Sequence analysis

Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison

Qi Dai*, Yanchun Yang and Tianming Wang

Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China

Received on April 15, 2008; revised on August 16, 2008; accepted on August 17, 2008

Advance Access publication August 18, 2008

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Many proposed statistical measures can efficiently compare biological sequences to further infer their structures, functions and evolutionary information. They are related in spirit because all the ideas for sequence comparison try to use the information on the k-word distributions, Markov model or both. Motivated by adding k-word distributions to Markov model directly, we investigated two novel statistical measures for sequence comparison, called \[ w_r k_r \] and \[ S_2 k_r \].

Results: The proposed measures were tested by similarity search, evaluation on functionally related regulatory sequences and phylogenetic analysis. This offers the systematic and quantitative experimental assessment of our measures. Moreover, we compared our achievements with these based on alignment or alignment-free. We grouped our experiments into two sets. The first one, performed via ROC (receiver operating curve) analysis, aims at assessing the intrinsic ability of our statistical measures to compare for similar sequences from a database and discriminate functionally related regulatory sequences from unrelated sequences. The second one aims at assessing how well our statistical measure is used for phylogenetic analysis. The experimental assessment demonstrates that our similarity measures intending to incorporate k-word distributions into Markov model are more efficient.

Availability: The software, data and supplement material are freely available at http://math.dlut.edu.cn/daiqi/mlplusd.html.

Contact: daiali2004@yahoo.com.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Among biological sequence analysis, some important computational methods are similarity search, phylogenetic analysis and sequence clustering. The similarity search (Altschul et al., 1990, 1997; Pham, 2007; Pham and Zuegg, 2004) is to search a database of known function sequences and uses the structures and functions of the most closely matched known sequences to analyze the query sequence. Phylogenetic analysis (Felsenstein, 1996; Huelsenbeck and Roquist, 2001; Kumar et al., 2004; Lin et al., 2002; Ronquist and Huelsenbeck, 2003; Waddell and Steel, 1997; Waddell et al., 1999a, b, 2001) is the study of the evolutionary history among species. It can also provide useful information for pharmaceutical researchers to determine which medicinal species share the same medical qualities (Komatsu et al., 2001). Clustering sequences (Mohseni-Zadeh et al., 2004; Pipenbacher et al., 2002) is to get a biologically meaningful partition. When sequences are grouped into a family, it can give us some clues about the general features of that family. However, these efficient computational methods rely heavily on the similarity measures defined among biological sequences.

Because of the importance of research into similarity measure, numerous efficient algorithms have been developed, but challenges remain. Moreover, we believe that further improvements in the similarity measures will allow us to design more effective tools. One kind of similarity measures in this area is alignment-based measure. Waterman (1995) and Durbin et al. (1998) provided comprehensive reviews about this method. The search for optimal solutions using alignment-based method encounters difficulties: (i) computational aspect with regard to large databases (Pham and Zuegg, 2004); (ii) the scoring schemes chosen (Vinga and Almeida, 2003). Therefore, the emergence of research into alignment-free measure is apparent and necessary to overcome critical limitations of alignment-based measure.

Up to now, many efficient alignment-free measures have been proposed, but they are still in the early development compared with alignment-based measure. One of the comprehensive reviews (Vinga and Almeida, 2003) reported several similarity measures, such as the Euclidean distance (Blaisdell, 1986), Mahalanobis distance (Wu et al., 1997), Kulback-Leibler discrepancy (Wu et al., 2001), Cosine distance (Stuart et al., 2002), Pearson’s correlation coefficient (Fichant and Gautier, 1987), Kolmogorov complexity (Li et al., 2001) and Lempel-Ziv (LZ) complexity (Otu and Sayood, 2003). Recently, several novel statistical measures have been designed for DNA sequence comparison based on Markov model, such as SimMM (Pham and Zuegg, 2004) and D2z (Kantorovitz et al., 2007).

These statistical measures are often associated with k-word distributions, Markov model or both. Markov model and k-word distributions contain the important information about biological sequences. But, they are not absolutely equivalent. Indeed, in k-word distributions, the number of occurrence of each word is fixed, whereas in Markov model it is given on average. The idea for k-word analysis with background of Markov model was used by Van Helden (2004) and Kantorovitz et al. (2007). Whereas, our work presents two novel statistical measures for sequence analysis, which directly

*To whom correspondence should be addressed.
adds k-word distributions to Markov model. The contents can be summarized as follows:

1. Two statistical measures wre.k.r and S2.k.r, as the extended Jensen–Shannon divergence, are proposed. They are defined with the k-word concerns and calculated under Markov chains which may be different for two sequences. Particularly, the statistical measure, S2.k.r, is proved to be a valid distance measure.

