International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 3, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Probabilistic Measures of Similarity/Dissimilarity Between Markov Models for Construction of Guide Tree for Multiple Sequence Alignment

M. T. Somashekara
Dept of Computer Science Engg.
Bangalore University,
Email: somashekara_mt@hotmail.com

R. R. Siva Kiran
Department of Biotech,
MSRIT, Bangalore
Email: reddykiran.rsr@gmail.com

B. L. Muralidhara
Dept of Computer Science Engg.
Bangalore University
Email: muralidharabl@yahoo.com

***Abstract* – In this paper, we explore the consequences of changing the method of construction of guide tree used for the progressive alignment. The 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. We evaluated both the leading MSA method, ClustalW as well as the new MSA method which we have developed on benchmark DNA datasets. Improvement in alignment accuracy (i.e. sum of pair's scores) is observed when compared with ClustalW method.**

***Keywords* – Hierarchical Clustering, Progressive Alignment, ClustalW, Column Score, Markov Model, Distance Matrix, Guide Tree**

I. Introduction

The most widely used approach to multiple sequence alignment is hierarchical clustering. The hierarchical clustering method uses guide trees in conjunction with the progressive alignment technique to generate multiple sequence alignment. In the progressive alignment technique, the first step is to compute a pair-wise distance matrix which is then used to construct guide tree. The Multiple Sequence Alignment (MSA) is built by adding the sequences sequentially to the growing MSA according to the guide tree. The algorithm used for pair-wise distance matrix construction in progressive alignment (ClustalW) is Fast algorithm developed by Lipman and Pearson (1985) [1]. The guide tree is constructed based on the pair-wise distance matrix scores using hierarchical clustering algorithm developed by Florence Corpet (1988) [2]. The improvement in the final multiple sequence alignment using progressive alignment technique can be done in three different methods. The first method is to improve the pair-wise distance matrix and second is to develop a method for guide tree construction using pair-wise distance matrix and the third is to construct the final alignment from guide tree.

The effect of pair-wise distance matrix and guide tree on multiple sequence alignment was studied by Nelesen et al (2007) [3]. Nelesen et al applied Fixed Tree Alignment (FTA) technique based on the Sankoff problem for construction of distance matrix and guide tree. They found that the guide tree estimation can be improved by changing the algorithm used for pair-wise distance matrix. Gordon et al (2010) [4] studied the application of sequence embedding technique for fast construction of guide trees. In this paper, we introduce a probabilistic measure developed by Tuan et al (2004) [5] for calculating the pair-wise distance matrix for guide tree construction for multiple sequence alignment. The distance matrix construction method is based on the concept of comparing the similarity/dissimilarity between two Markov models using Kullback–Leibler divergence. The MSA is performed with three benchmark datasets downloaded from http://csl.cs.byu.edu/mdsas [6]. The alignment accuracy of these datasets is compared with ClustalW software (progressive alignment). Our results using the new method showed good agreement with the reference MSA from the benchmark dataset and also showed significant improvement in the accuracy when compared with ClustalW.

II. 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 π = {π_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:

λ = (A, π)

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) [7]. Detailed explanation of KLD is available from: http://bioinformatics.oxfordjournals.org/content/20/18/3455.full.pdf.

III. Materials and Methods

The DNA benchmark datasets were downloaded from http://dna.cs.byu.edu/mdsas/download.shtml. There are two versions of the datasets available in the benchmark database, one with 100 percentage sequence similarity and the other with threshold E-value of 0.001 or better. The dataset with 100 percentage sequence similarity were selected and downloaded from the link http://dna.cs.byu.edu/mdsas/zipped/balibase_mdsa_100s.zip. We have considered only low complexity sequences (100 percentage similar sequences) for checking accuracy of our program. There are six datasets available in the downloaded link in which we have taken only three datasets, each containing, four, eight and four sequences respectively (Table 1).

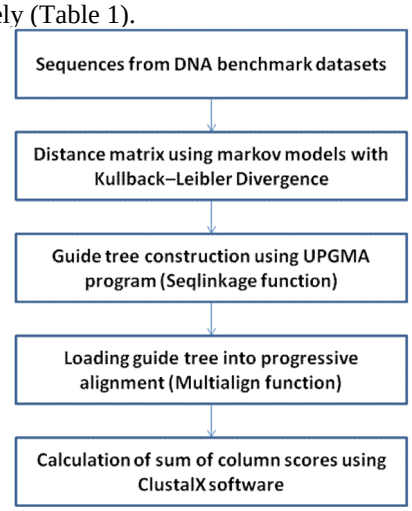

Fig.1. Flow diagram showing the steps involved in the new algorithm

The progressive alignment contains three steps [2] (Fig 1)

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 six sequences are found using similarity measure by comparing Markov models with KL divergence. The Markov model is implemented in Matlab software (MathWorks, USA). The guide tree is constructed using the UPGMA program available in 'Seqlinkage' function in Matlab software.

