BIOINFORMATICS

Vol. 20 no. 18 2004, pages 3455–3461
doi:10.1093/bioinformatics/bth426

A probabilistic measure for alignment-free sequence comparison

Tuan D. Pham[1,*] and Johannes Zuegg[2]

[1]School of Computing and Information Technology, Griffith University, Nathan Campus, QLD 4111, Australia and [2]Alchemia Ltd, PO Box 6242, Upper Mount Gravatt, QLD 4122, Australia

Received on March 1, 2004; revised on June 28, 2004; accepted on July 26, 2004
Advance Access publication July 22, 2004

ABSTRACT

Motivation: Alignment-free sequence comparison methods are still in the early stages of development compared to those of alignment-based sequence analysis. In this paper, we introduce a probabilistic measure of similarity between two biological sequences without alignment. The method is based on the concept of comparing the similarity/dissimilarity between two constructed Markov models.

Results: The method was tested against six DNA sequences, which are the *thrA*, *thrB* and *thrC* genes of the threonine operons from *Escherichia coli* K-12 and from *Shigella flexneri*; and one random sequence having the same base composition as *thrA* from *E.coli*. These results were compared with those obtained from CLUSTAL W algorithm (alignment-based) and the chaos game representation (alignment-free). The method was further tested against a more complex set of 40 DNA sequences and compared with other existing sequence similarity measures (alignment-free).

Availability: All datasets and computer codes written in MATLAB are available upon request from the first author.

Contact: t.pham@griffith.edu.au

INTRODUCTION

There have been a number of computational and statistical methods for the comparison of biological sequences developed over the past decade. It still remains a challenging problem for the research community of computational biology (Ewens and Grant, 2001; Miller, 2001). Two distinct bioinformatic methodologies for studying the similarity/dissimilarity of sequences are known as alignment-based and alignment-free methods. The search for optimal solutions using sequence alignment-based methods is encountered with difficulty in computational aspect with regard to large biological databases. Therefore, the emergence of research into alignment-free sequence analysis is apparent and necessary to overcome critical limitations of sequence analysis by alignment. Till date, alignment-free sequence analysis is still in its early development with regard to alignment-based sequence comparison. One of the most recent review (Vinga and Almeida, 2003) on published methods for alignment-free sequence comparison of biological sequences reports several concepts of distance measures, such as the Euclidean distance (Blaisdell, 1986), Euclidean and Mahalanobis distances (Wu *et al*., 1997), Markov chain models and Kullback–Leibler discrepancy (KLD) (Wu *et al*., 2001), cosine distance (Stuart *et al*., 2002), Kolmogorov complexity (Li *et al*., 2001) and chaos theory (Almeida *et al*., 2001). Our present work exhibit some strong similarity to the work by Wu *et al*. (2001), in which statistical measures of DNA sequence dissimilarity as the Mahalanobis distance and the standardized Euclidean distance under Markov chain model of base composition, as well as the extended KLD. The KLD extended by Wu *et al*. (2001) was computed in terms of two vectors of relative frequencies of n-words over a sliding window from two given DNA sequences. Whereas, our work presented here derives a probabilistic distance between two sequences using a symmetrized version of the KLD, which directly compares two Markov models built for the two corresponding biological sequences.

Thus, among alignment-free methods for computing distances between biological sequences, there seems rarely any work that directly computes distances between biological sequences using Markov-chain-based measures. If a Markov model can be constructed for each sequence, we can measure the similarity between any two sequences by computing the log-likelihood difference between two Markov models with the same observation data. We have tested the proposed method, which we call SimMM (SIMilarity of Markov Models), with six DNA sequences and one randomly generated sequence and then compared the results with other methods including CLUSTAL W algorithm (Thompson *et al*., 1994) and the chaos game representation (Almeida *et al*., 2001). Our results using SimMM were in good agreement with results obtained from the above two methods, but showed better discrimination to a random sequence. We have also further tested our method against a more complex set of 40 DNA sequences, and the results obtained from our proposed

*To whom correspondence should be addressed.

method were found to be more favorable than those obtained from other distance measures (Wu *et al*., 2001).

