J Biochem Tech (2012) 3(4): 354-357
ISSN: 0974-2328

Discovery of evolutionary patterns in ribosomal RNA data using markov models

Somashekara M T*, Muralidhara B L, Manjunatha D

Received: 25 February 2012 / Received in revised form: 26 February 2012, Accepted: 27 February 2012, Published online: 28 February 2012,

Abstract

The paper deals with the construction of the phylogenetic tree and multiple sequence alignment of ribosomal RNA datasets downloaded from Silva comprehensive rRNA database using markov models. The new method is based on the concept of comparing the similarity/dissimilarity between two Markov models using Kullback–Leibler divergence for construction of pair-wise distance matrix. The alignment accuracy (sum of pair's scores) of the multiple sequence alignment of these datasets is compared with ClustalX software (progressive alignment). The statistical significance and the deviation of the alignment accuracy from ClustalX software was found by analysis of variance (ANOVA). Our results using the new method showed good agreement with the ClustalX-Multiple Sequence Alignment and phylogenetic tree and confirming the use of probabilistic Markov models in rRNA dataset analysis. All datasets and computer codes written in MATLAB are available upon request from the first author (somashekar_mt@hotmail.com).

Key words: Markov models, phylogeny, ClustalW, progressive alignment, guide tree.

Introduction

Molecular studies of phylogenetic relationships within higher taxonomic groups, e.g. at the intra-ordinal level, still rely on individual genes, among which the nuclear and mitochondrial ribosomal RNA genes are the most frequently sequenced (Harald & Karl, 2011). This, combined with the ease of amplification, has lead to a widespread use of rRNA genes in phylogenetics and furthermore uncovered several specific properties of these genes, which should be considered, using these sequences as phylogenetic markers. There are more than 500 000 publicly available small and large subunit (SSU and LSU) rRNA sequences, maintained by specialized quality controlled databases and software tools like ARB suite (Wolfgang et al. 2004) and Silva online resource (Elmar et al. 2007). Comparative sequence analysis by constructing phylogenetic trees and multiple sequence alignment of the small subunit rRNAs has already been established as the most commonly applied approach for phylogeny inference as well as microbial taxonomy and identification.

Pelin et al. (2011) studied rRNA genes in analyzing the global ocean sampling expedition, Thomas et al. (2011a) constructed rRNA phylogenetic tree for quantifying Korarchaeota and similar studies were widely found in literature (Dmitriy et al. 2012; Thomas et al. 2011b; Xu et al. 2011). Current studies on the topic of modeling rRNA data in tree construction have utilized simulation analyses, which can generally be seen to be a sophisticated complement to empirical studies (Harald and Karl 2011). Patrick and Sarah (2011) proposed a new heuristic method that has a minimal effect on the robustness of operational taxonomic units and significantly reduces the necessary time and memory requirements for rRNA sequence analysis. Their results opened a new dimension of thought process that heuristic algorithms can be successfully implemented in rRNA dataset analysis.

Tuan et al. (2004) classified the study of similarity/dissimilarity of sequences into two distinct groups as alignment-based (Progressive methods) and alignment-free methods (Probabilistic methods). Our new method is the combination of both methods i.e. using alignment free methods (probabilistic Markov models) in alignment based algorithms. Using our new method, we replaced the FAST algorithm used in the guide tree construction for multiple sequence alignment in progressive alignment technique with the new guide tree constructed using the distance matrix calculated by comparing the similarity/dissimilarity between two Markov models. The principle of using rRNA sequences to characterize microorganisms and its

Somashekara M T*

Department of Computer Science and Applications, Bangalore University, Bangalore- 560056, Karnataka, India

Email: somashekar_mt@hotmail.com;

Muralidhara B L

Department of Computer Science and Applications, Bangalore University, Bangalore- 560056, Karnataka, India

Manjunatha D

Department of Electronic Science, Tumkur University, Tumkur-572103, Karnataka, India

Table 1: Details of the datasets downloaded from Selvi-Comprehensive rRNA database

Datasets	Taxon (SSUr108)	No of sequences
Dataset 1	Archaea;Euryarchaeota;Methaomicrobia;Methaomicrobiales;02-02-504;	4
Dataset 2	Archaea;Euryarchaeota;Methaomicrobia;Methaocellales;MidArch4;	7
Dataset 3	Archaea;Euryarchaeota;Methanomicrobia;Methanosarcinales;Methermicoccaceae; Methermicoccus;	16
Dataset 4	Achaea;Euachaeota;Methaomicobia;Methaosaciales;GoM-Ach87;	3
Dataset 5	Archaea;Euryarchaeoa;Mehaomicrobia;Mehaomicrobiale;Rice Cluer II;	19
Dataset 6	Archaea;Euryarchaeota;Methaomicrobia;Methaomicrobiales;Methaomicrobiaceae;Methaoplaus;	12
Dataset 7	Archaea;Euryarchaeota;Methaomicrobia;Methaomicrobiales;C19A;	5

general application can be anticipated if methods for finding evolutionary patterns of rRNA sequences can be further improved by using novel algorithms.