2. Our measures are applied to extensive tests, e.g., similarity search and evaluation on functionally related regulatory sequences. The performance of our measures is compared with other methods based on alignments or alignment-free. The results demonstrate that they are promising statistical measures for sequence comparison with future possible improvement on structure and function prediction.

3. The distance measure, S2.k.r, is further used to construct phylogenetic tree. The results are in good agreement with the authoritative phylogenies, which indicates that S2.k.r is an efficient measure for phylogenetic analysis.

2 METHODS

2.1 The construction of Markov model for DNA sequences

Markov chain involves two probabilistic measures: transition probability matrix P and initial state distribution π, and can be denoted in a compact form:

\[ \lambda=(P,\pi). \]  

(1)

\[ P=[p_{ij}] \] denotes the transition matrix of a discrete-time Markov chain. Each state transition probability \( p_{ij} \) is defined as follows:

\[ p(i,j) = \tau(a_{i+1}S_j | a_i = S_i), \]  

(2)

where \( a_i \) stands for the actual state at time \( t_i \) (\( n = 1, 2, \ldots \)), and \( S_i \) and \( S_j \) are the \( i \)-th and \( j \)-th states of a space with \( N \) distinct states. In the context of DNA sequences, its state space is \([A,C,G,T]\). The state transition probabilities are subject to

\[ p(i,j) \geq 0 \quad \forall i, j, \]

(3)

\[ \sum_{j=1}^{N} p(i,j) = 1 \quad \forall i. \]

(4)

Additional \( \pi = \pi_i \) denotes the initial state distribution where

\[ \pi_i = p(a_1 = S_i), \]  

(5)

What we have presented above is the 1-step Markov chain, and the k-step Markov chain can be deduced from the 1-step Markov chain (Isaascson and Madsen, 1976). \( p_k \) is the probability that Markov chain is in state \( j \) after \( k \) steps from state \( i \), so

\[ p^k_{ij} = p[a_{i+k} = j | a_0 = i]. \]

(6)

For any three events \( A, B \) and \( C \), the following identity is known

\[ p(A \cap B | C) = p(A | B \cap C)p(B | C). \]

(7)

Interpreting \( A \) as \( x_k = j \), \( B \) as \( x_{k-1} = r \) and \( C \) as \( x_0 = l \), we have

\[ p^k_{ij} = \sum_{r=1}^{n} p[x_k = j | x_{k-1} = r, x_0 = l]p[x_{k-1} = r | x_0 = l] \]

\[ = \sum_{r=1}^{n} p[x_k = j | x_{k-1} = r]p[x_{k-1} = r | x_0 = l] \]

\[ = \sum_{r=1}^{n} p^r_{ij} p^k_{jr}, \]

(8)

which is known as the Chapman–Kolmogorov equation. Hence, the matrix with the elements \( p^k_{ij} \) are

\[ [p^k_{ij}] = P^k. \]

(9)

2.1.1 Estimate of the parameters Since the transition probabilities are priori unknown, we have to estimate them according to the observed sequences. Here, we estimate the parameters of Markov model by using the maximum likelihood method (Durbin et al., 1998).

2.1.2 Test of hypotheses A DNA sequence can be regarded as a discrete-time Markov chain. We test this hypotheses based on the following test statistic (Durbin et al., 1998)

\[ \Lambda = 2[\hat{L}(\hat{P}) - L(P^0)] \]

(10)

\[ = 2 \sum_{x \in S} \sum_{y \in S} N(xy) \log \left( \frac{N(xy)}{p^0_{xy} \sum_{z \in S} N(xz)} \right), \]

where \( P^0 \) is the given transition probability matrix, \( N(x, y) \) is the number of appearance of the pair states \((x, y)\) in the given sequence, and \( S = \{A, C, G, T\} \).

2.2 Statistical measures

2.2.1 Previous similarity measures We now describe seven similarity measures for biological sequences. Many similarity measures, for sequence comparison, are to fix a short word with length \( k \), compute the frequency distributions of all \( k \)-length words (also called ‘k-word’) in each sequence and assess the similarity of the two frequency distributions.

1. Euclidian distance: the Euclidian distance is one of the simple and common similarity measures of biological sequences. The similarity scores between two biological sequences is the Euclidian distance between their \( d^2 \)-dimensional vectors of \( k \)-word distributions (Blaisdell, 1986).

2. Cosine of the angle (cos): in order to derive estimate of relatedness from the vector definitions of biological sequences, Stuart et al. (2002) proposed the pairwise cosine for phylogenetic analysis from whole genome sequences.


4. Kullback–Leibler discrepancy (kdl): the relative entropy, also known as the Kullback–Leibler’s distance or the divergence, was introduced by Kullback and Leibler (1951). Wu et al. (2001) applied kdl to compare DNA sequences based on the frequencies of all \( k \)-words.