SIMILARITY MEASURE BY COMPARING MARKOV MODELS

Let $A = [a_{ij}]$ denote the state transition probability matrix of a discrete Markov process. Each state transition probability a_{ij} is defined as:

$$a_{ij} = P[q_{t_n} = S_j | q_{t_{n-1}} = S_i], \quad 1 \le i, \ j \le N, \tag{1}$$

where q_{t_n} stands for the actual state at time t_n ($n = 1, 2, \ldots$), S_j a state j of a set of N distinct states. In the context of DNA sequences, the number of states $N = 4$, which correspond to the four nucleotide symbols {a, c, g, t}. The state transition probabilities are subject to

$$a_{ij} \ge 0 \quad \forall\, i, j, \tag{2}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \quad \forall\, i. \tag{3}$$

Also, let $\pi = \{\pi_i\}$ be the initial state transition distribution where

$$\pi_i = P(q_{t_1} = S_i), \quad 1 \le i \le N. \tag{4}$$

This Markov chain involves two probabilistic measures A and π, that can be denoted in a compact form as:

$$\lambda = (A, \pi). \tag{5}$$

What we have presented above is the first-order Markov model, the higher-order Markov models can be determined as follows. Let $\{\pi_j^{(n=1)}\}$ ($j = 1, 2, \ldots, N$) be the absolute (initial) probabilities that the system is in state S_j at t_1. Given the parameters $\{\pi_j^{(1)}\}$ and A of a Markov chain, the absolute probabilities after a specified number of transitions (n-order) are determined as follows:

$$\pi_j^{(n)} = \sum_i \pi_i^{(1)} a_{ij}^{(n)}, \tag{6}$$

where $a_{ij}^{(n)}$ is the n-step or n-order transition probability given by the recursive formula

$$a_{ij}^{(n)} = \sum_k a_{ik}^{(n-m)} \, a_{kj}^{(m)}, \quad 0 < m < n. \tag{7}$$

Expressions (6) and (7) are known as Chapman–Kolomogorov equations (Taha, 1982). Thus, for $n = 2, 3, 4$ we have the first-order, second-order, third-order Markov models, respectively and so on. We can see that the second-order Markov model over the four nucleotide symbols {a, c, g, t} is equivalent to the first-order Markov model with 16 states of the dinucleotides: aa, ac, ag, at, ca, cc, cg, ct, ga, gc, gg, gt, ta, tc, tg and tt. From now on, we will restrict our discussion to the context of the first-order Markov model.

Let $\lambda_1 = (A_1, \pi_1)$ and $\lambda_2 = (A_2, \pi_2)$ be two Markov models of the two bio-sequences, where each model is constructed by the observed symbols of each corresponding DNA sequence. Our interest is to find a similarity or dissimilarity measure between two Markov models λ_1 and λ_2. A well-known dissimilarity measure between two probability distributions is the Kullback–Leibler divergence (Cover and Thomas, 1991).

Let P_1 and P_2 be two probability distributions on a universe X, the Kullback-Leibler divergence (KLD) or the relative entropy, denoted as $H(P_1, P_2)$, of P_1 with respect to P_2 is defined by the Lebesgue integral (Kroupa, 2003).

$$H(P_1, P_2) = \int_X \frac{dP_1}{dP_2} \log \frac{dP_1}{dP_2} \, dP_2. \tag{8}$$

Expression (8) is equivalent to

$$H(P_1, P_2) = \int_X \log \frac{dP_1}{dP_2} \, dP_1. \tag{9}$$

The discrete version of the KLD defined in (8) is

$$H(p_1, p_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)}. \tag{10}$$

The integral (8) exists provided that $P_1 \ll P_2$. Although $H(p_1, p_2)$ is often called a distance, it is not a metric because $H(p_1, p_2) \neq H(p_2, p_1)$. Moreover, $H(p_1, p_2) = 0$ iff $p_1 = p_2$.

