1 \) and \( A2 \) and \( B^* \) into \( B1 \) and \( B2 \). Point mutations were introduced 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, \( A^* \) into \( A1 \) and \( A2 \) were evolved using point mutations only. Sequences \( A^* \), \( B^* \), \( A1 \), \( A2 \), \( B1 \) and \( B2 \) were used to build phylogenetic trees both using existing methods (maximum likelihood and parsimony) and the proposed method. The results are shown in Fig. 2. In Fig. 2, Phylogenetic trees obtained from the simulated sequences \( A^*, B^*, A1, A2, B1 \) and \( B2 \) using (a) Maximum Likelihood, (b) parsimony and (c) proposed methods are shown.

The trees obtained by five of the proposed distance measures resulted in identical topologies. In Fig. 2, the consensus tree obtained by those five trees along with the trees obtained by maximum likelihood and parsimony methods is shown. 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.

The proposed distance matrix used to build the NJ tree was compared 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, below, \( M \), \( R \) and \( M - R \) using \( d^* \) for the simulated data set are presented. The corresponding results were omitted 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.

**TABLE 1.**

<table>
<thead>
<tr>
<th>Scq.</th>
<th>A2</th>
<th>A1</th>
<th>B1</th>
<th>B2</th>
<th>A'</th>
<th>B'</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.7348</td>
<td>0.8093</td>
<td>0.8104</td>
<td>0.7535</td>
<td>0.8056</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>0.8000</td>
<td>0.8046</td>
<td>0.7767</td>
<td>0.8139</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>0.7488</td>
<td>0.8000</td>
<td>0.7674</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>0.8056</td>
<td>0.7488</td>
<td>0.7740</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A'</td>
<td>0.7488</td>
<td>0.8015</td>
<td>0.8161</td>
<td>0.7630</td>
<td>0.7180</td>
<td></td>
</tr>
<tr>
<td>A'</td>
<td>0.8056</td>
<td>0.8203</td>
<td>0.7672</td>
<td>0.7221</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>0.7488</td>
<td>0.7878</td>
<td>0.7740</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>0.8024</td>
<td>0.7886</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A'</td>
<td>0.7042</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A'</td>
<td>0.0000</td>
<td>0.0077</td>
<td>0.0057</td>
<td>0.0095</td>
<td>0.0686</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>0.0056</td>
<td>0.0156</td>
<td>0.0095</td>
<td>0.0918</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>0.0000</td>
<td>0.0121</td>
<td>0.0063</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>0.0024</td>
<td>0.0063</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A'</td>
<td>0.0702</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

**EXAMPLE 3**

[0082] In this example, it is shown 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 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 [9] Studies using the whole mtDNA sequences suggest the outgroup status of rodents relative to eutherians 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.

[0083] The first group of sequences was chosen from the controversial data set using the following mtDNA sequences from Gen-Bank: National Center for Biotechnology Information, National Library of Medicine, Building 38A, Bethesda, Md.: 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 (Balaenoptera physalus, X61145), blue whale (Balaenoptera musculus, X72204), cow (Bos taurus, V00654), rat (Rattus norvegicus, X14848), mouse (Mus musculus, V00711), opossum (Didelphis virginiana, Z29573), wallaroo (Macropus robustus, Y10524) and platypus (Ornithorhynchus anatinus, X83427). Rodent species were kept to murids only and marsupials and monotremes were used as outgroup.

[0084] The proposed distance measures to the complete mitochondrial genomes listed above were applied. All five metrics (d, d±, d*, d±*, and d±**) resulted in identical trees. In FIG. 3, the consensus of these five trees is shown. The tree is in complete agreement with Cao et al. (1998) [9] confirming the outgroup status of rodents relative to eutherians and primates. In FIG. 3, the topology for eutherians using whole mtDNA where wallaroo, opossum and platypus were 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 Gen-Bank accession codes for these additional mtDNA sequences are as follows: squirrel (Sciurus vulgaris, A2238588), fat dormouse (Glis glis, A001562), guinean pig (Cavia porcellus, A222767), donkey (Equus asinus, X97337), Indian rhinoceros (Rhinoceros unicornis, X97336), dog (Canis familiaris, U96659), sheep (Ovis aries, AF010400), pig (Sus scrofa, A002189), and hippopotamus (Hippopotamus amphibius, A010957). 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 polyphly) and the phylogenetic position of guinea pigs. [10][43]

[0085] 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). The consensus tree of the five trees obtained using the proposed metrics is shown in FIG. 4. In FIG. 4, the consensus tree for the proposed distance metrics using complete mtDNA. The consensus phylogeny is in agreement with Reyes et al. (2000) [43] except for the position of guinea pig which remains an open question. [10] FIG. 4 groups squirrel with dormouse, which has shown to be based on strong molecular, palaeontological and morphological evidence. See [43] 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) [43].