In contrast to the methods described above, several similarity measures are defined based on Markov model. DNA and protein sequences have been realized to comprise a mixture of local regions that consist of compositional characteristics and pseudo-periodic sequence patterns. Markov chain takes into account this ‘periodical’ behavior of the bio-signal by making use of state transition probability.

5. D2: the ‘D2’ score (Lippert et al., 2002) is a natural alignment-free similarity measure between two sequences, which is defined as the number of \( k \)-word matches between the sequences \( A \) and \( B \). Let \( I = \{(i, j): 1 \leq i \leq n_1, 1 \leq j \leq n_2, i \leq j \} \) be the index set, where \( n_1 = n_k - k + 1 \) and \( n_2 = n_k - k + 1 \) and \( Y(i, j) \) is the indicator variable for a match between the \( k \)-words starting at position \( i \) in \( A \) and at position \( j \) in \( B \), then the D2 score can be computed by

\[ D2(A, B) = \sum_{(i,j) \in I} Y(i,j). \]

(11)

6. D2c: in an application where several different pairs of sequences drawn from different distributions are to be compared, the previous
measure $D_2$ is not suitable. Kantorovitz et al. (2007) proposed and demonstrated the use of the ‘$D_2^*$’ score as such a ‘normalized’ measure which captures the statistical significance of the $D_2$ score as follows:

$$D_2(A, B) = \frac{D_2(A, B) - E(D_2)}{\sigma(D_2)}$$  (12)

In Equation (13), $E(D_2)$ and $\sigma(D_2)$ are the expectation and the SD of $D_2(A, B)$, respectively. The $D_2^*$ score is calculated under the general assumption of the two sequences generated by Markov chains, which may be different for the two sequences.

The symmetric form of $rre$ of comparing the similarity/dissimilarity between two constructed Markov models (Pham and Zuegg, 2004). If $x_1 = (p_1, \pi_1)$ and $x_2 = (p_2, \pi_2)$ denote two Markov models of two biological sequences, their probabilistic distance, denoted by $D(\lambda_1, \lambda_2)$, can be computed as:

$$D(\lambda_1, \lambda_2) = 1 - \exp\left(\frac{d(\lambda_1, \lambda_2) + d(\lambda_2, \lambda_1)}{2}\right),$$  (13)

$$d(\lambda_1, \lambda_2) = \frac{1}{T_j} \log \frac{P(O_{\lambda_1} | \lambda_1)}{P(O_{\lambda_2} | \lambda_2)}$$  (14)

where $O_{\lambda_j} = a_1 \cdots a_j$ is the biological sequence $j$, and $T_j$ is the length of the sequence $O_{\lambda_j}$.

2.2.2 Novel statistical measures

In this subsection, we present two novel statistical measures for biological sequence comparison.

1. Revised relative entropy ($rre$): relative entropy is the most important concept in both statistical biology and information theory. It has been explored as similarity measures, such as kl and SimMM to compare biological sequences. However, in an application where $p_1$ or $P(O_{\lambda_1} | \lambda_1)$ is equal to 0 or 1, $kl(P, \lambda_1)$ or $D(\lambda_1, \lambda_2)$ $\rightarrow$ $\infty$. So kl and SimMM measures become unsuitable. For such an application, we propose a similar statistical measure between two Markov models of order $r$ with the help of Jensen–Shannon divergence, denoted by $rre_k r(\lambda_1, \lambda_2)$, as follows:

$$rre_k r(\lambda_1, \lambda_2) = \sum_{k \in S_2} \left| \frac{P(O_{\lambda_1} | \lambda_1) \ln 2P(O_{\lambda_2} | \lambda_2) + P(O_{\lambda_1} | \lambda_1)}{P(O_{\lambda_1} | \lambda_2) + P(O_{\lambda_2} | \lambda_2)} \right|,$$  (15)

where the set $S_2$ consists of all possible sequences of length $k$ with symbol from the alphabet $\mathcal{A}$ of the context in DNA sequences, $\mathcal{A}$ is $\{A, C, G, T\}$. Moreover, $P(O_{\lambda_1} | \lambda_1)$ denotes the probability of sequence $O$ under the Markov model $\lambda$ of order $r$. From now on, we will restrict our discussion to Markov models with order from 0 to 2. Given Markov model $\lambda_1$, the probability of sequence $O = a_1 \cdots a_k$, $1 \leq k \leq N$, can be computed as follows:

$$P(O | \lambda_1) = P(a_1 | \lambda_1)P(a_2 | a_1, \lambda_1)\cdots P(a_k | a_{k-1}, \lambda_1)$$  (16)

