An Advanced Statistical Method of Multiple Sequence Alignment

Mian Arif Shams Adnan*, M. Shamsuddin
Department of Statistics, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh
Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia.

ABSTRACT

Motivation: Although sequence alignment has become one of the most influential techniques to discover the most probable functional and structural form of any evolutionary biological system, the study of sequence alignment can be missleaded very frequently either by introducing gap penalty calculation or by algorithm accuracy level.

Results: A method of multiple sequence alignment based on the difference among transition probabilities of the multiple sequences has been developed which is relatively complete and invariant to the aforementioned problems. It accomplishes not only an overall decision of the significant dissimilarity or similarity but also the dissimilarity or similarity of all possible individual and group wise transitions that help the biotechnologists to quickly identify the portion of the total infrastructure of the entire transitions that is significantly differing from those of the other sequences and detect the core fact(s) for possible differences between bio-organisms. Hence it reduces the incompleteness due to the comparison among the multiple sample sequences from the several populations.

*To whom correspondence should be addressed at julias284atyahoo.com,

Key words: Gap penalty, Matrix of chi-squares, Matrix of p-values, Transition probability.

1. INTRODUCTION

Sequence alignment is one of the most important techniques for discovering functional, structural, and evolutionary information of the concern biological sequences. Sequences that are very much alike, or “similar” in the parlance of sequence analysis, probably have the same function, or a similar biochemical and three dimensional structures in the case of proteins. If sequences from different organisms are similar, there may have been a common ancestor sequence, and the sequences are then defined as being homologous. The alignment indicates the changes that could have occurred among the homologous sequences and a common ancestor sequence during evolution.

Most modern programs (developed by more than two hundred authors [Chuong and Kathoh (2008)]) for constructing multiple sequence alignments (MSAs) consist of two components: an objective function for assessing the quality of a candidate alignment of a set of input sequences, and an optimization procedure for identifying the highest scoring alignment with respect to the chosen objective function [Notredame (2002)].

While most alignment techniques rely abstractly on a scoring scheme that uses substitution scores and gap penalties, they do not develop an explicit model of the evolutionary process rather the probabilistic methods for aligner construction has recently become more interesting. These techniques for multiple sequence alignment generally come in three main varieties: complex evolutionary models of insertion, deletion, and mutation in multiple sequences; fixed dimensionality profile models for representing specific protein families; and hybrid methods that combine probabilistic models with traditional ad hoc alignment techniques. Of the three approaches, evolutionary models for statistical alignment provide the most explicit representation of change in biological sequences as a stochastic process [Bishop and Thompson (1986), Hein *et al* (2000)]. Research in statistical alignment typically derive from the classic Thorne–Kishino–Felsenstein (TKF) pairwise alignment model [Thorne, Kishino and Felsenstein (1991)] in which amino acid substitutions follow a time-reversible Markov process and single-gap creation and deletion are treated as birth or death processes over imaginary "links" separating letters in a sequence. Subsequent work on statistical alignment has focused on modeling multiresidue, overlapping indels [Thorne ,Kishino and Felsenstein (1992), Miklos and Toroczkai (2001), Miklos (2003), Miklos, Lunter and Holmes (2004), Knudsen and Miyamoto (2003), Metzler (2003)], extending the TKF model to multiple alignment [Hein (2001), Hein, Jensen and Pedersen (2003), Holmes and Bruno (2001), Holmes (2003), Steel and Hein (2001), Miklos (2002), Lunter *et al* (2003), Jensen and Hein (2005)], and the even more complex task of coestimating alignment and sequence phylogeny [Steel and Hein (2001), Hein (1990), Vingron and Haeseler (1997), Fleissner, Metzler and Haeseler (2005), Lunter *et al* (2005), Redelings and Suchard (2005)]. Unlike traditional score-based alignment approaches, statistical alignment methods provide a natural framework for estimating the parameters underlying stochastic evolutionary processes [Metzler *et al* (2001)]. However, the resulting models are often quite complex. While dynamic programming is sometimes possible, these models often require sampling-based inference procedures [Allison and Wallace (1994)] that share many of the disadvantages of simulated annealing approaches discussed earlier. The accuracy of TKF-based techniques in alignment construction is unclear as few methods based on this approach have been comparatively benchmarked against standard programs; one exception is the Handel [Holmes and Bruno (2001), Holmes (2003)] program for statistical multiple alignment, which achieves substantially lower accuracy (i.e., 13% fewer correctly aligned residue pairs) than CLUSTALW, the prototypical score-based modern sequence aligner.

A second class of probabilistic modeling techniques is the profile hidden Markov model (profile HMM), a sophisticated variant of the character frequency profile matrices that takes into account position-specific indel probabilities [Durbin *et al* (1999), Krogh *et al* (1994), Krogh (1998), Hughey and Krogh (1996), Eddy (1996)]. To construct a profile HMM given a set of unaligned sequences, a length is chosen for the initial profile, as well as initial emission probabilities for each position in the profile and transition probabilities for indel creation and extension after each position. Next, the model is optimized according to a likelihood criterion using an expectation–maximization (EM)-based Baum–Welch procedure [Durbin *et al* (1999)], simulated annealing [Eddy (1995)], deterministic annealing [Mamitsuka (2005)], or approximate gradient descent (Baldi and Chauvin (1994), Baldi *et al* (1994)]. Finally, all sequences are aligned to the profile using the Viterbi algorithm [Viterbi (1967)] for finding the most likely correspondence between each individual sequence and the profile, and the correspondences of each sequence to the profile are accumulated to form the multiple alignment. Profile HMMs and their variants [Grundy *et al* (1997)] form the basis of many remote homology detection techniques [Bucher et al (1996), Karplus, Barrett and Hughey (1998), Park *et al* (1998)] and have been used to characterize protein sequence families [Sonnhammer et al (1998)]. Empirically, profile HMMs [Hughey and Krogh (1996), Eddy and HMMER] have great appeal in practice as they provide a principled probabilistic framework, and, when properly tuned [Sjolander *et al* (1996), Barrett, Hughey and Karplus (1997)], achieve good empirical performance close to that of CLUSTALW[McClure, Smith and Elton (1996), Karplus and Hu (2001)].