Given two Markov models, we now can define a probabilistic distance between two sequences, denoted by $d(\lambda_1, \lambda_2)$, as:

$$d(\lambda_1, \lambda_2) = 1 - \exp[D_s(\lambda_1, \lambda_2)], \tag{11}$$

where D_s is the symmetrized version of the approximate KLD divergence of λ_1 and λ_2, which is expressed as:

$$D_s(\lambda_1, \lambda_2) = \frac{D(\lambda_1, \lambda_2) + D(\lambda_2, \lambda_1)}{2}, \tag{12}$$

where $D(\lambda_1, \lambda_2)$ is the empirical KLD between λ_1 and λ_2, which was originally introduced by Juang and Rabiner (1985) using the Monte Carlo simulations. The models are assumed to be ergodic, having arbitrary observation probability distributions; and the dissimilarity is defined as the mean divergence of the observation sample. This approximate KLD is given by

$$D(\lambda_1, \lambda_2) = \frac{1}{T_2} \log \frac{P(O_{\lambda_2} | \lambda_1)}{P(O_{\lambda_2} | \lambda_2)}, \tag{13}$$

where $O_{\lambda_2} = (o_1 o_2 \cdots o_{T_2})$ is a sequence of observations generated by model λ_2, and T_2 is the length of sequence O_{λ_2}.

It can be interpreted that expression (13) implies how well model λ_1 scores the observation sequence that is used to construct model λ_2, relative to how well model λ_2 scores the

observations used to construct itself (Rabiner, 1989; Rabiner and Juang, 1993).

Because $D(\lambda_1, \lambda_2)$ and $D(\lambda_2, \lambda_1)$ are not symmetrical, we can define $D(\lambda_2, \lambda_1)$ as:

$$D(\lambda_2, \lambda_1) = \frac{1}{T_1} \log \frac{P(O_{\lambda_1}|\lambda_2)}{P(O_{\lambda_1}|\lambda_1)}, \qquad (14)$$

where $O_{\lambda_1} = (o_1 o_2 \cdots o_{T_1})$ is a sequence of observed symbols generated by model λ_1.

Finally, the probability of the observation sequence $O = \{o_t,\ t = 1, \ldots, T\}$, given a Markov model λ can be evaluated as:

$$\begin{aligned} P(O|\lambda) &= P(o_1, \ldots, o_T|\lambda) \\ &= P(o_1) \cdots P(o_t|o_{t+1}) \cdots P(o_{T-1}|o_T) \\ &= \pi_{(q_{t_1}=o_1)} \cdots a_{q_{t_n}, q_{t_{n+1}}} \cdots a_{q_{t_{T-1}}, q_{t_T}}. \end{aligned} \qquad (15)$$

RESULTS

Experiment no. 1

The algorithm was tested with six DNA sequences, taken from the threonine operons of *Escherichia coli* K-12 (gi:1786181) and *Shigella flexneri* (gi:30039813). The three sequences taken from each threonine operon are *thrA* (aspartokinase I-homoserine dehydrogenase I), *thrB* (homoserine kinase) and *thrC* (threonine synthase), using the open reading frames (ORFs) 337–2799 (*ec*-thrA), 2801–3733 (*ec*-thrB) and 3734–5020 (*ec*-thrC) in the case of *E.coli* K-12, and 336–2798 (*sf*-thrA), 2800–3732 (*sf*-thrB) and 3733–5019 (*sf*-thrC) in the case of *S.flexneri*. All the sequences were obtained from GenBank (www.ncbi.nlm.nih.gov/Entrez). In addition, we compared all six sequences with a randomly generated sequence (*rand*-thrA), using the same length and base composition as *ec*-thrA.