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:

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

where q_{tn} stands for the actual state at time t_n (n = 1, 2, . . .), 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

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

$$\sum_{j=1}^{N} a_{ij} = 1 \ \ \forall \ \ i,$$

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

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

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

$\lambda = (A, \pi)$.

The above model is called the first order Markov model. We can also define second, third and higher order markov models, but our process is based only on the first order markov model.

Let λ1 = (A1, π1) and λ2 = (A2, π2) be two Markov first order 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 λ1 and λ2. A well-known dissimilarity measure between two probability distributions is the Kullback–Leibler Divergence (KLD) (Tuan et al. 2004). Detailed explanation of KLD is available from: http://bioinformatics.oxfordjournals.org/content/20/18/3455.full.pdf.

Materials and methods

Seven rRNA datasets are randomly selected from SELVI-comprehensive ribosomal database (http://www.arb-silva.de/). The rRNA sequences are downloaded from each dataset and the sequences with non nucleotide alphabets like 'n', 's', 'w', 'k' etc., are removed using Matlab Software. The details of the seven Small Sub Units (SSUr108) rRNA datasets are shown in Table 1.

The progressive alignment contains three steps (Pearson & Lipman 1988)

1. Initialization: all pairwise comparisons are performed by a fast algorithm and their scores are recorded.
2. A hierarchical clustering of the sequences is done using these scores.
3. The hierarchical tree is climbed with the pair-wise alignment of clusters to obtain the complete alignment. The alignment is shown, recorded or printed. A score is given for the multiple alignments: it is the sum of the scores of all the pairwise alignments included in the multiple one. A new hierarchical clustering is done with these new scores. if the new clustering is different from the old one, a new multiple alignment can be done following the new clustering (step 2). This process can be repeated until the clustering of the sequences is unchanged.

Our aim is to use a new algorithm instead of FAST algorithm in the first step for improving the pair-wise scores. The probabilistic distances among seven datasets are found using similarity measure by comparing Markov models with KL divergence. The Markov model is implemented in Matlab software (MathWorks, USA). The phylogenetic tree was constructed using UPGMA (Phylip). The phylogenetic tree is loaded into the third step of progressive alignment algorithm as a guide tree for constructing multiple sequence alignment using Matlab. ClustalX v2.1 software (http://www.ncbi.nlm.nih.gov) is used for calculating the sum of pair's column scores using standard IUB DNA weight matrix. The deviation of the column scores between Markov method and progressive alignment is found by using Analysis of variance (ANOVA) (Statistica V7.0, Statsoft, USA).

Results and discussion

The new method had been tested successfully with seven datasets taken from Silva comprehensive rRNA database. The probabilistic distances of the sequences present in each dataset were obtained using the markov method. The probabilistic distance matrix was used in construction of phylogenetic tree with the help of UPGMA program (PHYLIP package). The guide tree was loaded into the third step of progressive alignment (See materials and methods) with the help of Matlab bioinformatics toolbox. The final multiple sequence alignment constructed using Markov model-progressive alignment was saved as a text document and uploaded into ClustalX software for calculating the column scores. To compare our Markov method-progressive alignment **with** other methods, we calculated the sequence similarity or sequence distances using alignment-based methods (FAST algorithm-Progressive alignment). All seven datasets had been aligned and column scores were calculated using

Table 2: Comparison of sum of column scores of the multiple sequences constructed using Markov model-progressive alignment and Fast algorithm-progressive alignment

Datasets	Quality scores of multiple sequence alignment	
	Markov Model combination with progressive alignment	Fast Algorithm and progressive alignment
Dataset 1	93708	93787
Dataset 2	101244	101528
Dataset 3	92173	92636
Dataset 4	125884	125956
Dataset 5	80022	84006
Dataset 6	**123140**	122098
Dataset 7	88729	88886

ClustalX (Thompson et al., 1994). The one way ANOVA was applied to check the deviation between the values given by the two models for all datasets. The ANOVA for all seven datasets was highly significant with an F value of 465.4 as shown by Fisher's F test, along with a very low probability value ($P<0.05$), which was significant at 95% confidence interval. This confirms that there was a significant difference between the results of each dataset and there is no much deviation between the Markov model-progressive alignment and FAST algorithm-progressive alignment.