The symmetric form of $rre_k r$, denoted by $S2.k.r(\lambda_1, \lambda_2)$, is defined by

$$S2.k.r(\lambda_1, \lambda_2) = \begin{cases} 0 & \text{if } \lambda_1 = \lambda_2 \\ \sum_{k \in S_2} \left( \frac{\ln 2P(O_{\lambda_2} | \lambda_2) + P(O_{\lambda_1} | \lambda_1)}{P(O_{\lambda_1} | \lambda_2) + P(O_{\lambda_2} | \lambda_2)} \right) & \text{if } \lambda_1 \neq \lambda_2 \end{cases}$$  (17)

2. Adding $k$-word distributions to Markov model: if $a_1 \cdots a_k$ is a given sequence of length $n$, $O = a_0 \cdots a_{n+k}$, $1 \leq i \leq n-k$, $1 \leq k \leq n$ is a $k$-word of the given sequence. Its distributions can be estimated by

$$W_0 = C_0 / (n-k+1),$$  (18)

where $C_0$ is the count of the $k$-word $O$ in the given sequence.

By adding the $k$-word distributions to Markov model, we present a novel statistical model in a compact form $\zeta = (\lambda, W)$. It contains the information from both $k$-word distributions and Markov model.

$$wre.k.r(\zeta_1, \zeta_2) = \sum_{k \in S_2} \left[ \frac{\ln 2P(O_{\zeta_1} | \zeta_1) + P(O_{\zeta_2} | \zeta_2)}{P(O_{\zeta_1} | \zeta_2) + P(O_{\zeta_2} | \zeta_2)} \right].$$  (19)

where the set $S_2$ consists of all possible sequences of length $k$ with symbol from the alphabet $\mathcal{A}$, $1 \leq k \leq \min[n_1, n_2]$. Moreover, $P(O_{\zeta_1} | \zeta_1)$ denotes the entry of sequence $\zeta$ under the statistical model $\zeta_1$, which can be computed by

$$P(O_{\zeta_1} | \zeta_1) = \pi_{1} \cdots \pi_{k}$$  (20)

Similarly, the symmetric form, denoted as $S2.k.r(\zeta_1, \zeta_2)$

$$S2.k.r(\zeta_1, \zeta_2) = \begin{cases} 0 & \text{if } \zeta_1 = \zeta_2 \\ \sum_{k \in S_2} \left( \frac{\ln 2P(O_{\zeta_2} | \zeta_2) + P(O_{\zeta_1} | \zeta_1)}{P(O_{\zeta_1} | \zeta_2) + P(O_{\zeta_2} | \zeta_2)} \right) & \text{if } \zeta_1 \neq \zeta_2 \end{cases}$$  (22)

where $|S_2|$ is the size of the set $S_2$.

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 the supplementary material, we prove that $S2.k.r$ satisfies the three conditions and, therefore, a valid distance metric.

2.3 Evaluation methods

Receiver operating curve (ROC) goes back to signal detection and classification problems and is now widely used (Egan, 1975). This approach is employed in binary classification of continuous data, usually categorized as positive (1) or negative (0) cases. The classification accuracy can be measured by plotting, for different threshold values, the number of true positives (TP), also named sensitivity or coverage versus false positives (FP), or (1-specificity), encountered for each threshold, properly normalized (Equation 23).

$$\text{sensitivity} = \frac{\text{TruePositives}}{\text{Positives}} = \frac{TP}{TP+FN};$$

$$\text{specificity} = \frac{\text{TrueNegatives}}{\text{Negatives}} = \frac{TN}{TN+FP};$$  (23)

$$1 - \text{specificity} = \frac{FP}{TN+FP}.$$  (24)

A ROC curve is simply the plot of sensitivity versus (1-specificity) for different threshold values. The area under a ROC curve (AUC) is a widely employed parameter to quantify the quality of a classifier because it is a threshold independent performance measure and is closely related to the Wilcoxon signed-rank test (Bradley, 1997). For a perfect classifier, the AUC is 1 and for a random classifier the AUC is 0.5. For additional results and comprehensive discussion on AUC measure, see Green and Bremner (2002).
3 RESULT AND DISCUSSION