The probabilistic distances among the seven sequences obtained using SimMM are shown in Table 1. To compare SimMM with other methods, we calculated the sequence similarity or sequence distance using alignment-based methods. All seven sequences have been aligned using CLUSTAL W (Thompson *et al*., 1994). The multiple sequence alignment has then been used to calculate an identity matrix, which is represented in Table 2 as a distance (converted identity) matrix; and the distance matrix, shown in Table 3, using DNADist from the PHYLIP package (Felsenstein, 1993) and the modification of the Kimura distance model (Kimura, 1980). The DNADist program uses nucleotide sequences to compute a distance matrix, under the modified Kimura model of nucleotide substitution. Being similiar to the Jukes and Cantor (1969) model, which constructs the transition probability matrix based on the assumption that a base change is independent of its identity, the Kimura '2-paramter' model allows for a difference between transition and transversion rates in the construction of the DNA distance matrix. Figures 1–3 show the phylogenetic trees of SimMM, CLUSTAL W using the identity matrix and CLUSTAL W using the distance matrix, respectively, which were plotted using the KITSCH program in the PHYLIP package. The KITSCH program, based on the Fitch–Margoliash and least-squares methods with an evolutionary clock, deals with matrices of pairwise distances between all pairs of taxa. The method may be considered as providing an estimate of the phylogeny, using a phenetic clustering of the tip species. By minimizing an objective function, this method not only sets the levels of the clusters, but also rearranges the hierarchy of the clusters in order to find alternative clusterings that give a lower global sum of squares of differences between the observed distance matrix and the expected one (http://www.cmbi.kun.nl/bioinf/PHYLIP/kitsch-1.html).

Table 1. Probabilistic distance matrix (symmetric) using SimMM

	ec-thrA	*ec*-thrB	*ec*-thrC	*sf*-thrA	*sf*-thrB	*sf*-thrC	*rand*-thrA
ec-thrA	0	0.0074	0.0053	0.0002	0.0072	0.0054	0.0272
ec-thrB		0	0.0103	0.0074	0.0004	0.0109	0.0331
ec-thrC			0	0.0059	0.0110	0.0001	0.0401
sf-thrA				0	0.0073	0.0058	0.0208
sf-thrB					0	0.0116	0.0214
sf-thrC						0	0.0318
rand-thrA							0

Table 2. Distance (converted identity) matrix (symmetric) using CLUSTAL W alignment

	ec-thrA	*ec*-thrB	*ec*-thrC	*sf*-thrA	*sf*-thrB	*sf*-thrC	*rand*-thrA
ec-thrA	0	0.8490	0.8030	0.0250	0.8510	0.8040	0.7140
ec-thrB		0	0.6710	0.8450	0.0170	0.6740	0.8620
ec-thrC			0	0.8000	0.6690	0.0090	0.8050
sf-thrA				0	0.8470	0.8020	0.7100
sf-thrB					0	0.6710	0.8620
sf-thrC						0	0.8070
rand-thrA							0

Table 3. Distance matrix (symmetric) calculated by DNADist using Kimura model and CLUSTAL W alignment

	ec-thrA	*ec*-thrB	*ec*-thrC	*sf*-thrA	*sf*-thrB	*sf*-thrC	*rand*-thrA
ec-thrA	0	1.5246	1.5866	0.0133	1.5691	1.6123	2.9377
ec-thrB		0	1.0938	1.5007	0.0163	1.1065	1.5711
ec-thrC			0	1.5982	1.0833	0.0086	1.6227
sf-thrA				0	1.5447	1.6243	2.9317
sf-thrB					0	1.0960	1.5853
sf-thrC						0	1.6330
rand-thrA							0

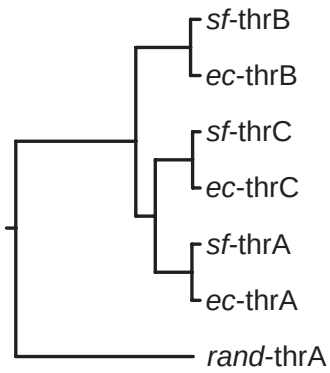

Fig. 1. Phylogenetic tree obtained from probabilistic distance matrix using SimMM, plotted by KITSCH program of PHYLIP package using Fitch–Margoliash criterion and evolutionary clock.