[0086] The results presented in FIGS. 3 and 4 are in accord with (Li et al., 2001)[36], 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. [40][42] 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) [40] and Murphy et al. (2001) [42] use individual genes to build the trees. 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.

[0087] Both of these analyses used the whole mitochondrial genomes of the species. The results are not based on the phylogenies inferred using coding regions or individual genes. Instead the complete sequences as opposed to partial genome data were used. The phylogenies inferred using the proposed distances confirm that our method can successfully construct evolutionary histories using whole genome sequences.

**EXAMPLE 4**

[0088] The LZ-complexity of a sequence was obtained by counting the number of steps needed to generate a copy of the primary sequence starting from a null state. Each step involved a process of copying a nucleotide or a series of nucleotides for a sequence and then adding the next nucleotide from the sequence being analyzed. The number of steps needed to obtain the exhaustive library was identified as the LZ-complexity value of the given sequence. For example, the LZ-complexity of the simple sequence \('ATGSGAAAG'\) would be obtained as shown in FIG. 5. Since five steps were needed to generate the exhaustive library, the LZ-complexity value of the analyzed sequence was \(5^5\). If long fragments of repeat sequence had been present, the number of steps required to construct the exhaustive library would have been small compared to a random nonrepetitive sequence. A random nonrepetitive sequence generates a high LZ-complexity value. The complexity of a sequence \('X'\) compared to a sequence \('Y'\) is known as the RCM of \('X'\) with respect to \('Y'\). This is the number of steps required to construct sequence \('X'\) beginning with the set of nucleotide sequences used to construct \('Y'\). The RCM of sequence \('X'\)
with respect to 'Y' is defined as the number of production steps required to construct an exhaustive library of nucleotide combinations of sequence 'X' beginning with the exhaustive library of nucleotide combinations within sequence 'Y'. The exhaustive library of a sequence is defined as the smallest number of distinct nucleotides or nucleotide combinations required to construct the sequence using a copying process. Mathematically, the distance 'D' between sequences 'X' and 'Y', is defined as: $D(X, Y) = \min\{RCM(X, Y), RCM(Y, X)/\max(\{RCM(X, Y), RCM(Y, X)\})\}$, where _ represents the empty sequence. The distance matrix is a two dimensional table where the element in the 'ith' row and the 'jth' column are the distances between the 'ith' and 'jth' sequence.

The method may also be described with reference to FIG. 5 where steps involved in generating an exhaustive library of a given sequence ATGTGAATGT are shown. In step 1, the first nucleotide from the analyzed sequence is added to generate a new sequence 'Z'. In step 2, the analyzed sequence is scanned and any new nucleotides not present in the generated sequence are added, (CT is added to the sequence 'Z' generated in Step 1 to give A,T). In step 3, the analyzed sequence is scanned for the presence of new nucleotide combinations. Finding none in the example, the nucleotide 'G' is added to 'Z' to generate A,T,G. In step 4, the next nucleotide in the analyzed sequence is scanned to copy TG from the most recently generated sequence 'Z' and 'A' added to generate the sequence A,T,G,T,G,A.T.G. In step 5, ATG is copied from the newly generated sequence 'Z' to give the final sequence of A,T,G,T,G,A.T.G. The A,T,G,T,G,A.T,G constitute the exhaustive library of nucleotides for the analyzed sequence ATGTGAATGT, resulting in the LZ-complexity value to equal 5 based on the number of steps required.