3.1 Similarity search

The proposed statistical measure is used to search for similar sequences of a query sequence from a database of 39 library sequences, of which 20 sequences are known to be similar in biological function to the query sequence, and the remaining 19 sequences are known as being not similar in biological function to the query sequence. This dataset has been studied in Wu et al. (2001), Pham and Zuegg (2004). These 39 sequences were selected from mammals, viruses, plants, etc., of which lengths vary from 322 to 14121 bases. The query sequence is HSLIPAS (Human mRNA for lipoprotein lipase), which has 1612 bases. These sequences are known as being not similar in biological function to the query sequence, and the remaining 19 sequences of a query sequence from a database of 39 library sequences are considered as targets for similarity search. The ROC curves obtained from 2 to 8. For each measure, the best combination is chosen to represent that score in the performance. The ROC curves obtained from all the similarity measures. Among the similarity measures based on Markov model, D outperforms D2.8.0, D2.5.0, wre.2.1, S1.4.1 and S2.3.1. Among the similarity measures based on k-word distributions, pcc.3 is clearly more efficient than other measures. Among the similarity measures based on Markov model, D outperforms D2.8.0, D2.5.0, wre.2.1, S1.4.1 and S2.3.1. The main surprise of this analysis is that when we add the information on k-word distributions to Markov model in our way, wre.2.1 performs better than other similarity measures based on k-word distributions, Markov model or both. But the symmetrical form S2.3.1 has no significant improvement in similarity search. The inspection of the ROC curves themselves (Fig. 1) further illustrates this comparison between similarity measures.

3.2 Evaluation on functionally related regulatory sequences

Regulatory sequence comparison plays an important role in the ab initio discovery of cis-regulatory modules (CRMs) with a common function. If a set of co-regulated genes in a single species is given, we wish to find, in their upstream and downstream regions (henceforth called the ‘control regions’), the CRMs that mediate the common aspect of their expression profiles. The control regions may be tens of Kilobase long for each gene (especially for metazoan genomes), while the CRMs to be discovered are often only hundreds of base pair long. One must therefore search in the control regions for subsequences (the candidate CRMs) that share some functional similarity (Kantorovitz et al., 2007). The proposed statistical measures are further tested to evaluate if functionally or evolutionarily related sequence pairs are scored better than unrelated pairs of sequences randomly chosen from the genome. To assess the performance on functionally related sequences, we construct each dataset as follows. A set of CRMs, known to regulate expression in the same tissue, is taken as the ‘positive’ set. A set of equally many randomly chosen non-coding sequences, with lengths matching the CRMs, is taken as the ‘negative’ set. The following seven datasets are chosen to analyze: FLY BLASTODERM (82 CRMs with expression in the blastoderm-stage embryo of the fruitfly, Drosophila melanogaster);

![Fig. 1. ROC curves for similarity search dataset, ROC(a) denotes the comparison of the similarity measures based on alignment or k-word distributions. ROC(b) denotes the comparison of the similarity measures based on Markov model. Similarity measure names are presented in Section 2, with parameter values as suffix. A random classifier would generate equal proportions of FP and TP classifications, which corresponds to the ROC diagonal (dashed line).](https://academic.oup.com/bioinformatics/article-abstract/24/20/2296/259406)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Area</th>
<th>Tstd. Error</th>
<th>Asymptotic Sig.</th>
<th>Asymptotic 95% confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustal W</td>
<td>0.922</td>
<td>0.048</td>
<td>0.000</td>
<td>0.827</td>
</tr>
<tr>
<td>cos.2</td>
<td>0.900</td>
<td>0.053</td>
<td>0.000</td>
<td>0.796</td>
</tr>
<tr>
<td>eu.3</td>
<td>0.897</td>
<td>0.058</td>
<td>0.000</td>
<td>0.785</td>
</tr>
<tr>
<td>pcc.3</td>
<td>0.945</td>
<td>0.035</td>
<td>0.000</td>
<td>0.876</td>
</tr>
<tr>
<td>kld.4</td>
<td>0.900</td>
<td>0.051</td>
<td>0.000</td>
<td>0.800</td>
</tr>
<tr>
<td>D2.8.0</td>
<td>0.776</td>
<td>0.076</td>
<td>0.003</td>
<td>0.628</td>
</tr>
<tr>
<td>D2.5.0</td>
<td>0.929</td>
<td>0.040</td>
<td>0.000</td>
<td>0.851</td>
</tr>
<tr>
<td>D</td>
<td>0.938</td>
<td>0.046</td>
<td>0.000</td>
<td>0.848</td>
</tr>
<tr>
<td>wre.2.1</td>
<td>0.897</td>
<td>0.052</td>
<td>0.000</td>
<td>0.795</td>
</tr>
<tr>
<td>S1.4.1</td>
<td>0.918</td>
<td>0.047</td>
<td>0.000</td>
<td>0.827</td>
</tr>
<tr>
<td>S2.3.1</td>
<td>0.980</td>
<td>0.018</td>
<td>0.000</td>
<td>0.944</td>
</tr>
</tbody>
</table>

The bold values indicate that they are the highest ones in each experiment.
FLY PNS [23 CRMs (average length 998 bp) driving expression in the peripheral nervous system in the fruitfly]; FLY TRACHEAL [9 CRMs (average length 1220 bp) involved in regulation of the tracheal system in the fruitfly]; FLY EYE [17 CRMs (average length 894 bp) expressing in the Drosophila eye]; HUMAN MUSCLE [28 human CRMs (average length 450) regulating muscle specific gene expression]; HUMAN LIVER [9 CRMs (average length 201) driving expression specific to the human liver]; HUMAN HBB [17 CRMs (average length 453) regulating the HBB complex]. They are well studied by Gallo et al. (2006) and Kantorovitz et al. (2007).