Finally, hybrid techniques combine the rigor of probabilistic model parameter estimation with standard heuristics for multiple alignments. The ProAlign [Loytynoja and Milinkovitch (2003)], COACH [Edgar and Sjolander (2004)], and SATCHMO [Edgar and Sjolander (2003), Edgar and Sjolander (2003)] progressive alignment tools, for instance, all achieve CLUSTALW accuracy; the recent PRANK aligner [Loytynoja and Goldman (2005)] has revealed the benefits of scoring insertions and deletions differently for the purposes of indel distribution estimation. A separate promising direction has been the development of the maximum expected accuracy (MEA) algorithm for pairwise alignment based on posterior match probabilities [Holmes and Durbin (1998)], which was generalized to consistency-based multiple alignment in the PROBCONS algorithm [Do *et al* (2005)]. Other programs based on the public domain PROBCONS source code include AMAP [Schwartz, Myers and Pachter (2006)], which optimizes an objective function that rewards for correctly placed gaps, and ProbAlign [Roshan and Livesay (2006)], which uses a physics-inspired modification of the posterior probability calculations in PROBCONS. Finally, the MUMMALS program [Pei and Grishin (2006)], which extends the PROBCONS approach to allow for more sophisticated HMM structures, has achieved the highest reported accuracies to date of all modern stand-alone multiple alignment programs.

In studies of multiple sequence alignment, the algorithms used can be important, but they are not the only consideration that must be made. Techniques for assessing aligner performance typically have one of four goals: (1) demonstrating the effectiveness of a particular heuristic strategy for SP objective optimization; showing that a particular software package achieves good accuracy relative to “gold standard” reference alignments of either (2) real or (3) simulated proteins; or (4) quantifying alignment accuracy on real data in a reference-independent manner. For comparing software packages relying on different objective functions, the first validation scheme is not applicable.

In real protein sequences, the true alignment of a set of sequences based on structural considerations is not necessarily the same as the true alignment based on evolutionary or functional considerations. In practice, structural alignments are relatively easy to obtain for proteins of known structure, and hence, are the de facto standard in most real-world benchmarks of alignment tools. Popular databases of hand-curated structural alignments include BAliBASE version 2 [Thompson, Plewniak and Poch (1999), Thompson, Plewniak, and Poch (1999)] and HOMSTRAD [Mizuguchi *et al* (1998)]. Because of the difficulty and lack of reproducibility of hand curation, a number of modern alignment databases rely on automated structural alignment protocols, including SABmark [Walle, Lasters and Wyns (2005)], PREFAB [Edgar (2004)], OxBench [Raghava *et al* (2003)], and to a large extent, BAliBASE version 3 [Thompson *et al* (2005)]. Because the correct protein structural alignment can sometimes also be ambiguous, most alignment databases annotate select portions of their provided alignments as “core blocks”—regions for which structural alignments are known to be reliable—and measures of accuracy such as the Q score [defined as the proportion of pairwise matches in a reference alignment predicted by the aligner; other measures of accuracy also exist (Sauder, Arthur and Dunbrack, 2000)] are computed with respect to only core blocks.

Finally, it is possible to avoid dealing with ambiguities in reference alignments using techniques that directly assess the quality of an alignment in terms of the resulting structural superposition. For a pair of proteins, the coordinate root-mean-square-distance (coordinate RMSD) between positions identified as “equivalent” according to an alignment (after the two protein structures have been appropriately rotated and translated) is a common measure for evaluating structural alignment quality. Several RMSD variants exist [Eidhammer, Jonassen and Taylor (2000)], including variants that account for protein length [Carugo and Pongor (2001)], that examine pairwise distances between residues in a protein [Armougom et al (2006)], or that rely on alternate representations of protein backbones [Chew *et al* (1999)]. Another recently proposed metric is the APDB measure [Sullivan *et al* (2003)], an approximation of the Q score that judges the “correctness” of aligned residue pairs based on the degree to which nearby aligned residues have similar local geometry in the sequences being aligned.

For traditional score-based sequence alignment procedures, estimation of substitution matrices and gap penalties are usually treated separately. Briefly, substitution matrices are generally estimated from databases of alignments known to be reliable. Statistical estimation procedures for constructing log-odds substitution matrices vary in their details, but most methods nonetheless tend to generate sets of matrices approximately parameterized by some notion of evolutionary distance for which that matrix is optimal. Popular matrices include the BLOSUM [Henikoff and Henikoff (1992)], PAM [Dayhoff, Eck and Park (1972), Dayhoff, Schwartz and Orcutt (1978)], JTT [Jones, Taylor and Thornton (1992)], MV [Muller and Vingron (2000)], and WAG [Whelan and Goldman (2001)] matrices; matrices derived from structural alignments for use with low-identity sequences also exist [Prlic, Domingues and Sippl (2000)]. For gap parameters, an empirical trial-and-error approach [Reese and Pearson (2002)] is common as the number of parameters to be estimated is low.

Probabilistic models have the advantage that the maximum likelihood principle provides a natural mechanism for estimating gap parameters when example alignments are available [Arribas-Gil, Gassiat and Matias (2006)]; when only unaligned sequences are available, unsupervised estimation of gap parameters can still be effective [Do et al (2005)]. Alternatively, Bayesian methods [Liu, Neuwald and Lawrence (1995), Zhu, Liu and Lawrence (1998)] automatically combine the results obtained when using multiple varying parameter sets and thus avoid the need for deciding on fixed parameter sets.

Recently, the problem of parameter estimation has been the subject of renewed attention, stemming from the influence of the convex optimization and machine learning communities. Kececioglu and Kim (2007) described a simple cutting-plane algorithm for inverse alignment—the problem of identifying a parameter set for which an aligner aligns each sequence in a training set correctly. Their algorithm is fast in practice, though the biological accuracy of the resulting alignments on unseen test data is unclear. Do *et al*. (2006) developed a machine learning-based method based on pair conditional random fields (pair-CRFs) called CONTRAlign, which achieves significantly better generalization performance than existing methods for pairwise alignment of distant sequences. Most recently, Yu et al. (2007) described a fast approach for training protein threading models based on support vector machines [Tsochantaridis *et al* (2005)], which shares many of the generalization advantages of CONTRAlign.