Identical taxon samples for the cytochrome b gene ranging in size from 396 to 1158 bp were used from published source. The 18S rDNA sequences ranging from 1683 to 1800 bp was identical to the published source. The 467 to 709 bp ITS sequences (which included the 5.8S genomic region) used in this study were obtained using the methods of Henry et al. (2000) [64] and included the following (Accession numbers from GenBank, National Center for Biotechnology Information, National Library of Medicine, Building 8A, Bethesda, Md. 20894): Ajellomyces capsulatus (AB071828 to 31 and AF038353), Ajellomyces dermatitidis (AF183912 and AF038358), Anthrodema benhamiae (AF038359), Aspergillus flavus (AF138287), A. fumigatus (AF138288), A. niger (AF138904), A. sp. (AJ001332), A. terreus (AF138290), A. ustus (AJ157507), Auaxthrorn umbrinum (AY177308), A. umbrinum (AY177309), Byssomchlamys fujva (AY306014), Cladophialaphora bantiana (AJ131079), Cordyceps capitata (U57668), Emericella nidulans (AF138289), Esperinichium pinetorum (AY354240), Eurotium repens (AY360405), Filobasidiella neoformans (AF162916), Fusarium oxysporum (AF165875), Geosmithia argeliae (AF033389), G. cylindrospora (AF033386), G. emersonii (AF033387), G. lavendula (AF033385), Gymnascella halyinospora (AF129853 and AF129854), Hamigera avellanea (AB105350), H. striata (AF454074), H. striata (AF454073), Hypoxylon cohaerens (AJ390399), Malbranchea dendritica (AY177310), Monascus purpureus (AF458473), Monascus sp. (AF458474-76), Neocellula reticulata (AF165874), Paracoccidioides brasiliensis (AF038360), Pseudallescheria boydii (AF181558), and Trichophyton tonsurans (AB094675, AB094674, AB094659, AB094658). The Coccidioides immitis sequence was obtained.

Generation of distance matrix and phylogenetic tree construction The relative complexity measure (RCM) for creation of the distance matrix was utilized as previously described. The distance-based trees were generated with the neighbor-joining option of the phylogeny inference package, PHYLYP 3.5c using the RCM algorithm generated distance matrix as input. [17]

For the set of fungal sequences, distances between all pairs were calculated and used to construct a distance matrix. The RCM was used to determine the distance matrices from each of the three molecular targets (cytochrome b gene, 18S rDNA gene and the combined ITS regions). These distance matrices were used by the neighbor-joining algorithm to obtain the respective phylogenetic trees. To determine the discriminatory power of RCM, phylogenetic trees were compared with previously published trees for cytochrome b and 18S rDNA sequences. For the analysis of cytochrome b sequences, the distance between the sequences was calculated using Kimura’s two-parameter model. [58] A total of 26 datasets were included in the analysis. The cytochrome b sequences were analyzed using 1000 bootstrap-resampled data sets by neighbor-joining method to access the branch point support values. [58] However, the phylogenetic analysis of aligned 18S rDNA sequences of 35 Otomyces, four Eurotiales species, and one Chlaetothyriales (as outgroups) were done using a maximum-likelihood multiple-hit correction with an empirical transition/transversion ratio, empirical base frequencies, a gamma distribution of 0.5, and four categories of variations. [26] The support values for the branches were obtained from 1000 bootstrap-resampled data sets used for analysis by both the neighbor-joining and parsimony methods. [26]

The robustness of the proposed RCM approach was tested by examining the impact on the topology of the phylogenetic tree by reducing the overall length of the cytochrome b gene, ITS regions, and 18S rDNA gene sequences. Seven medically relevant fungal sequences were chosen and phylogenetic trees were generated by progressive removal of 10, 20, 30, 40 and 50% of the three original target sequences from either the 5' or the 3' ends. A comparison of the topology of phylogenetic trees obtained with the reference method and the RCM approach was done.

The cytochrome b gene sequences from multiple isolates of the order Saccharomycetales together with the sequence for Malassezia furfur as an outgroup, were evaluated to test the discriminatory power of the relative complexity measure (RCM) approach (FIG. 6). Referring to FIG. 6, a neighbor-joining tree of the most common Candida species based on nonaligned nucleotide sequences for the cytochrome b gene using the RCM approach is shown. Items designated with (*) refer to nucleotide sequence showing 100% identity to C. lusitaniae cytochrome b sequence. Malassezia furfur sequence was used as an outgroup for the analysis. Cl. Clavispora, C. = Candida, Is. = Issatchenkia, M. = Malassezia. The GenBank accession numbers are included for each species/strain. Consistent with the previously published phylogenetic relatedness of Candida sp.
using aligned sequences for construction of a tree, RCM showed similar topology for the multiple strains of this species. However, Candida tropicalis (AB044917) grouped together with Clavispora lusitaniae and subsequent analysis showed that both have 100% sequence identity. The RCM approach was applied to identical taxon samples of previously published 18S rDNA sequences, a commonly used and widely accepted target gene for phylogenetic analysis.