Table 1: Details of the Datasets downloaded from DNA Benchmark Database collected by Hyrum *et al.* 2004 [6]

S. No.	Dataset file Name	List of NCBI accession numbers of all DNA sequences present in the dataset file	Total number of sequences in the dataset
1	RV11_BBS11025.afa	DQ891020.1, BX640422.1, CP000077.1, DQ778054.1	4
2	RV11_BBS11002.afa	NM_007313.2, NM_001031685.2, NM_181523.1, AY779561.1, NM_001046541.1, NM_006533.1, DQ893150.1, NM_147152.1	8
3	RV11_BBS11008.afa	NM_001024955.1, DQ893861.1, XM_635569.1, NM_009283.3	4

'Multialign' function is also present in Matlab Bioinformatics toolbox which is used to load the guide tree into the third step of progressive alignment algorithm. The detailed algorithm is shown in the below flow diagram. ClustalX v2.1 software is used for calculating the column scores using standard IUB DNA weight matrix. The column scores are imported into Microsoft Excel 2007 (Microsoft, USA) for finding the sum of all column scores.

IV. Results

The new algorithm has been tested with three datasets taken from DNA Benchmark dataset. The dataset file name, NCBI accession numbers and the sequences information are shown in Appendix. The sequence information has been obtained from GenBank (www.ncbi.nlm.nih.gov/Entrez). The probabilistic distances among the three datasets obtained using Markov model are shown in Table 2.

Nelson *et al.* [3] showed that the effect of phylogenetic method used for guide tree construction from the distance matrix is very less on the accuracy of the multiple sequence alignment. In this paper, we have used UPGMA method for construction of the trees from the distance matrices. The guide tree drawn using UPGMA method using Seqlinkage function (Matlab v7.0, Mathworks, USA) is shown in Figure 2 for all three datasets.

Multiple sequence alignment is done using Matlab software, *Multialign* function, which loads the above guide trees and constructs the MSA using the final step of progressive alignment. The Figure 3 shows the screenshot of MSA constructed using Matlab software. *Multialignviewer* function is used for viewing the MSA.

International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 3, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Table 2: Probabilistic distance matrix (symmetric) using Markov Model for all three dataset files

Dataset file name: RV11_BBS11025.afa

	DQ891020.1	BX640422.1	CP000077.1	DQ778054.1
DQ891020.1	0	0.0013186442	0.0021480946	0.0019817755
BX640422.1		0	0.0036053049	0.0031551228
CP000077.1			0	0.0023303281
DQ778054.1				0

Dataset file name: RV11_BBS11002.afa

	NM_007313.2	NM_001031685.2	NM_181523.1	AY779561.1	NM_001046541.1	NM_006533.1	DQ893150.1	NM_147152.1
NM_007313.2	0	0.001194818	0.000847384	0.00071014	0.001337198	0.001199849	0.00071165	0.000454735
NM_001031685.2		0	0.000563804	0.00153049	0.00084853	0.001111508	0.00115621	0.001224757
NM_181523.1			0	0.00083132	0.001104656	0.001390397	0.001365221	0.000618528
AY779561.1				0	0.001625538	0.002023913	0.001345772	0.000238816
NM_001046541.1					0	0.000590175	0.000631444	0.001251272
NM_006533.1						0	0.00059489	0.001536727
DQ893150.1							0	0.001045977
NM_147152.1								0

Dataset file name: RV11_BBS11008.afa (Fig. 3)

	NM_001024955.1	DQ893861.1	XM_635569.1	NM_009283.3
NM_001024955.1	0	0.00249289	0.001270263	0.000959415
DQ893861.1		0	0.00368434	0.001002342
XM_635569.1			0	0.001916808
NM_009283.3				0

The accuracy of the multiple sequence alignment is assessed by calculating the sum of column scores using quality menu option present in ClustalX software. The multiple sequence alignment is performed using progressive alignment (ClustalW software) and the scores are determined using ClustalX. The sum of column scores of the progressive alignment (ClustalW), reference alignment (Benchmark DNA dataset) and the markov method are compared.