In the own work Hanus *et al* (2009) proposed an asymmetric source coding scheme for such alignments using evolutionary prediction in combination with lossless black and white image compression. Moreover, Hong *et al* (2008), Lu Y *et al* (2009), Singh *et al* (2010), contributed some new ideas on pairwise sequence alignment. Recently Hongwei *et al* (2008) proposed a hybrid algorithm based on artificial immune system and hidden Markov model for multiple sequence alignment. Prakash *et al* (2009) assessed the discordance of MSA by proposing a log likelihood score considering a multiple alignment with a length and related to phylogenetic tree. Sahraein *et al* (2011) proposed PicXAA-R as an extension to PicXAA for greedy structural alignment of ncRNAs. PicXAAR efficiently grasps both folding information within each sequence and local similarities between sequences. It uses a set of probabilistic consistency transformations to improve the posterior base-pairing and base alignment probabilities using the information of all sequences in the alignment. Using a graph-based scheme, we greedily build up the structural alignment from sequence regions with high base-pairing and base alignment probabilities.

Rajasekaran *et al* (2004) presented a randomized algorithm for distance matrix calculations in MSA. It deals with randomly sampling sequences and aligning to achive nearly the same result in terms of distance matrix calculation and achieve a significant routine improvement. Furthermore, they have extended the randomization approach to include non-uniform length sequences and also taken segmented approach to improve accuracy. Unfortunately, their paper does not show any mathematical presentation of the entire method and the proposed distance matrix is traditional and not known by following any the statistical distribution. Pena *et al* (2007) proposed nonparametric KS test in multiple hypothesis testing of transition matrices.

Maximum literature for quantifying the disorderness of two sequences in case of alignment algorithm has been suffering from either calculating superficial gap penalty or obtaining unsatisfactory accuracy or discordance matrix is not statistically sound in with respect to the mechanism or methodology or even the distribution of the discordance matrix is not found. After 1962, from Watson and Crick to Toshihide Hara *et al* (2010), many researchers have been investigating for knowing the most accurate way of pair wise sequence alignment as well as multiple sequence alignment. T. Hara again gives a flavor of improving the pair-wise sequence algorithm by introducing core analysis of transition probabilities of the sequence. Dannemann *et al* (2007) proposed a method of testing the equality of transition parameters based on transition probabilities and likelihood ratio test statistic that simply gives the significant dissimilarity of the total transition but not that of the individual transition. The present study aims to improve the pair-wise sequence alignment considering the more analysis of transition probabilities of the nucleotides from two sequences. The author introduces a new idea of using the difference of pair wise transition probabilities of the two sequences which will ensure three advantages at least. Firstly, it will find the degree disorderness between all possible individual and groupwise transition probabilities of nucleotides of two sequences; and secondly, will reduce the loss of comparison between the two sequences from the two unknown populations. Thirdly, it clearly identifies the portion of the total infrastructure of the entire transition that is significantly differing from that of the other sequence. The paper is organized as follows. Section 2 briefly describes proposed approach of the DNA sequence Alignment and section 3 evaluates it through some real life examples. The performance and advantages are referred to section 4 and the final section draws the conclusion.

Usual methods for aligning DNA sequence in the recent years use a measure empirically determined. As an example, a measure is usually defined by a combination of two quantities: (*i*) the sum of substitutions between two residue segments (*ii*) the sum of the gap penalties in insertions or deletion region. But it is true that the efficiency of the available alignment procedures are not up to the level desired. Improving pairwise sequence alignment procedure is an initial step of improving multiple sequence alignment procedure.

Recently a new statistical method of Pair-wise sequence alignment has been developed by Adnan *et al* (2011). It accomplishes not only an overall decision of the significant similarity/dissimilarity but also the similarity/dissimilarity of all possible individual and group wise transitions that help the biotechnologists to quickly identify the portion of the total infrastructure of the entire transitions that is significantly differing from that of the other sequence and detect the core fact(s) for possible differences between bio-organisms. Hence it reduces the loss due to the comparison between the two sample sequences from the two populations.

With an aim of developing an extension of the pairwise DNA sequence alignment the authors demonstrate an alternative statistical approach Multiple Sequence Alignment (MSA).

2. ALTERNATIVE METHOD OF MULTIPLE SEQUENCE ALIGNMENT

Let the stochastic process is $\{X(t); t \in T\}$, then for each value of t, $X(t)$ is a random variable. So, the process is a sequence of outcomes for discrete states and time space. These outcomes may be dependent on earlier ones in the sequence. A Markov chain is collection of random variables $X(t)$ (where the index runs through 0, 1, ...) having the property that, given the present, the future is conditionally independent of the past. So, the stochastic process $\{X_n, n \geq 0\}$ is called a Markov chain, if for $j, k, j_1, \ \ldots \ j_{n-1} \in J$

$$\Pr[X_n = k \mid X_{n-1} = j, X_{n-2} = j_1, \ \ldots \ X_0 = j_{n-1}] = \Pr[X_n = k \mid X_{n-1} = j] = P_{jk}$$

The outcomes are called the states of the Markov Chain; if X_n has the outcome j ($i.e.$, $X_n = j$) the process is said to be at state j at n^{th} trial. The conditional probability $P[X_{n+1} = j | X_n = i] = P_{ij}$ is known as transition probability referring the probability that the process is in stat i and will be in state j in the next step and the transition probability P_{ij} satisfy the properties $(i)\ P_{ij} \geq 0$ and $(ii) \sum_j P_{ij} = 1$ for the transition probability matrix $P = [P_{ij}]\ \forall\ i,\ j = 1, 2, \cdots, n$.

Here, two states i and j are said to be communicate state if each is accessible from the other, it is denoted by $i \leftrightarrow j$; then there exist integer m and n such that $P_{ij}^{(n)} > 0$ and$P_{ij}^{(m)} > 0$. If state i communicate with state j and state j communicate with state k then state i communicate with state k.