Table 3: One Way Analysis of Variance of the sum of column scores

	Sum of Squares	DF	Mean Squares	F	p
Intercept	1.42E+11	1	1.423E+11	115611	0.0
Datasets	3.448E+09	6	5.749E+08	465.4	0.0
Error	8.644E+06	7	1.237E+06		

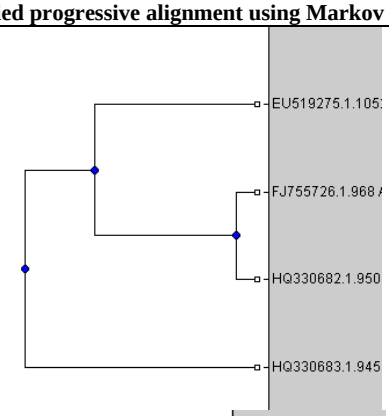
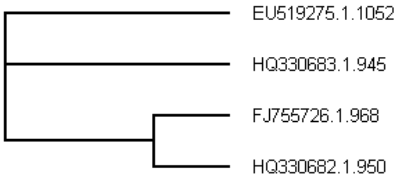
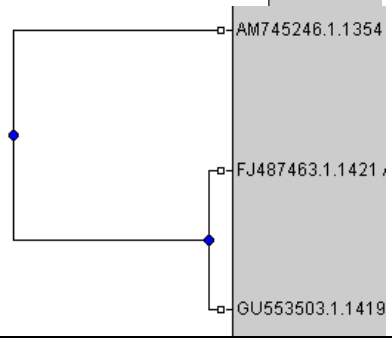
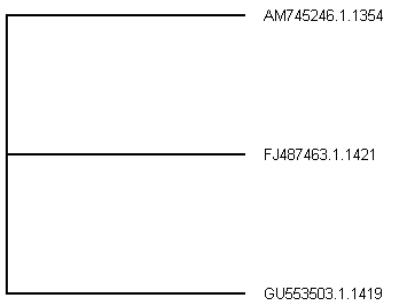
DF: Degree of Freedom; F: F-value; p: p-value

Difference between Modified progressive alignment using Markov model and conventional progressive alignment (FAST Algorithm-Progressive alignment) can also be illustrated by analyzing 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 ClustalX tree (Last column - Table 3) than in the tree using the new method (Second Column - Table 3). We had shown only two datasets in the Table 4 but the phylogenetic trees of remaining datasets (Dataset 3-dataset 8) were also found to be the same. This confirms the usefulness of our new method, when compared with ClustalX.

Conclusion

The proposed method can be considered as another useful tool among other alignment methods for rRNA sequence comparison. The trees are more informative when compared with the conventional progressive alignment technique. The deviation of the sum of column scores between the progressive alignment and

Table 4: Comparison of phylogenetic trees constructed using the two methods

Dataset name	Phylogenetic trees	
	Modified progressive alignment using Markov Model	Fast algorithm and progressive alignment
Dataset 1		
Dataset 2		

Markov model is less with respect to ANOVA table. The experiments show clearly that tree estimation can be improved through the use of improved guide trees using Markov method. It is also clear that these improvements require some additional computational effort. The wealth of novel data analyzing techniques should help to answer many open questions concerning the structure, function and evolution patterns of ribosomal RNAs.

Acknowledgement

We thank Dr. Tuan D, School of Computing and Information Technology, Griffith University, Australia, for helping in the development of the algorithm. Our sincere thanks are also due to Dr. Pradeep G. Siddheshwar, Professor of Mathematics, Bangalore University, Bangalore, India.

References

Dmitriy VV, Vahan S, Maureen KD, Vladimir EC (2012) RNA polymerase beta subunit (rpoB) gene and the 16S–23S rRNA intergenic transcribed spacer region (ITS) as complementary molecular markers in addition to the 16S rRNA gene for phylogenetic analysis and identification of the species of the family Mycoplasmataceae. Mol Phylogenetics Evol 62(1):515-528

Elmar P, Christian Q, Katrin K et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35(21):7188-7196

Harald OL, Karl MK (2011) Potential pitfalls of modelling ribosomal RNA data in phylogenetic tree reconstruction: Evidence from case studies in the Metazoa. BMC Evol Biol 11:146

Patrick DS, Sarah LW (2011) Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. Appl Environ Microbiol 77:6908-6917

Pearson WR, Lipman, DJ (1988) Improved tools for biological sequence comparison. In: Proceedings of the National Academy of Sciences Mar 1; USA. pp. 2444–8.

Pelin Y, Renzo K, Elmar P et al. (2011) Analysis of 23S rRNA genes in metagenomes – A case study from the Global Ocean Sampling Expedition. Sys App Microbiol 34(6):462-469

Thomas AA, Galina S, Colleen MC (2011a) 16S rRNA phylogenetic analysis and quantification of Korarchaeota indigenous to the hot springs of Kamchatka, Russia. Extremophiles 15(1):105-116

Thomas MG, Stefan JG, Christopher WS (2011b) Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. Env Microbiol 1462-2920

Tuan D Pham, Johannes Z (2004) A probabilistic measure for alignment-free sequence comparison. Bioinformatics 20(18):3455–3461

Wolfgang L, Oliver S, Ralf W et al. (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32(4):1363-1371

Xu ZH, Jiang SD, Wang GZ et al. (2011) DNA extraction, amplification and analysis of the 28S rRNA portion in sediment-buried copepod DNA in the Great Wall Bay and Xihu Lake, Antarctica. J Plankton Res 33(6):917-925