A comparison between the trees generated by non-aligned sequences using the RCM approach to the previously published trees that used aligned sequences demonstrated an overall similar topology except for the addition of a branch carrying Eremascus albicus (FIG. 7). Referring to FIG. 7, neighbor-joining tree of nonaligned 18S rDNA sequence of Onygenales, Eurotioidae and Chaetothyriales generated using the RCM approach. The differences in the relatedness that is observed compared to the previously published tree that used maximum-likelihood multiple-hit correction methods on aligned sequences are marked (*, branch; **, within the lade) and summarized in results. The orders and morphological groupings are shown on the right with vertical bars. A. = Aspergillus, Ap. = Apilanoascus, Ar. = Arthrodema, As. = Ascoscytum, B. = Blastomyces, Co. = Coccioides, Cl. = Clutonia, E. = Eurotiomycetidae, Em. = Emmonsia, Er. = Eremascus, Eu. = Euaspergillus, G. = Gymnoascideus, Gr. = Gymnascideus, H. = Histoplasma, I. = Lactisia, L. = Malbranchea, M. = Malbranchea, N. = Neosartorya, O. = Orygena, P. = Pectinotrichum, Pa. = Paracoccidioides, P. = Pselaphomyces, Ro. = Roridula, S. = Sporormia, T. = Trichophyton, U. = Uncinocarpus. GenBank accession numbers are indicated for each species/strain.

The Eurotiales (Aspergillus fumigatus, Neosartorya fischeri, Eurotiomycetidae, Emmonsia parva, and Histoplasma capsulatum) and the lade containing the saprophytic Sporormia warcupii and the keratinophilic fungus Malbranchea gypseae showed a close relatedness to the dimorphic fungi. In contrast to the phylogenetic tree depicting the relatedness of Eremascus albicus to the dermatophytes including Tidemakeria verrucosa, Arthrodema incurvum, Trichophyton rubrum and Arthrodema ciferrii, the RCM resulted in a distinct phylogenetic position for E. albics. Further, this close relatedness with the dermatophytes was replaced by a clad containing the members of Gymnoascideus (Gymnoascideus aurantiaca, Roridula hyladina and Gymnoascideus petalospora) and Onygenales.

The RCM approach was also applied to ITS sequences from 47 taxa of the subphylum Pezizomycotina (Enascomycotina) including 39 Eurotiomycetes (20 Eurotioidae, 19 Onygenales), 6 Sordariomycetes, 1 Chaetothyriomycetes, and a member of Basidiomycota as the outgroup. These included 17 medically important fungi, representing these different fungal orders. Comparison of the trees using the RCM approach with the trees constructed with identical taxon samples using standard methods requiring sequence alignment demonstrated an overall similar topology with some variation in outer branching. The distance-based tree generated with the unweighted pair group method with arithmetic mean (UPGMA) option of the PHYLIP 3.5c using the Jukes-Cantor algorithm gave the best bootstrap values for branching (FIG. 8) throughout the tree compared to the trees generated with neighbor-joining option or the use of other distance matrix algorithm.

Referring to FIG. 9, phylogenetic relationship of 17 medically relevant fungi among 47 genera of Pezizomycotina based on rDNA ITS1-5.8S gene-TTS2 sequence is shown. Distance based trees were generated with the (a) UPGMA using Jukes-Cantor algorithm and the (b) Neighbor-joining method using RCM algorithm. The significance of the branches in the UPGMA tree was tested by bootstrap analysis using 1000 bootstrap replications. Branches showing bootstrap values less than 50% are designated by *. Sequences were obtained from clinical isolates are: A. = A. fumigatus, Ap. = A. clavatus, Ar. = A. danubialis, As. = A. terreus, B. = B. dermatitidis, Co. = C. bovis, C. = C. albicans, Cl. = C. tropicalis, E. = E. verginae, Em. = E. earleii, Er. = E. gramineus, Eu. = E. penicillus, F. = F. pedrosoi, Fm. = F. mackenziae, G. = G. fujicola, Gs. = G. asahinae, H. = H. capsulatum, I. = I. farinose, J. = J. curvula, K. = K. lactis, L. = L. thermotolerans, M. = M. lanuginosa, N. = N. crassa, O. = O. salmonicolor, P. = P. butleri, Q. = Q. thermodurica, R. = R. oryzae, S. = S. sciuri, T. = T. rubrum, U. = U. maydis. GenBank accession numbers for these genera are listed in the Material and Methods section.