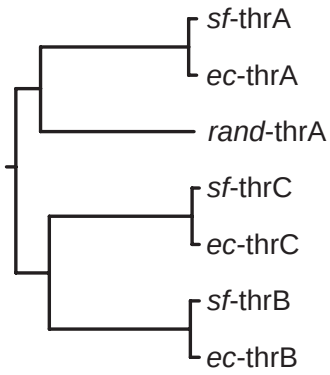

Fig. 2. Phylogenetic tree obtained from distance (converted identity) matrix using CLUSTAL W, plotted by KITSCH program of PHYLIP package using Fitch–Margoliash criterion and evolutionary clock.

The results obtained using SimMM agree with those obtained using the chaos game representation (Almeida *et al*., 2001) even though we used seven sequences as test sets. In the chaos game method the sequence *ec*-thrA is closer to *ec*-thrC than to *ec*-thrB, and *ec*-thrB is closer to *ec*-thrA than to *ec*-thrC. Using the probabilistic similarity values calculated by our SimMM, we obtained the same relationships, of thrA being closer to thrC, and thrB being closer to thrA. This relationship was found within both species, *E.coli* K-12 (gi:1786181) and *S.flexneri*. We need to point out that this agreement between the two models does not confirm any hypothesis about the relationships of these threonine operons since we have found no current phylogenetic study of these threonine operons in the literature. The alignment-based methods, on the other hand, show a slightly different relationship between the three different sequences. The calculations from both the identity and distance matrices place the thrA sequences closer to thrB than to thrC, and thrB closer to thrC than to thrA (Tables 2 and 3). The phylogenetic trees illustrate the difference between SimMM and the alignment-based

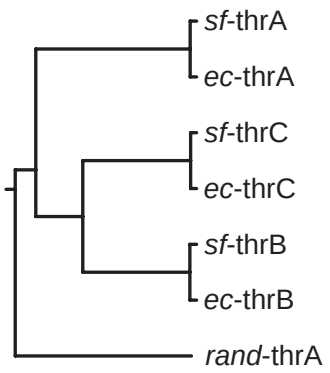

Fig. 3. Phylogenetic tree obtained from distance matrix (DNADist) using CLUSTAL W and Kimura model, plotted by KITSCH program of PHYLIP package using Fitch–Margoliash criterion and evolutionary clock.

DNADist. Using SimMM, sequences thrC and thrA form a subcluster; whereas using DNADist, the subcluster is formed by thrC and thrB.

Another difference between SimMM and DNADist can also be illustrated by the phylogenetic trees. All the trees were drawn to an equivalent overall size, and based on a relative scale it can be observed that all the real sequences appear to be less related to each other in the DNADist tree (Fig. 3) than in the tree using the new method (Fig. 1). In other words, all real sequences are more closely clustered with the new method than the alignment-based DNADist method. This result suggests that the SimMM method is better at distinguishing a randomized sequence from a group of related natural sequences as a whole. This can be explained in that DNA and protein sequences have been realized to comprise a mixture of local regions with distinct genetic functions and evolutionary origins, i.e. DNA and protein sequences do not represent strings of random symbols. These local regions rather consist of compositional characteristics and pseudo-periodic sequence patterns (Li *et al*., 2004). Our proposed method, based on the Markov chain, takes into account this 'periodical' behavior of the biosignal by both using the state transition probability distribution, and the probabilistic distance measure defined in (11). Whereas, this stochastic information content is absent from the solution provided by the multiple sequence alignment. The DNADist considers the transition probability of the distance matrix which is still derived from the multiple sequence alignment.

Experiment no. 2

The proposed KLD of Markov models was further 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. These 39 sequences

were selected from mammals, viruses, plants, etc., of which lengths vary between 322 and 14 121 bases. All of these sequences can be obtained from the GenBank sequence database (http://www.ncbi.nlm.nih.gov/Entrez/). The query sequence is HSLIPAS (Human mRNA for lipoprotein lipase), which has 1612 bases.