1.1 *Proposed method*

With an aim of developing a test procedure of testing the equality of several transition probability matrices or several evolutionary rates from several Markov chains or several sequences, let us demonstrate our method assuming that we have several population transition frequency matrices or several population transition probability matrices or several Markov chains each of which having *r* states and let the hypothesis be

$$H_0: N_1 = \ N_2 = \cdots = N_m$$

$$\Rightarrow H_0: \begin{pmatrix} N_{111} & N_{121} \cdots & N_{1r1} \\ N_{211} & N_{221} & N_{2r1} \\ & \vdots & \\ N_{r11} & N_{r21} & N_{rr1} \end{pmatrix} = \begin{pmatrix} N_{112} & N_{122} \cdots & N_{1r2} \\ N_{212} & N_{222} & N_{2r2} \\ & \vdots & \\ N_{r12} & N_{r22} & N_{rr2} \end{pmatrix} = \cdots$$
$$= \begin{pmatrix} N_{11m} & N_{12m} \cdots & N_{1rm} \\ N_{21m} & N_{22m} & N_{2rm} \\ & \vdots & \\ N_{r1m} & N_{r2m} & N_{rrm} \end{pmatrix}$$

$$H_0: P_1 = \ P_2 \ = \cdots . = P_m;$$

$$\therefore H_0: \begin{pmatrix} p_{111} & p_{121} \cdots & p_{1r1} \\ p_{211} & p_{221} & p_{2r1} \\ & \vdots & \\ p_{r11} & p_{r21} & p_{rr1} \end{pmatrix} = \begin{pmatrix} p_{112} & p_{122} \cdots & p_{1r2} \\ p_{212} & p_{222} & p_{2r2} \\ & \vdots & \\ p_{r12} & p_{r22} & p_{rr2} \end{pmatrix} = \cdots = \begin{pmatrix} p_{11m} & p_{12m} \cdots & p_{1rm} \\ p_{21m} & p_{22m} & p_{2rm} \\ & \vdots & \\ p_{r1m} & p_{r2m} & p_{rrm} \end{pmatrix}.$$

where, N_l ($\forall\, l = 1,2,\dots,m)$) is the population transition frequency matrix of the *l*th population such that $N_l = (n_{ijl})_{r\times r}$; P_l is the population transition probability matrix of the *l*th population such that $P_l = (p_{ijl})_{r\times r}$, where $p_{ij} = \frac{N_{ijl}}{N_{i.l}}$ whereas N_{ijl} is the population transition frequency of the (*i,j*)th element of the *l*th population transition frequency matrices N_l and $N_{i.l} = \sum_{j=1}^{r} N_{ijl}\,; \forall\, i,j = 1,2,\dots,r$.

k pairs of sample sequences from *m* populations (a total of *k* sample-sequences are collected from each population) have been collected and on the basis of these samples we want to test whether they come from the same population. After collecting *k* sample-sequences we obtain *k* transition frequency matrices from each of the *m* populations. The maximum likelihood estimators of the transition relative frequency or probability matrices are obtained as $\hat{P}_l = (\hat{p}_{ijl})_{r\times r}$ where $\hat{p}_{ijl} = \frac{n_{ijl}}{n_{i.l}}$ whereas n_{ijl} is the average frequency of the (*i,j*)th element of the average transition frequency matrix n_l constructed from *k* sample-transition frequency matrices drawn from the *l*th population. Here, $n_{i.l} = \sum_{j=1}^{r} n_{ijl}\,; \forall\, i,j = 1,2,\dots,r$.

For large $n_{i.l}$ the asymptotic distribution of each element of estimated transition probability matrices, according to the Central Limit Theorem, is normal such that

$$\hat{p}_{ijl} \ \sim N\left(p_{ijl}, \frac{p_{ijl}\ \ (1 - p_{ijl})}{kn_{i.l}}\right).$$

$$\therefore \sum_{l=1}^{m} \frac{\left(\hat{p}_{ijl} - \bar{p}_{ij.}\right)^2}{\frac{\bar{p}_{ij.}(1 - \bar{p}_{ij.})}{kn_{i.l}}} \sim \boldsymbol{\chi}^2{}_{(m-1)} \forall i,j = 1,2,\dots,r;$$

where $\bar{p}_{ij.} = \frac{n_{i.1}\hat{p}_{ij1} + \cdots + n_{i.l}\hat{p}_{ijl}}{n_{i.1} + \cdots + n_{i.l}}; \forall\, i,j = 1,2,\dots,r.$

However, we obtain an element-chi-square-matrix $\chi 2$ of the following form

$$\chi 2 = \begin{array}{c} \\ \\ \end{array}\begin{array}{cccccc} & 1 & 2 & & r & \\ 1 & \sum_{l=1}^{m} \frac{(\hat{p}_{11l}-\bar{p}_{11.})^2}{\frac{\bar{p}_{11.}(1-\bar{p}_{11.})}{kn_{1.l}}} & \sum_{l=1}^{m} \frac{(\hat{p}_{12l}-\bar{p}_{12.})^2}{\frac{\bar{p}_{12.}(1-\bar{p}_{12.})}{kn_{1.l}}} & \cdots & \sum_{l=1}^{m} \frac{(\hat{p}_{1rl}-\bar{p}_{1r.})^2}{\frac{\bar{p}_{1r.}(1-\bar{p}_{1r.})}{kn_{1.l}}} & \\ 2 & \sum_{l=1}^{m} \frac{(\hat{p}_{21l}-\bar{p}_{21.})^2}{\frac{\bar{p}_{21.}(1-\bar{p}_{21.})}{kn_{2.l}}} & \sum_{l=1}^{m} \frac{(\hat{p}_{22l}-\bar{p}_{22.})^2}{\frac{\bar{p}_{22.}(1-\bar{p}_{22.})}{kn_{2.l}}} & \cdots & \sum_{l=1}^{m} \frac{(\hat{p}_{2rl}-\bar{p}_{2r.})^2}{\frac{\bar{p}_{2r.}(1-\bar{p}_{2r.})}{kn_{2.l}}} & , \\ & \cdots & \cdots & & \cdots & \\ r & \sum_{l=1}^{m} \frac{(\hat{p}_{r1l}-\bar{p}_{r1.})^2}{\frac{\bar{p}_{r1.}(1-\bar{p}_{r1.})}{kn_{r.l}}} & \sum_{l=1}^{m} \frac{(\hat{p}_{r2l}-\bar{p}_{r2.})^2}{\frac{\bar{p}_{r2.}(1-\bar{p}_{r2.})}{kn_{r.l}}} & \cdots & \sum_{l=1}^{m} \frac{(\hat{p}_{rrl}-\bar{p}_{rr.})^2}{\frac{\bar{p}_{rr.}(1-\bar{p}_{rr.})}{kn_{r.l}}} & \end{array}$$

$$\therefore \chi 2 = \begin{pmatrix} \chi_{11}^2 & \cdots & \chi_{1r}^2 \\ \vdots & \ddots & \vdots \\ \chi_{r1}^2 & \cdots & \chi_{rr}^2 \end{pmatrix}.$$

The above matrix of chi-squares can also be called as element-chi-square-matrix. From this matrix we basically can test three types of hypotheses which are as follows:

(i) $H_0 : p_{ij1} = p_{ij2} = \ldots = p_{ijm}$; or, the hypothesis of testing the equality of the each individual ((i,j)$^{\text{th}}$) transition probability of the multiple (*m*) population transition probability matrices $P_1, P_2\ , \ldots., P_m$ for all values of $i, j = 1, 2, \ldots, r$.