Each pair of sequences in the positive set is compared, and so is each pair in the negative set. The evaluation procedure is based on a binary classification of each sequence pair, where 1 corresponds to the pairs from positive set, 0 corresponds to the pairs from negative set. Let $n$ be the number of sequences in the positive set, all the pairs constitute a vector of length $2\binom{n}{2}$, which is used as prediction. Also, we can get a vector of length $2\binom{k}{2}$ consisting of 1 and 0 as class labels. A perfect measure would completely separate negative from positive set. Of course, this does not happen in practice, and the classes are interspersed. The ROC curves permit to assess the level of accuracy of this separation without choosing any distance threshold for the separation point. In particular, the AUC will give us a unique number of the relative accuracy of each measure.

The similarity measures evaluated here should satisfy the symmetry condition. The evaluated measures are: the similarity measures based on alignment, $S_{1.3.2}$, $S_{2.8.2}$ and the seven measures outlined in Section 2, where the similarity measures based on alignment are Needleman–Wunsch (global alignment) or Smith–Waterman (local alignment) raw scores, with no correction for statistical significance, using linear gap penalties or affine gap penalties, with a gap penalty of 2. All statistical measures based on $k$-word distributions run with $k$-word from 2 to 8. The $D_{2.8.2}$, $D_{2.2.1}$, $S_{1.3.2}$ and $S_{2.8.2}$ scores run with Markov order $r$ from 0 to 2 and word length $k$ from 2 to 8. For each measures, separate tests are done with each combination of parameter values, and the best combination is chosen to represent that score in the performance. ROC curves are computed to evaluate and compare the performances of our measures and other measures. The ROC curves obtained for PNS and EYE are presented in Figure 2. All the color figures for seven datasets are presented in the Supplementary Material.

Table 2 summarizes the AUCs obtained from all the similarity measures for seven datasets. In the BLASTODERM experiment, $S_{2.7.2}$ performs better than other alignment-based or alignment-free measures, with the AUC 0.664. The next best measure is $D_{4}$, and the other measures lag behind. In the PNS experiment, $S_{2.8.2}$ measure is significantly better than all other scores, its AUC is 0.854. The AUCs of all other measures are below 0.7. In the Fly TRACHEAL experiment, $S_{2.8.2}$ significantly outperforms other methods, and its AUC is 0.975. It is followed by $D_{4}$ and $D_{2.7.4}$. This is a highly significant result demonstrating that the $S_{2.8.2}$ measure is successful at detecting the functional similarity of tracheal CRMs. In the EYE experiment, the AUC of $S_{2.8.2}$ is 0.833, significantly better than other statistical measures. The next best measures are $D_{4}$ and $D_{2.7.4}$. In the MUSCLE experiment, $D_{2.2.0}$ outperforms other methods, and its AUC is 0.703. It is followed by $S_{2.8.0}$. In LIVER experiments, $S_{2.5.1}$ performs significantly better than other measures. The next best measure is the $D_{2.2.0}$. In HBB experiments, $S_{2.2.0}$ achieves the best performance, followed by $D_{2.7.3.0}$. From the seven experiments, we can see that $S_{2.8.2}$ performs significantly better than other measures among six experiments, because it incorporates the $k$-word information into Markov model directly. The inspection of the ROC curves themselves (in the Supplementary Material) further illustrates these comparisons among similarity measures. Since the different CRMs are only functionally related and not orthologous, the CRM search algorithm requires a method that can discern functional similarity among candidate CRMs based on their sequence similarity. From Table 2, we can note that the alignment-based methods lag behind some alignment-free methods.

### 3.3 Phylogenetic analysis

Since the proposed measure $S_{2.8.2}$ is a statistical distance measure, it can be further tested by phylogenetic analysis. Given a set of DNA, their phylogenetic relationship can be obtained through the following main operations: first, we construct the statistical models for biological sequences and calculate their similarity distance by using our statistical distance measure $S_{2.8.2}$; second, by arranging all the similarity distance into a matrix, we obtain a pairwise distance matrix and finally, we put the pairwise distance matrix into the neighbor-joining program in the PHYLIP package (Felsenstein, 1989).

The phylogeny of eutherian orders has been unresolved due to conflicting results obtained from comparison of whole mtDNA sequences and individual proteins encoded by mtDNA (Cao et al., 1998; Lin et al., 2002; Waddell et al., 1999a, b, 2001). We choose a more controversial dataset (29 mammalian species, of which five are rodents deals) that have been widely studied in Reyes et al. (2000), Li et al. (2001) and Otu and Sayaood (2003). These sequences are described in the Supplementary Material.