The 20 sequences, which are known as being similar in biological function to HSLIPAS are as follows: OOLPLIP (*Oestrus ovis* mRNA for lipoprotein lipase, 1656 bp), SSLPLRNA (pig back fat *Sus scrofa* cDNA similar to *S.scrofa* LPL mRNA for lipoprotein lipase, 2963 bp), RATLLIPA (*Rattus norvegicus* lipoprotein lipase mRNA, complete cds, 3617 bp), MUSLIPLIP (*Mus musculus* lipoprotein lipase gene, partial cds, 3806 bp), GPILPPL (guinea pig lipoprotein lipase mRNA, complete cds, 1744 bp), GGLPL (chicken mRNA for adipose lipoprotein lipase, 2328 bp), HSHTGL (human mRNA for hepatic triglyceride lipase, 1603 bp), HUMLIPH (human hepatic lipase mRNA, complete cds, 1550 bp), HUMLIPH06 (human hepatic lipase gene, exon 6, 322 bp), RATHLP (rat hepatic lipase mRNA, 1639 bp), RABTRIL [*Oryctolagus cuniculus* (clone TGL-5K) triglyceride lipase mRNA, complete cds, 1444 bp], ECPL (*Equus caballus* mRNA for pancreatic lipase, 1443 bp), DOGPLIP (canine lipase mRNA, complete cds, 1493 bp), DMYOLK [*Drosophila* gene for yolk protein I (vitellogenin), 1723 bp], BOVLDLR [bovine low-density lipoprotein (LDL) receptor mRNA, 879 bp], HSBMHSP (*Homo sapiens* mRNA for basement membrane heparan sulfate proteoglycan, 13 790 bp), HUMAPOAICI (human apolipoprotein A-I and C-III genes, complete cds, 8966 bp), RABVLDLR (*O.cuniculus* mRNA for very LDL receptor, complete cds, 3209 bp), HSLDL100 (human mRNA for apolipoprotein B-100, 14 121 bp) and HUMAPOBF (human apolipoprotein B-100 mRNA, complete cds, 10 089 bp).

The other 19 sequences known as being not similar in biological function to HSLIPAS are as follows: A1MVRNA2 [alfalfa mosaic virus (A1M4) RNA 2, 2593 bp], AAHAV33A [*Acanthocheilonema viteae* pepsin-inhibitor-like-protein (Av33) mRNA sequence, 1048 bp], AA2CG (adeno-associated virus 2, complete genome, 4675 bp), ACVPBD64 (artificial cloning vector plasmid BD64, 4780 bp), AL3HP (bacteriophage alpha-3 H protein gene, complete cds, 1786 bp), AAABDA [*Aedes aegypti* abd-A gene for abdominal-A protein homolog (partial), 1759 bp], BACBDGALA [*Bacillus circulans* beta-D-galactosidase (bgaA) gene, complete cds, 2555 bp], BBCA (*Bos taurus* mRNA for cyclin A, 1512 bp), BCP1 (bacteriophage Chp1 genome DNA, complete sequence, 4877 bp) and CHIBATPB (sweet potato chloroplast F1-ATPase beta and epsilon-subunit genes, 2007 bp), A7NIFH (Anabaena 7120 nifH gene, complete CDS, 1271 bp), AA16S (*Amycolatopsis azurea* 16S rRNA, 1300 bp), ABGACT2 (*Absidia glauca* actin mRNA, complete cds, 1309 bp), ACTIBETLC (Actinomadura R39 DNA for beta-lactamase gene, 1902 bp), AMTUGSNRNA (*Ambystoma mexicanum* AmU1 snRNA gene, complete sequence, 1027 bp), ARAST18B (cloning vector pAST 18b for *Caenorhabditis elegans*, 3052 bp), GCALIP2 (*Geotrichum candidum* mRNA for lipase II precursor, partial cds, 1767 bp), AGGGLINE (*Ateles geoffroyi* gamma-globin gene and L1 LINE element, 7360 bp) and HUMCAN (*H.sapiens* CaN19 mRNA sequence, 427 bp).