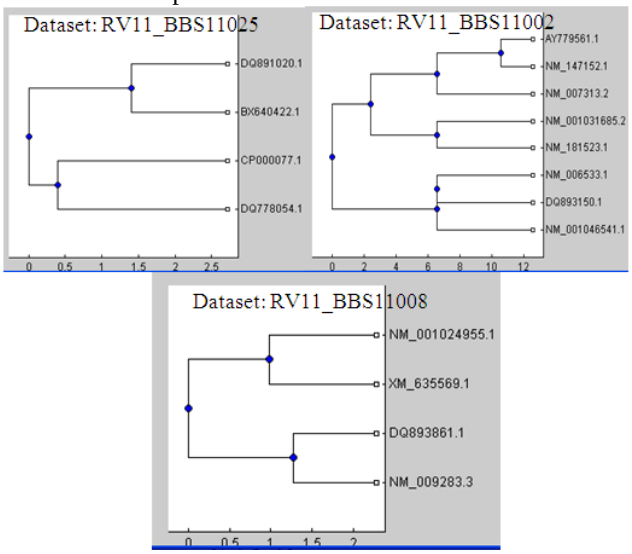

Fig.2. Guide tree using UPGMA method

Table 3: Two way ANOVA calculated using Statistica V7.0 (Statsoft USA)

	Sum of Squares	Degree of freedom	Mean Squares	F-value	p-value
Intercept	1.61E+09	1	1.61E+09	6089.8	0.0
Datasets	1.22E+08	2	6.14E+07	232.2	0.73E-4
Error	1.05E+06	4	2.64E+05		

Analysis of variance is applied to check the deviation between the values given by the three models for all datasets. The ANOVA for all three datasets is highly significant with an F value of 232.251 as shown by Fisher's F test, along with a very low probability value (P-model>F=0.0001), which is significant at 95% confidence interval. This confirms that the deviation of column scores between three models is very less according to the Table 3. The Markov model showed significant increase in the sum of column scores when compared with the progressive alignment (ClustalW).

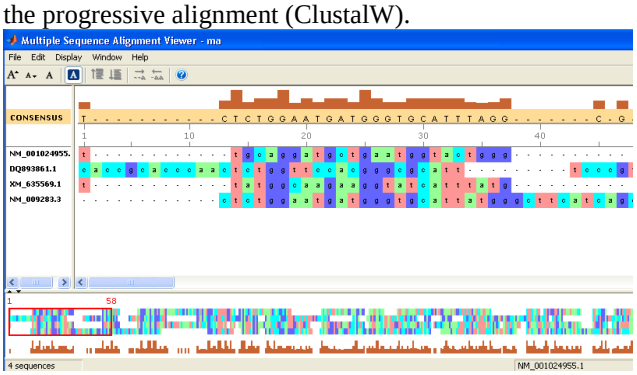

Fig.3. Matlab Version 2010, Screenshot of MSA drawn for dataset (RV11_BBS11008).

V. Conclusion

The proposed method can be considered as another useful tool among other multiple sequence alignment tools. Significant increase in the sum of column scores (Table 4) is observed when compared with ClustalW confirming the validity of the new algorithm on par with ClustalW. The deviation between the sum of column scores between the reference dataset, the progressive alignment and 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.

Table 4: Comparison of quality scores of the multiple sequences constructed using ClustalW and Markov Model

Datasets	Guide tree calculations		Reference Dataset
	Markov Model	Fast Algorithm in ClustalW	
RV11_BBS11025.afa	13576	13258	12275
RV11_BBS11002.afa	9156	8809	9130
RV11_BBS11008.afa	18884	18260	17041