The topology presented by this tree compared well with the tree generated with neighbor-joining option using the distance matrix generated by using RCM algorithm, except for the positioning of Coccidioides immitis, Eurotiomycetes, A. fumigatus sp. CCF76 and B. dermatitidis (FIG. 9). However, a close relationship among molds and the dimorphic fungi was evident (FIG. 9). With an exception of Eurotiomycetes pinetorum and Hamigera avellanea, the RCM approach separated Eurotioidae, Onygenales and the members Sordariomycetes into distinct clusters (FIG. 9). The Arthrodema benhamiae, a teleomorph of Trichophyton grouped together with Trichophyton tonsurans. Similarly, Malbranchea dentriglum, a teleomorph of A. fumigatus grouped together with A. clavatus.

The reliability and robustness of the RCM approach was tested by examining the impact upon topology integrity by progressively deleting portions of the target genes. The phylogenetic trees generated after removal of 10, 20, 30, 40 and 50% of the three original target sequences from seven fungal species showed target sequence-dependent difference in the level of robustness (FIG. 10). Neighbor-joining trees of non-aligned cytochrome b gene (A, B), ITS1-5.8S-TTS2 region (C, D) and 18S gene (E, F) sequences generated using the RCM approach are shown. The trees on the left of FIG. 10 used full-length sequences and the trees on the right used randomly deleted partial sequences for analysis (deletion of 50% of cytochrome b gene sequences, 40% of ITS1-5.8S-TTS2 region sequence and 30% of the 18S target gene sequence). Vertical lines indicate change in topology.

The RCM generated trees retained appropriate topology after removal of up to 50% of the cytochrome b
gene sequence, 40% of the ITS sequence, and 30% of the 18S rDNA gene target sequence. In comparison, both reference methods failed to maintain baseline topology after removal of only 10% of the sequence. Deletion from either the 5' or 3' end had no impact on the outcome.

[0104] The relative complexity measure (RCM) evaluates the relatedness of DNA sequences with an advantage to not require alignment of sequences prior to analysis. Three genetic targets from fungi, including the mitochondrial cytochrome b gene, the 18S rDNA gene, and the ITS-1 and ITS-2 regions of the rDNA gene complex, were evaluated using the RCM approach to depict species relatedness.

[0105] Using the cytochrome b gene sequence, the RCM approach separated closely related Candida species. Consistent with an earlier report, strains of C. tropicalis, C. albicans and C. glabrata showed interspecies variation. RCM successfully detected the reported two DNA types due to 1.8% sequence variation for C. tropicalis. In addition, a 0.3% sequence variation reported for C. albicans and C. glabrata were apparent (FIG. 6). Candida dubliniensis showed a distinct phylogenetic position in relation to Candida albicans. The topologies obtained with the RCM approach closely resembled UPGMA trees described by Yokoyama et al. (2000) [58]. These observations reflect the sensitivity and comparability of RCM approach with the standard algorithmic approaches.

[0106] The RCM approach when applied to 18S rDNA sequences from 33 different fungal species (28 Onygenales, 4 Eurotiales, and 1 Chaetothyriales) generated similar topology to that obtained from the method using multiple aligned sequences. [26] Based on fruiting body structures, the phylogenetic positioning of Eremascus albus had been problematic due to a lack of diagnostic clavistothecial morphology. Eremascus albus has been alternatively classified with yeasts or in the class Plectomycetes, which includes genera producing clavistothecia. [5] The RCM approach showed a distinct phylogenetic position for Eremascus albus. The phylogenetic positioning of Capronia pilosella and the species in the Eurotiales strengthened the discriminating power of the RCM approach. The Capronia pilosella sequence was used as an outgroup sequence in the standard approach [26] and was appropriately outgrouped without the need for multiple sequence alignment. Consistent with earlier observations, species within the Eurotiales were grouped appropriately into one clade (FIG. 7). In contrast to the close phylogenetic affinities observed between Paracoccidioides brasiliensis and L. lobo of [26] the RCM approach separated L. lobo from the remaining dimorphic pathogens. The observed differences may be due to inaccuracies in the DNA sequence as submitted to GenBank. [58]