Sensitivity and selectivity were computed to evaluate and compare the performance of the proposed KLD of Markov models with other distance measures studied by Wu *et al*. (2001). Sensitivity is expressed by the number of HSLIPAS-related sequences found among the first closest 20 library sequences; whereas selectivity is expressed in terms of the number of HSLIPAS-related sequences of which distances are closer to HSLIPAS than others and are not truncated by the first HSLIPAS-unrelated sequence. Among several distance measures introduced by Wu *et al*. (2001), they concluded that the standardized Euclidean distance under the Markov chain models of base composition was generally recommended, of which sensitivity and selectivity were of 18 and 17 sequences, respectively, of order one for base composition, and 18 and 16 sequences, respectively, of order two for base composition; when all the distances of nine different word sizes were combined. Whereas both sensitivity and selectivity obtained from our proposed method were of 18 sequences. The false acceptances given by the proposed method are HUMCAN and AMTUGSNRNA; whereas the false rejections are GPILPPL and DMYOLK. Table 4 shows the sensitivity and selectivity obtained from the combined distance measures introduced by Wu *et al*. (2001) and from our proposed KLD of Markov models. The KLD of Markov models produced the highest number of HSLIPAS-related sequences on selectivity, and equal to the highest result obtained from the other methods (Mahalanobis and standardized Euclidean distances) on sensitivity. However, the computation of the proposed method is more efficient in comparison with both the Mahalanobis and the standardized Euclidean distances under the Markov chain models of base composition.

CONCLUSIONS

Comparison between sequences is a key step in bioinformatics when analyzing similarities of functions and properties of different sequences. Similarly, evolutionary homology is analyzed by comparing DNA and protein sequences. So far, most such analyses are conducted by aligning first the sequences and then comparing at each position the variation or similarity of the sequences. Multiple sequence alignments of several hundred sequences is thereby always a bottleneck, first due to long computational time, and second due to possible bias of multiple sequence alignments for multiple occurrences of highly similar sequences. An alignment-free comparison method is therefore of great value as it reduces the technical constraints as only pairwise comparisons are necessary, and

Table 4. Comparisons of sensitivity and selectivity

	Sensitivity	Selectivity
No assumption on model for base composition (Wu *et al*., 2001)		
Modified KLD1	17	14
Modified KLD2	17	15
Combined Euclidean distance	17	12
Under independent-uniform model for base composition (Wu *et al*., 2001)		
Combined Mahalanobis distance	18	17
Combined standardized Euclidean distance	18	16
Under Markov chain model of order one for base composition (Wu *et al*., 2001)		
Combined Mahalanobis distance	18	17
Combined standardized Euclidean distance	18	17
Under Markov chain model of order two for base composition (Wu *et al*., 2001)		
Combined Mahalanobis distance	18	16
Combined standardized Euclidean distance	18	16
KLD of Markov models (proposed SimMM)	18	18

is free of bias. For this alignment-free method, each pairwise comparison is unrelated to other pairwise comparisons.

Our SimMM method or KLD of Markov models presented in this work, is one of the methods, which is able to calculate the distances between DNA sequences without prior alignment. It calculates the similarity between two sequences by computing the log-likelihood difference between the two Markov models with the same observation sequence. The SimMM method has been tested with seven DNA sequences (one random, three from *E.coli* K-12 and three from *S.flexneri*), and another complex set of 40 DNA sequences. In the first experiment, the proposed alignment-free method showed good agreement with alignment-based sequence comparison, as it was able to cluster sequence families (thrA, thrB and thrC) of different species, and was able to discriminate the sequences against a randomly generated sequence with the same length and base composition as one of the sequence family (thrA). In addition, SimMM was able to discriminate real sequences from synthetic sequences (randomly generated) even better than the alignment-based method. The results obtained from SimMM have shown a slightly different orientation of the family clusters [(thrA,thrC),thrB], compared with the alignment-based method that gives [thrA,(thrB,thrC)]; but interestingly, SimMM yielded the same orientation as that given by another alignment-free method. However, we wish to mention that we have not analyzed any correlation between the scale of similarity in the SimMM and the distance in any evolutionary or mutational dimension. At this stage, the SimMM similarity can only be used for a topological analysis of the similarity or distance between DNA sequences. In the second experiment, our proposed method generally outperformed other alignment-free distance measures of sequence similarity on selectivity, but its computational cost is much less and its procedure is much simpler for computer implementation.