ACKNOWLEDGMENT

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

APPENDIX

NCBI accession numbers, organism names and few details of the sequences

Dataset File Name	NCBI Accession No.	Organism Name	Details of The Sequences
RV11_BBS11025.afa	DQ891020.1	*Homo sapiens*	Synthetic construct Homo sapiens clone FLH168582.01X; RZPDo839A1294D PPBP mRNA,
	BX640422.1	*Bordetella pertussis*	Bordetella pertussis strain Tohama I, complete genome; segment 12/12
	CP000077.1	*S. acidocaldarius DSM 639*	Sulfolobus acidocaldarius DSM 639, complete genome
	DQ778054.1	*Escherichia coli strain Eco412*	Escherichia coli strain Eco412 heat-labile enterotoxin A subunit and heat-labile enterotoxin B subunit genes, complete cds
RV11_BBS11002.afa	NM_007313.2	*Homo sapiens*	Homo sapiens c-abl oncogene 1, non-receptor tyrosine kinase (ABL1), transcript variant b, mRNA
	NM_001031685.2	*Homo sapiens*	Homo sapiens tumor protein p53 binding protein, 2 (TP53BP2), transcript variant 1, mRNA.
	NM_181523.1	*Homo sapiens*	Homo sapiens phosphoinositide-3-kinase, regulatory subunit 1 (alpha) (PIK3R1), transcript variant 1, mRNA
	AY779561.1	*HIV-1 isolate*	HIV-1 isolate CANC5FULL from Canada, complete genome
	NM_001046541.1	*Bos taurus*	Bos taurus bridging integrator 1 (BIN1), mRNA
	NM_006533.1	*Homo sapiens*	Homo sapiens melanoma inhibitory activity (MIA), mRNA
	DQ893150.1	*Homo sapiens*	Synthetic construct Homo sapiens clone FLH194366.01X; RZPDo839A0780D NCF1 mRNA, complete sequence
	NM_147152.1	*Homo sapiens*	Homo sapiens intersectin 2 (ITSN2), transcript variant 2, mRNA
RV11_BBS11008.afa	NM_001024955.1	*Mus musculus*	Mus musculus phosphatidylinositol 3-kinase, regulatory subunit, polypeptide 1 (p85 alpha) (Pik3r1), transcript variant 1, mRNA
	DQ893861.1	*Homo sapiens*	Homo sapiens clone FLH165033.01L; RZPDo839F01157D GRB7 mRNA, partial sequence
	XM_635569.1	*Dictyostelium discoideum AX4*	Dictyostelium discoideum AX4 signal transducer and activator of transcription family protein (dstA) mRNA, complete cds
	NM_009283.3	*Mus musculus*	Mus musculus signal transducer and activator of transcription 1 (Stat1), mRNA

REFERENCES

[1] W. R. Pearson, D.J. Lipman, "Improved tools for biological sequence comparison" In *Proceedings of the National Academy of Sciences*; Mar 1, 1988; USA. pp. 2444–2448.

[2] Florence. Corpet, "Multiple sequence alignment with hierarchical clustering", *Nucleic Acids Research*, Volume 16, 1988, Pages 10881-10890.

[3] S. Nelesen, K. Liu, D. Zhao, C. R. Linder, T. Warnow. "The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses", *Pacific Symposium on Biocomputing*, Volume 13, 2008, Pages 25-36.

[4] B. Gordon, S. Fabian, Shi Weifeng, "Sequence embedding for fast construction of guide trees for multiple sequence alignment". *Alg Mol Biol*, Volume 5, Issue 21, 2008, Pages 1-11.

[5] Tuan. D. Pham, Z. Johannes, "A probabilistic measure for alignment-free sequence comparison", *Bioinformatics* Volume 20, Issue 18, 2004, Pages 3455–3461.

[6] Hyrum. Carroll, Wesley. Beckstead, Timothy. O'Connor "DNA reference alignment benchmarks based on tertiary structure of encoded proteins", *Bioinformatics*, Volume 23, Issue 19, 2007, Pages 2648–2649.

[7] T.M. Cover, J.A. Thomas, "Elements of Information Theory", 1st ed. John Wiley & Sons; USA 1991.

AUTHOR'S PROFILE

M. T. Somashekara
is currently working as Assistant Professor in the Dept of Computer Science and Applications, Bangalore University. His areas of research include Bioinformatics, Pattern Recognition and Image Processing. Email: somashekara_mt@hotmail.com

R. R. Siva Kiran
is currently working as an Assistant Professor in the Department of Biotechnology at M.S. Ramaiah Institute of Technology, Bangalore. His research includes mathematical modeling, statistical optimization and bioinformatics.
Email: reddykiran.rsr@gmail.com

B. L. Muralidhara
is currently working as Associate professor in the Dept of Computer Science and Applications, Bangalore University, Bangalore. Earlier, he has even taught at University of Hyderabad. He has many research articles to his credit. His areas of research include Bioinformatics, Parallel Computation and e-Governance. Email: muralidharabl@yahoo.com