(*ii*) $H_0 : (p_{i11}\quad p_{i21}\quad \cdots\quad p_{ir1}) = (p_{i12}\quad p_{i22}\quad \cdots\quad p_{ir2}) = \cdots = (p_{i1m}\quad p_{i2m}\quad \cdots\quad p_{irm})$; or, the hypothesis of checking the equality of the i-th row vector of all population transition probability matrices $P_1, P_2\ , \ldots., P_m$ for all values of $i = 1, 2, \ldots, r.$. Actually, it tests the equity of the frequentness of the transition of the random movement of multiple population sequences from each state to all states.

(*iii*) $H_0 : P_1 = P_2\ = \cdots . = P_m$; or the hypothesis of testing the equity of the total transitions for all population sequences. It tests the similarity of multiple population sequences or whether the *m* sample sequences are drawn from same population.

For the aforementioned tests the concern test statistics are given below respectively.

(i) Comparing each χ_{ij}^2 ($\forall\, i, j = 1, 2, \ldots, r$) with the tabulated $\chi^2_{(m-1,.\propto)}$ of (m-1) degree of freedom,

(ii) Comparing each $\sum_{j=1}^{r} \chi_{ij}^2$ ($\forall\ i = 1, 2, \ldots, r$) with the tabulated $\chi^2_{[\mathrm{r(m-1)-1,.\propto}]}$ of [r(m-1)-1] degrees of freedom,

(iii) Comparing Chi-squares' matrix sum $= \chi_{11}^2 + \cdots + \chi_{1r}^2 + \cdots + \chi_{r1}^2 + \cdots + \chi_{rr}^2$ with the tabulated $\chi^2_{(r(r\mathrm{m}-\mathrm{r}-1),.\propto)}$ of $[r(r\mathrm{m}-\mathrm{r}-1)]$ degrees of freedom.

3. REAL LIFE EXAMPLE

Since it is stated that in a Markov process all possible states and transitions have been assumed in such a way that there is always a next state and the process goes on forever; the characteristics of the DNA, the basic genetic material in living organisms and having a double standed-helical structure each of which is consisting of very long sequence from four letters/alphabets (nucleotides), *a, g, c,* and *t* (for adenine, guanine, cytosine, and thymine, respectively), sequence that undergoes the change within any population over the course of many generations, as random mutations arise and become fixed in the population can easily be treated as a Markov Chain. If two sequences from different organisms are similar, there may have been a common ancestor sequence, and the sequences are then defined as being homologous. The alignment indicates the changes that could have occurred between the two homologous sequences and a common ancestor sequence during evolution. So, a common gauge is to check whether the two sequences show significant similarity, to assess, for example, whether they have a remote common ancestor. As a result, sequence alignment is one of the most important techniques to analyze biological system.

Suppose we have three small DNA sequences such as those in the book of 'Statistical Methods in Bioinformatics' by Ewens, W. *et al* (2004), 30 pairs of sample sequences from same species have been considered. The average transition frequency matrices cum average transition probability matrices (one average transition probability matrix has been obtained from the 30 sample sequences accessed first population, another average transition probability matrix form 30 sample sequences of second population and the third average transition probability matrix from 30 sample sequences collected from the third population) are estimated as follows:

$$\hat{P}_1 = \begin{array}{c c} & \begin{array}{cccc} a & t & c & g \end{array} \\ \begin{array}{c} a \\ t \\ c \\ g \end{array} & \begin{pmatrix} 0.19 & 0.17 & 0.16 & 0.47 \\ 0.20 & 0.03 & 0.22 & 0.56 \\ 0.38 & 0.34 & 0.19 & 0.09 \\ 0.27 & 0.11 & 0.29 & 0.33 \end{pmatrix} \end{array}; \hat{P}_2 = \begin{array}{c c} & \begin{array}{cccc} a & t & c & g \end{array} \\ \begin{array}{c} a \\ t \\ c \\ g \end{array} & \begin{pmatrix} 0.34 & 0.21 & 0.26 & 0.19 \\ 0.11 & 0.15 & 0.26 & 0.49 \\ 0.22 & 0.39 & 0.28 & 0.11 \\ 0.18 & 0.25 & 0.13 & 0.45 \end{pmatrix} \end{array}; \hat{P}_3 = \begin{array}{c c} & \begin{array}{cccc} a & t & c & g \end{array} \\ \begin{array}{c} a \\ t \\ c \\ g \end{array} & \begin{pmatrix} 0.09 & 0.29 & 0.32 & 0.30 \\ 0.14 & 0.13 & 0.33 & 0.40 \\ 0.27 & 0.32 & 0.32 & 0.10 \\ 0.14 & 0.14 & 0.30 & 0.42 \end{pmatrix} \end{array}$$

We first want to observe the properties of three average transition probability matrices to judge the comparability of them as well as the samples. As such the following calculations have been performed.