[0107] Recent studies have demonstrated the discriminatory power of the ITS regions to identify medically important yeasts. The RCM approach when applied to the ITS regions of the rDNA gene complex generated similar topology to that obtained from the method using multiple aligned sequences that showed strong bootstrap values for the branching. Differences in the outer branching could be attributed to the varying positions of Byssoschlamys fulva, Aspergillus sp., CCF76 and Eurotium repens. Consistent with earlier reports, the branching order suggests that the anamorphic genus Geosmithia is a polyphyletic taxon. The dimorphic fungal pathogens including P. brasiilensis clustered together with in the Onygenales as defined by a basal branch supported by 100% of the UPGMA bootstrapped data sets. Other members of the Onygenales including Coccidioides immitis show some variation in the evolutionary origins and the bootstrap values for these branches are relatively weaker. It is interesting to note that H. avellanea and E. pinctorum are outgrouped by both methods of analysis suggesting a possible misidentification of these sequences.

[0108] The robustness and reliability of the RCM approach was demonstrated by trees generated with appropriate topology after removal of up to 50% of the cytochrome b gene, 40% of the ITS sequences and 30% of the 18S gene target sequence.

[0109] The examples above demonstrate the multiple advantages of the RCM approach over standard phylogenetic tree construction approaches. These advantages included the elimination of the requirement for sequence alignment to obtain a distance matrix, the ability to analyze short diverse sequences, the lack of need for operator involvement in choosing the aligned sequence, and the avoidance of reliance on a predetermined, arbitrary decision tree. These advantages are expected to contribute towards whole genome phylogeny.

[0110] The present invention relates to 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. A sequence from a different sequence is generated 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, comparisons can be performed 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, no approximations and assumptions were used in calculating the distance between sequences. The method utilizes 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 will be useful as genome level phylogeny construction becomes important with the arrival of such data.

[0111] Although the invention has been described with reference to embodiments of the invention and the attached drawing figures, it is noted that substitutions may be made and equivalents employed herein without departing from the scope of the invention as claimed. For example, additions steps may be added and other steps omitted without departing from the scope of the invention.

REFERENCES


SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 6

<210> SEQ ID NO 1
<211> LENGTH: 10
<212> TYPE: DNA
ORGANISM: Artificial Sequence

FEATURE:

OTHER INFORMATION: Designed nucleic acid sequence to act as a template to create a dictionary of words.

SEQUENCE: 1

aacgtaacctg

SEQ ID NO 2
LENGTH: 11
TYPE: DNA
ORGANISM: Artificial Sequence
FEATURE:
OTHER INFORMATION: Designed nucleic acid sequence to act as a template to create a dictionary of words.

SEQUENCE: 2

aacgaatattg

SEQ ID NO 3
LENGTH: 12
TYPE: DNA
ORGANISM: Artificial Sequence
FEATURE:
OTHER INFORMATION: Designed nucleic acid sequence to act as a template to create a dictionary of words.

SEQUENCE: 3

cagggactatt

SEQ ID NO 4
LENGTH: 11
TYPE: DNA
ORGANISM: Artificial Sequence
FEATURE:
OTHER INFORMATION: Designed nucleic acid sequence to act as a template to create a dictionary of words.

SEQUENCE: 4

acgtaaccaaaa

SEQ ID NO 5
LENGTH: 23
TYPE: DNA
ORGANISM: Artificial Sequence
FEATURE:
OTHER INFORMATION: Designed nucleic acid sequence to act as a template to create a dictionary of words.

SEQUENCE: 5

aacgtaaccttgacgtaaccaaaa

SEQ ID NO 6
LENGTH: 23
TYPE: DNA
ORGANISM: Artificial Sequence
FEATURE:
OTHER INFORMATION: Designed nucleic acid sequence to act as a template to create a dictionary of words.