The phylogenetic tree constructed by our measure is presented in Figure 3. The bootstrap technique is employed to evaluate the tree topologies by resampling the sequence 100 times. We obtain the phylogenetic trees drawn by MEGA program (Kumar et al., 2004); bootstrap values, lower than 50, are hidden. Generally, an independent method can be developed to evaluate the accuracy of a phylogenetic tree, or the validity of a phylogenetic tree can be tested.
Table 2. Comparison of AUCs obtained from all the similarity measures for seven datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>Blastoderm AUC</th>
<th>Method</th>
<th>Pns AUC</th>
<th>Method</th>
<th>Tracheal AUC</th>
<th>Method</th>
<th>Eye AUC</th>
<th>Method</th>
<th>Muscle AUC</th>
<th>Method</th>
<th>Liver AUC</th>
<th>Method</th>
<th>HBB AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NW-linear</td>
<td>0.505</td>
<td>NW-linear</td>
<td>0.503</td>
<td>NW-linear</td>
<td>0.553</td>
<td>NW-linear</td>
<td>0.508</td>
<td>NW-linear</td>
<td>0.506</td>
<td>NW-linear</td>
<td>0.442</td>
<td>NW-linear</td>
<td>0.534</td>
</tr>
<tr>
<td>NW-affine</td>
<td>0.526</td>
<td>NW-affine</td>
<td>0.490</td>
<td>NW-affine</td>
<td>0.671</td>
<td>NW-affine</td>
<td>0.540</td>
<td>NW-affine</td>
<td>0.561</td>
<td>NW-affine</td>
<td>0.477</td>
<td>NW-affine</td>
<td>0.701</td>
</tr>
<tr>
<td>SW-linear</td>
<td>0.417</td>
<td>SW-linear</td>
<td>0.450</td>
<td>SW-linear</td>
<td>0.239</td>
<td>SW-linear</td>
<td>0.448</td>
<td>SW-linear</td>
<td>0.509</td>
<td>SW-linear</td>
<td>0.501</td>
<td>SW-linear</td>
<td>0.489</td>
</tr>
<tr>
<td>SW-affine</td>
<td>0.417</td>
<td>SW-affine</td>
<td>0.450</td>
<td>SW-affine</td>
<td>0.239</td>
<td>SW-affine</td>
<td>0.448</td>
<td>SW-affine</td>
<td>0.609</td>
<td>SW-affine</td>
<td>0.501</td>
<td>SW-linear</td>
<td>0.498</td>
</tr>
<tr>
<td>cos.2</td>
<td>0.547</td>
<td>cos.7</td>
<td>0.547</td>
<td>cos.2</td>
<td>0.750</td>
<td>cos.3</td>
<td>0.573</td>
<td>cos.8</td>
<td>0.647</td>
<td>cos.2</td>
<td>0.659</td>
<td>cos.2</td>
<td>0.710</td>
</tr>
<tr>
<td>eu.2</td>
<td>0.560</td>
<td>eu.6</td>
<td>0.601</td>
<td>eu.2</td>
<td>0.755</td>
<td>eu.3</td>
<td>0.582</td>
<td>eu.2</td>
<td>0.550</td>
<td>eu.2</td>
<td>0.695</td>
<td>eu.2</td>
<td>0.710</td>
</tr>
<tr>
<td>pcc.2</td>
<td>0.529</td>
<td>pcc.7</td>
<td>0.542</td>
<td>pcc.2</td>
<td>0.746</td>
<td>pcc.3</td>
<td>0.564</td>
<td>pcc.7</td>
<td>0.646</td>
<td>pcc.2</td>
<td>0.636</td>
<td>pcc.2</td>
<td>0.709</td>
</tr>
<tr>
<td>kld.2</td>
<td>0.543</td>
<td>kld.6</td>
<td>0.527</td>
<td>kld.2</td>
<td>0.818</td>
<td>kld.2</td>
<td>0.573</td>
<td>kld.2</td>
<td>0.547</td>
<td>kld.2</td>
<td>0.725</td>
<td>kld.2</td>
<td>0.711</td>
</tr>
<tr>
<td>D2.2.1</td>
<td>0.494</td>
<td>D2.1</td>
<td>0.495</td>
<td>D2.2.1</td>
<td>0.509</td>
<td>D2.8.0</td>
<td>0.500</td>
<td>D2.8.2</td>
<td>0.671</td>
<td>D2.7.1</td>
<td>0.641</td>
<td>D2.5.1</td>
<td>0.560</td>
</tr>
<tr>
<td>D2z.2.0</td>
<td>0.552</td>
<td>D2z.6.2</td>
<td>0.641</td>
<td>D2z.4.0</td>
<td>0.829</td>
<td>D2z.4.0</td>
<td>0.626</td>
<td>D2z.2.0</td>
<td>0.703</td>
<td>D2z.2.0</td>
<td>0.835</td>
<td>D2z.3.0</td>
<td>0.742</td>
</tr>
<tr>
<td>D</td>
<td>0.579</td>
<td>D</td>
<td>0.605</td>
<td>D</td>
<td>0.833</td>
<td>D</td>
<td>0.640</td>
<td>D</td>
<td>0.528</td>
<td>D</td>
<td>0.698</td>
<td>D</td>
<td>0.738</td>
</tr>
<tr>
<td>S1.3.2</td>
<td>0.557</td>
<td>S1.3.2</td>
<td>0.559</td>
<td>S1.3.2</td>
<td>0.807</td>
<td>S1.3.2</td>
<td>0.609</td>
<td>S1.3.2</td>
<td>0.566</td>
<td>S1.8.1</td>
<td>0.799</td>
<td>S1.3.2</td>
<td>0.739</td>
</tr>
<tr>
<td>S2.7.2</td>
<td>0.664</td>
<td>S2.8.2</td>
<td>0.854</td>
<td>S2.8.2</td>
<td>0.975</td>
<td>S2.8.2</td>
<td>0.833</td>
<td>S2.8.0</td>
<td>0.672</td>
<td>S2.5.1</td>
<td>0.880</td>
<td>S2.2.0</td>
<td>0.785</td>
</tr>
</tbody>
</table>