3.1 Comparability of the three matrices

From the transition probability graphs of the matrix $\hat{P}_1$ we can conclude that it's all the states are recurrent because all the states are accessible to each other and they are communicating class and the number of states is finite. The matrices $\hat{P}_2$, $\hat{P}_3$ give the same result. The random walks for the three types of sequences have been observed from where the suspect of the difference among the sequences is evident. The Eigen values and vectors of the transition probability matrices have been observed. One of the Eigen values of the 2nd matrix and two of the Eigen values of the 1st as well as 3rd matrices are negative whereas the maximum Eigen values of the three matrices are 1.010, 0.944 and 0.922

respectively. So we can say that there is difference among the transition probabilities of the tree types of samples. Determinant of the matrices are -0.002, -0.007 and 0.001. The ranks of them are same (loosely 4) which is a sign of justification of comparing the three matrices. The stationary probabilities are given as the solution of the equations $\pi_1 = 0.19\pi_1 + 0.20\pi_2 + 0.38\pi_3 + 0.27\pi_4$, $\pi_2 = .17\pi_1 + 0.03\pi_2 + 0.34\pi_3 + 0.11\pi_4$, $\pi_3 = 0.16\pi_1 + 0.22\pi_2 + 0.19\pi_3 + 0.29\pi_4$, $\pi_4 = 0.47\pi_1 + 0.56\pi_2 + 0.09\pi_3 + 0.33\pi_4$ and $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$.

Similarly for the second sample we get five equations solving those we obtain the solutions of the stationary probabilities. For the first types of samples the limiting probabilities are 0.26, 0.16, 0.22, 0.35; for the second types of samples 0.20, 0.25, 0.22, 0.33 and for the third types of samples 0.17, 0.22, 0.32, 0.29 respectively. To test the hypothesis of equality of the stationary probabilities for the samples the null hypothesis can be expressed as

$$H_0: \pi_{i1} = \pi_{i2} = \pi_{i3}$$

where, $\pi_{i1}, \pi_{i2}$and π_{i3} ($\forall\, i = 1, 2, 3,4$) are the stationary probabilities of *i*th state for the 1st , 2nd and 3rd average transition probability matrices respectively. The test statistic for the aforementioned test is

$$\sum_{l=1}^{3} \frac{(\hat{\pi}_{il} - \bar{\pi}_{i.})^2}{\frac{\bar{\pi}_{i.}(1-\bar{\pi}_{i.})}{kn_{i.l}}};$$

$\forall i = 1, 2, 3, 4$; where $\bar{\pi}_{i.} = \frac{\pi_{i1}n_{i.1} + \pi_{i2}n_{i.2} + \pi_{i3}n_{i.3}}{n_{i.1} + n_{i.2} + n_{i.3}}$;

which is distributed as chi-square with (3-1) degree of freedom. The result of equality tests gives the p-values of the aforementioned chi-square statistic as 0.133, 0.248, 0.048 and 0.392. As such at 1% level of significance the limiting probabilities for the same state for the three types of samples are similar. So, for the long run the randomness visit of the population sequence to the individual state or nucleotide is similar for all states over the three populations. Therefore, from the aforementioned results it seems to us that the three matrices are compare able.

3. 2 *Proposed approach*

According to the alternative approach, the chi-square matrix will be:

$$\chi 2 = \begin{array}{c|cccc} & a & t & c & g \\ a & 24.92 & 7.10 & 10.72 & 28.46 \\ t & 4.70 & 10.38 & 5.06 & 6.60 \\ c & 10.89 & 2.11 & 7.49 & 0.51 \\ g & 13.00 & 19.33 & 27.19 & 7.56 \end{array}$$

The tabulated value of chi – square at 1% level of significance with 2 degree of freedom is 9.21. There is one calculated value for each of the 16 chi-square test statistics for 16 types of transitions in the matrix of chi-squares. For the first transition (from adenine to adenine), the calculated value (= 24.92) of chi-square test statistic is greater than the tabulated value (= 9.21) which means the null hypothesis $H_0: p_{aa1} = p_{aa2} = p_{aa3}$ is rejected at 1 percent level of significance. So, we conclude that the probability of three population sequences for the transition from adenine to adenine is not similar and we denote the dissimilarity by a notation "DS". Again for the transition (from thymine to adenine), the null hypothesis $H_0: p_{ta1} = p_{ta2} = p_{ta3}$ is accepted at the same level of significance with a *p*- value of 0.10. It can be inferred that the frequentness of three population sequences for the transition from thymine to adenine is similar and we denote the similarity by a notation “S”. So the resultant decision matrix for the 16 various transitions is given below:

$$\text{the resultant decision matrix} = \begin{pmatrix} DS & S & DS & DS \\ S & DS & S & S \\ DS & S & S & S \\ DS & DS & DS & S \end{pmatrix}.$$

Moreover, the calculated value of overall chi – square, the sum of all individual chi-squares of the chi-squares’ matrix sum, is obtained as 186.009. Therefore, the null hypothesis $H_0\colon P_1 = P_2 = P_3$ of the equality of the entire transition probability matrices of three population sequences is rejected at 1 % level of significance (since the tabulated value of the chi-squares matrix sum with 27 degrees of freedom is 46.96). So, with an overall point of view it can be concluded that the two population sequences are dissimilar or do not belong to the same ancestor. Moreover, the row similarity can be found here. The sum of chi- squares for the 1st, 2nd 3rd and 4th rows are calculated as 71.19, 26.73, 21.01 and 76.09 respectively. The tabulated value of the row wise sum of chi-squares with 7 degree of freedom is 18.48 at 1 % level of significance. So, all rows are significantly varying among themselves for the three population sequences. The dissimilarity among all of the rows of the three transition probability matrices is also a potential evidence of ensuring the conclusion that the three population sequences are dissimilar.

4. ADVANTAGES

For the given example it is observed that the *p* - value of the proposed test is close to zero (since the *p*-values for the chi-square test is 10^{-25}) indicating bold rejection of the null hypothesis of the equality of the transition probability matrices whereas the samples were really drawn from three different populations. Therefore, the performance of the alternative method seems better.

The authors also checked the results of the proposed multiple sequence alignment with those obtained by combining the 3 pair-wise sequence alignments (3 pair for three populations) for the aforementioned samples (30 sample sequences drawn from each of the three populations). The 3 pair-wise sequence alignments test better (since the equality of the entire transition probability matrices of the three population sequences is rejected with a lower *p*-value of 10^{-36}). However, the alternative multiple sequence alignment method will be more amiable since it requires relatively less effort and time.

Besides, the alternative method is not affected by natural gap in the one or more sequences for the multiple sequence alignment. So, there is no need of penalization for a natural gap or even an artificial controversial gap.