In the study of DNA sequences, there are four states representing 4 nt symbols, which give 16 elements in the probability transition matrix. Whereas for protein sequences, there are 20 amino-acid symbols that give 400 elements in the 20×20 transition matrix. This implies that the construction of the transition matrix for a protein sequence may not provide an estimate of the model parameters as well as it equivalently does for a DNA sequence if the protein sequence does not cover sufficient occurrences of all events. We have not investigated the comparison of protein sequences in this present study. However, if zero or very low-probability events occur in the estimate due to insufficient data, a numeric floor value for these transition elements might be appropriately specified and all remaining parameters are rescaled so that they obey the stochastic constraints. A similar problem has been discussed in the implementation of hidden Markov models for speech recognition (Rabiner and Juang, 1993).

All in all, our main discussion here is to present a new and effective computational framework for sequence comparison to the research community of bioinformatics. The proposed method can be considered as another useful tool among other alignment-based and alignment-free methods for sequence comparison; however, there is s no single method for solving all the problems in biological sequence comparison. The proposed method is based on the mathematical theories of the Markov process and the Kullback-Leibler divergence. However, its mathematical modeling is much more simple for practical implementation in comparison with many other alignment-free methods, and its computation based on the first-order Markov chain is roughly on the linear order of the length of the sequence. Higher-order Markov chains are expected to capture more useful information by the better power of predicting probabilities but require more computational effort and training data. It can be foreseen that both

alignment-based and alignment-free methods can be combined at some different levels, as such a particular aspect is the fusion of multiple distance matrices for a better solution.

ACKNOWLEDGEMENTS

We thank the three anonymous referees for their constructive suggestions and critical comments, which helped to improve the quality of the manuscript. One of the referees suggested us to test our method against the same dataset used by Wu *et al*. (2001).

REFERENCES

Almeida,J.S., Carrico,J.A., Maretzek,A., Noble,P.A. and Fletcher,M. (2001) Analysis of genomic sequences by chaos game representation. *Bioinformatics*, **17**, 429–437.

Blaisdell,B.E. (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl Acad. Sci. USA*, **83**, 5155–5159.

Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory.* John Wiley and Sons, NY.

Ewens,W.J. and Grant,G.R. (2001) *Statistical Methods in Bioinformatics*. Springer, NY.

Felsenstein,J. (1993) PHYLIP (Phylogeny Inference Package), version 3.5c. Distributed by the Author, Department of Genetics, University of Washington, Seattle, WA.

Juang,B.H. and Rabiner,L.R. (1985). A probabilistic distance measure for Hidden Markov Models, *AT&T Technical Journal*, **64**, 391–408.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.), *Mammalian Protein Metabolism*. Academic Press, NY, pp. 21–132.

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

Kroupa,T. (2003) Measure of divergence of possibility measures. In *Proceedings of the 6th Workshop on Uncertainty Processing (WUPES'2003)*, Hejnice, Czech Republic, pp. 173–181.

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

Li,L., Jin,R., Kok,P.L. and Wan,H. (2004) Pseudo-periodical partitions biological sequences. *Bioinformatics*, **20**, 295–306.

Miller,W. (2001) Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics*, **17**, 391–397.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Rabiner,L. and Juang,B.H. (1993) *Fundamentals of Speech Recognition*. Prentice Hall, NJ.

Stuart,G.W., Moffett,K. and Baker,S. (2002) Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, **18**, 100–108.

Taha,H.A. (1982) *Operations Research: An Introduction*. Macmillan, NY.

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

Vinga,S. and Almeida,J. (2003) Alignment-free sequence comparison—a review. *Bioinformatics*, **19**, 513–523.

Wu,T.J., Burke,J.P. and Davison,D.B. (1997) A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics*, **53**, 1431–1439.

Wu,T.J., Hsieh,Y.C. and Li,L.A. (2001) Statistical measures of DNA dissimilarity under Markov chain models of base composition. *Biometrics*, **57**, 441–448.