SEQUENCE: 6

cagggactattacgtaaccaaaa
We claim:
1. A method of determining whether a set of nucleotides is within a first nucleic acid sequence, the method comprising: receiving a first and a second nucleotide of a second nucleic acid sequence, the second nucleotide being a nucleotide after the first nucleotide; combining said first and second nucleotide into a sequential set; and comparing the first set of nucleotides to a first nucleic acid sequence to determine whether the first set of sequential nucleotides is within the first nucleic acid sequence.
2. The method of claim 1, wherein if the first set of nucleotides is not within the first nucleic acid sequence, storing said first set as a unit in a database for the second nucleic acid sequence.
3. The method of claim 1, wherein if the first set of nucleotides is within the first nucleic acid sequence, receiving a third nucleotide of the second nucleic acid sequence, the third nucleotide being a nucleotide after the second nucleotide.
4. The method of claim 3, further comprising: combining the first set of nucleotides with the third nucleotide to make a second sequential set.
5. The method of claim 4, further comprising: comparing the second set of nucleotides to a first nucleic acid sequence to determine whether the second set of sequential nucleotides is within the first nucleic acid sequence.
6. The method of claim 5, wherein if the second set of nucleotides is not within the first nucleic acid sequence, storing said second set as a unit in a database for the second nucleic acid sequence.
7. The method of claim 6, further comprising: determining the sum of all units stored for the second nucleic acid sequence.
8. The method of claim 7, further comprising: determining the difference between total number of units stored for a first nucleic acid sequence and the total number of units stored for the second nucleic acid sequence.
9. The method of claim 8, further comprising: utilizing the difference to determine the distance between the first nucleic acid sequence and the second nucleic acid sequence.
10. A computer readable medium comprising the method of claim 1.
11. A method of creating a database of nucleotide units for a first nucleic acid sequence, the method comprising: receiving a first nucleotide of a first nucleic acid sequence; determining whether the first nucleotide has been stored in a database as a unit for the first nucleic acid sequence; and if not, storing the first nucleotide as a unit for the first nucleic acid sequence.
12. The method of claim 11, wherein if the first nucleotide has been stored in a database as a unit for the first nucleic acid sequence, receiving a second nucleotide of the first nucleic acid sequence, the second nucleotide being a nucleotide after the first nucleotide.
13. The method of claim 12, further comprising: combining the first and second nucleotides into a sequential set.
14. The method of claim 13, further comprising: determining whether sequential set has been stored in a database as a unit for the first nucleic acid sequence.
15. The method of claim 14, wherein if the sequential set has been stored, receiving a third nucleotide of the first nucleic acid sequence, the third nucleotide being the next sequential nucleotide after the second nucleotide.
16. The method of claim 14, wherein if the sequential set has not been stored, storing said set as a unit for the first nucleic acid sequence.
17. The method of claim 16, further comprising: determining the sum of all units stored for the first nucleic acid sequence.
18. A computer readable medium comprising the method of claim 11.
19. A system for determining whether a set of nucleotides is within a first nucleic acid sequence, the system comprising: a receiving component for receiving a first and a second nucleotide of a second nucleic acid sequence, the second nucleotide being a nucleotide after the first nucleotide; a combining component for combining said first and second nucleotide into a sequential set and a comparing component for comparing the first set of nucleotides to a first nucleic acid sequence to determine whether the first set of sequential nucleotides is within the first nucleic acid sequence.
20. The system of claim 19, further comprising: a storing component for storing said first set as a unit in a database for the second nucleic acid sequence if the first set of nucleotides is not within the first nucleic acid sequence.
21. The system of claim 20, comprising: a second receiving module for receiving a third nucleotide of the second nucleic acid sequence if is determined that the first set of nucleotides is within the first nucleic acid sequence.
22. A method of determining the distance between two nucleic acid sequences, the method comprising: determining the number of words in a first nucleic acid sequence; combining the first sequence with a second nucleic acid sequence to make a combined nucleic acid sequence; determining the number of words in the combined nucleic acid sequence; and determining the difference between the number of words in the combined nucleic acid sequence and the first nucleic acid sequence to determine the distance between the first nucleic acid sequence and the second nucleic acid sequence.
23. A computer readable medium comprising the method of claim 22.