The alternative method measures the comparison among the random frequentness of the individual or group-wise or entire transitions for multiple sequences rather than accumulating the distances between or among the similar positioned individual nucleotides of multiple sequences.

Unlike previously suggested multiple sequence alignment procedures, the proposed alternative approach gives not only an overall decision of the significant similarity/dissimilarity of the multiple population sequences but also the significant similarity/dissimilarity of all possible transitions. It clearly identifies the possible dissimilarity among all population sequences. The current method specifically detects for which transition(s) the overall dissimilarity for the multiple population sequences is being pragmatic. This idea of more specification can help the biotechnologist to quickly detect the core fact of the possible difference among bio-organisms more easily and more efficiently.

CONCLUDING REMARKS

Sequence alignment has been widely being studied by numerous authors. Most of them suggested their methods more complex by introducing algorithm's accuracy level. Our method quantifies the degree of disorderness among the transition probabilities of nucleotides of multiple sequences, and reduces the loss of comparison among the multiple sequences from the multiple unknown populations. It also ensembles the individual, group wise and overall transitions pattern of one type of sequences whether significantly differing from other types of sequences. Advanced multiple comparison test for the multiple sequence alignment can be the further scope of the proposed heuristic. Any inquiry and prove(s) of the mathematical development of the alternative approach can be accessible from author on demand. The further scope of the multiple sequence alignment is to find multiple comparisons of the multiple sequences after inferring that the multiple sequences come from multiple populations.

REFERENCES

Adnan M. A. S., Moinuddin, M, Roy, S, *et al*. (2011). An Alternative Approach of Pair-wise sequence Alignment. Proceedings. JSM 2011, American Statistical Association, p2941 - 2951.

Allison, L. and Wallace, C. S. (1994) The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments. *J. Mol. Evol*. **39,** 418–430.

Anfinsen Cb (1973). "Principles that govern the folding of protein chains". Science 181 (96): 223–230. doi:10.1126/science.181.4096.223. PMID 4124164.

Armougom, F., Moretti, S., Keduas, V., and Notredame, C. (2006) The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics* **22,** e35–39.

Arribas-Gil, A., Gassiat, E., and Matias, C. (2006) Parameter estimation in pairhidden Markov models. *Scand. J. Stat*. **33,** 651–671.

Baldi, P. and Chauvin, Y. (1994) Smooth on-line learning algorithms for hidden Markov models. *Neural Comput*. **6,** 307–318.

Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. A. (1994) Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA* **91,** 1059–1063.

Barrett, C., Hughey, R., and Karplus, K. (1997) Scoring hidden Markov models. *Comput. Appl. Biosci*. **13,** 191–199.

Bergeron, B. (2003). Bioinformatics Computing. Prentice Hall Publisher.

Bhat, U. N. (1972). Elements of Applied Stochastic Process, Wiley & Sons, Canada.

Bishop, M. J. and Thompson, E. A. (1986) Maximum likelihood alignment of DNA sequences. *J. Mol. Biol*. **190,** 159–165.

Blackshields G, Wallace I, Larkin M, Higgins D (2006): Analysis and comparison of benchmarks for multiple sequence alignment. In Silico Biology, 6(4)**:**321-339.

Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996) A flexible motif search technique based on generalized profiles. *Comput. Chem*. **20,** 3–23.

Carugo, O. and Pongor, S. (2001) A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci*. **10,** 1470–1473.

Chew, L. P., Huttenlocher, D., Kedem, K., and Kleinberg, J. (1999) Fast detection of common geometric substructure in proteins. *J. Comput. Biol*. **6,** 313–325.

Chuong, B. Do, Kathoh, K. (2008). Protein Multiple Sequence Alignment. Methods in Molecular Biology, vol. 484, DOI: 10.1007/978-1-59745-398-1.

Crooks G, Green R, Brenner S (2005): Pairwise alignment incorporating dipeptide covariation. Bioinformatics, 21(19)**:**3704.

Dannemann, J And Holzmann, H (2007). The likelihood ratio test for hidden Markov models in two-sample problems. Comp. Stat. & Dtat Analysis. V 52, P:1850 -1859.

Dayhoff, M. O., Eck, R. V., and Park, C. M. (1972) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.). National Biomedical Research Foundation,Washington, DC, pp. 89–99.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978)A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.). National Biomedical Research Foundation,Washington, DC, pp. 345–352.

Do, C. B., Gross, S. S., and Batzoglou, S. (2006) CONTRAlign: discriminative training for protein sequence alignment. RECOMB.

Do, C. B., Mahabhashyam, M. S., Brudno, M., and Batzoglou, S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*. **15,** 330–340.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1999) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Do C. B. et al. (2005): ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res, 15:330-340.

Eddy, S. R. (1995) Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. **3,** 114–120.

Eddy, S. R. HMMER: a profile hidden Markov modeling package, available from http://hmmer.janelia.org/.

Eddy, S. R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol*. **6,** 361–365. 179. Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics* **14,** 755–763.

Edgar, R. C. and Sjolander, K. (2004) COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics* **20,** 1309–1318.

Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32,** 1792–1797.

Edgar, R. C. and Sjolander, K. (2003) Simultaneous sequence alignment and tree construction using hidden Markov models. *Pac. Symp. Biocomput*. 180–191.

Edgar, R. C. and Sjolander, K. (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models. *Bioinformatics* **19,** 1404–1411.

Eidhammer, I., Jonassen, I., and Taylor, W. R. (2000) Structure comparison and structure patterns. *J. Comput. Biol*. **7,** 685–716.

Ewens, W. And Grant, G. R. (2004). Statistical Methods in Bioinformatics. Springer.

Fleissner, R., Metzler, D., and von Haeseler, A. (2005) Simultaneous statistical multiple alignment and phylogeny reconstruction. *Syst. Biol*. **54,** 548–561.

Grundy, W. N., Bailey, T. L., Elkan, C. P., and Baker, M. E. (1997) Meta-MEME: motif-based hidden Markov models of protein families. *Comput. Appl. Biosci*. **13,** 397–406.

Hanus, P. et al (2009). Source scoring scheme for Multiple Sequence Alignments. Web.

Hara, T, Sato, K, Ohya, M. (2010). Pairwise sequence alignment algorithm by a new measure based on transition probability between two consecutive pairs of residues. BMC Bioinformatics. 11: 235.