The bold values indicate that they are the highest ones in each experiment.

Fig. 3. Phylogenetic tree obtained by our measure using complete mtDNA. Bootstraps are based on 100 replications.

by comparing it with authoritative ones. Here, we adopt the latter one to test the validity of our phylogenetic tree.

It has been debated which two of the three main groups of placental mammals are more closely related: Primate, Ferungulates and Rodents. This is because by the maximum likelihood method, some protein support the (Ferungulates, (Primates, Rodents)) grouping, while other proteins support the (Rodents, (Ferungulates, Primates)) grouping (Cao et al., 1998). Our method has reconfirmed the hypothesis of (Rodents, (Primates, Ferungulates)). Figure 3 shows that our tree is quite consistent with the authoritative ones (Cao et al., 1998; Li et al., 2001; Otu and Sayood, 2003; Reyes et al., 2000) in the following four aspects: first of all, marsupials and monotremes are grouped closely, the bootstrap values (98% and 100%) are higher than Cao et al. (1998) (89% and 99%) and Reyes et al. (2000) (77% and 90%) and finally, ferungulates are grouped in a branch and clustered with primates with a high bootstrap value 98% [it is better than Cao et al. (1998) (94%), Reyes et al. (2000) (65%) and Li et al. (2001) (89%)]. These results confirm that S2.k.r can be considered as another efficient distance measure for phylogenetic analysis.

4 CONCLUSIONS

Many bioinformatics applications rely heavily on sequence comparison techniques, from searching a database with a query DNA sequence to the phylogenetic tree construction. Despite the prevalence of alignment-based methods, it is also noteworthy that alignment-based method is computationally intensive and consequently impractical for querying large datasets, which forces the use of some heuristics to reduce the running times, as exemplified by BLAST. Alignment-free comparison method is therefore of great value as it reduces the technical constraints of alignments.

The statistical measures wr.k.r and S2.k.r presented in this work are alignment-free methods for sequence comparison. By incorporating the k-word distributions into Markov model, we construct a novel statistical model $\zeta$ for each sequence. The similarity between two sequences can be obtained by computing the log-likelihood difference between their corresponding statistical models. Therefore, they are simple alignment-free methods that yield results reasonably. The test of our methods are to perform similarity search and evaluate the functionally related regulatory sequences. Our statistical measures are evaluated by comparison with methods based on alignment or not. The comparison demonstrates that our measures intending to incorporate k-word distributions into Markov model are more efficient (Tables 1 and 2). We also demonstrate that S2.k.r can be used as a novel distance measure to construct phylogenetic tree, and performs as well as the Kolmogorov complexity and LZ complexity.

Overall our results highlight the necessity for alignment-free measures to extract more information as possible. Thus, this understanding can then be used to guide development of more powerful measures for sequence comparison with future possible improvement on evolutionary study, structure and function prediction.
ACKNOWLEDGEMENTS

The authors thank all the anonymous referees for their valuable suggestions and support. In particular, the authors thank Prof. Tuan D. Pham for providing all MATLAB code for SimMM and Prof. Saurabh Sinha for providing the software of D2z, datasets and the technical help. The authors also thank Dr Weiping Wang for his kindly help.

Funding: National Natural Science Foundation of China (10571019).

Conflict of Interest: none declared.

REFERENCES