Hein, J. (2001) A generalisation of the Thorne-Kishino-Felsenstein model of statistical alignment to k sequences related by a binary tree. PSB.

Hein, J., Jensen, J. L., and Pedersen, C. N. (2003)Recursions for statistical multiple alignment. *Proc. Natl. Acad. Sci. USA* **100,** 14960–14965.

Hein, J., Wiuf, C., Knudsen, B., Moller, M. B., and Wibling, G. (2000) Statistical alignment: computational properties, homology testing and goodness-of-fit. *J.Mol. Biol*. **302,** 265–279.

Hein, J. (1990) Unified approach to alignment and phylogenies.*Methods Enzymol*. **183,** 626–645.

Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.

Holmes, I. and Bruno, W. J. (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* **17,** 803–820.

Holmes, I. (2003) Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics* **19(Suppl 1),** i147–157.

Hongwei, et al (2008). A Hybrid Algorithm Based on Artificial Immune System and Hidden Markov Model for Multiple Sequence Alignment. Web.

Huang J, Zhou W, Watson AM, et al. (2008) Efficient ends-out gene targeting in Drosophila. Genetics 180(1): 703-707.

Hughey, R. and Krogh, A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biosci*. **12,** 95–107.

Jensen, J. L. and Hein, J. (2005) Gibbs sampler for statistical multiple alignment. *Stat. Sin*. **15,** 889–907.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci*. **8,** 275–282.

Karplus, K., Barrett, C., and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14,** 846–856.

Karplus, K. and Hu, B. (2001) Evaluation of protein multiple alignments by SAMT99 using the BAliBASE multiple alignment test set. *Bioinformatics* **17,** 713–720.

Kececioglu, J. and Kim, E. (2007) Simple and fast inverse alignment. RECOMB.

Knudsen, B. and Miyamoto, M. M. (2003) Sequence alignments and pair hidden Markov models using evolutionary history. *J. Mol. Biol*. **333,** 453–460.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling.Thorne, J. L., Kishino, H., and Felsenstein, J. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol*. **33,** 114–124.

Krogh, A. (1998) An introduction to hidden Markov models for biological sequences. In *Computational Methods in Molecular Biology* (Salzberg, S.,Searls, D., Kasif, S., eds.). Elsevier Science, St. Louis, MO, pp. 45–63.

Liu, J. S., Neuwald, A. F., and Lawrence, C. E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc*. **90,** 1156–1170.

Loytynoja, A. and Milinkovitch, M. C. (2003) A hidden Markov model for progressive multiple alignment. *Bioinformatics* **19,** 1505–1513.

Loytynoja, A. and Goldman, N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. USA* **102,** 10557–10562.

Lunter, G. A., Miklos, I., Song, Y. S., and Hein, J. (2003) An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comput. Biol*. **10,** 869–889.

Lunter, G., Miklos, I., Drummond, A., Jensen, J. L., and Hein, J. (2005) Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinform*. **6,** 83.

Lu Y, Sze S (2009): Improving accuracy of multiple sequence alignment algorithms based on alignment of neighboring residues**.** Nucleic Acids Research, 37(2)**:**463.

Mamitsuka, H. (2005) Finding the biologically optimal alignment of multiple sequences. *Artif. Intell. Med*. **35,** 9–18.

McClure, M. A., Smith, C., and Elton, P. (1996) Parameterization studies for the SAM and HMMER methods of hidden Markov model generation. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. **4,** 155–164.

Metzler, D., Fleissner, R.,Wakolbinger, A., and von Haeseler, A. (2001) Assessing variability by joint sampling of alignments and mutation rates. *J. Mol. Evol*. **53,** 660–669. *J. Mol. Biol*. **235,** 1501–1531. 176.

Metzler, D. (2003) Statistical alignment based on fragment insertion and deletion models. *Bioinformatics* **19,** 490–499.

Miklos, I. and Toroczkai, Z. (2001) An improved model for statistical alignment. WABI. 156. Miklos, I. (2003) Algorithm for statistical alignment of sequences derived from a Poisson sequence length distribution. *Disc. Appl. Math*. **127,** 79–84.

Miklos, I., Lunter, G. A., and Holmes, I. (2004) A “Long Indel” model for evolutionary sequence alignment. *Mol. Biol. Evol*. **21,** 529–540.

Miklos, I. (2002) An improved algorithm for statistical alignment of sequences related by a star tree. *Bull. Math. Biol*. **64,** 771–779.

Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998) HOMSTRAD: a database of

Mount, D. W. (2003). Bioinformatics. CBS Publishers.

protein structure alignments for homologous families. *Protein Sci*. **7,** 2469–2471.

Muller, T. and Vingron, M. (2000) Modeling amino acid replacement. *J. Comput. Biol*. **7,** 761–776.

Notredame, C. (2002) Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics* **3,** 131–144.

Nuin, P. A., Wang, Z., and Tillier, E. R. (2006) The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinform*. **7,** 471.

O’Sullivan, O., Zehnder, M., Higgins, D., Bucher, P., Grosdidier, A., and Notredame, C. (2003) APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics* **19(Suppl 1),** i215–221.

Pang, A., Smith, A. D., Nuin, P. A., and Tillier, E. R. (2005) SIMPROT: using an empirically determined indel distribution in simulations of protein evolution. *BMC Bioinform*. **6,** 236.

Pena, V. D. L. (2007). Multiple hypotheses testing of transition matrices. Journal of Risk Model Validation. V 1 (3), 69-76.

Pei, J. and Grishin, N. V. (2006) MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. *Nucleic Acids Res*. **34,** 4364–4374.

Prakash, A. et al (2009). Assessing the discordance of MSA. IEEE/ACM Transactions on Computaional Biology and Bioinformatics. V 6, xx, xxxxxxx 2009.

Prlic, A., Domingues, F. S., and Sippl, M. J. (2000) Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng*. **13,** 545–550.

Raghava, G. P., Searle, S. M., Audley, P. C., Barber, J. D., and Barton, G. J. (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinform*. **4,** 47.

Rajasekaran, S et al (2004). A Randomized Algorithm for Distance Matrix Calculations in Multiple Sequence Alignment. KELSI 2004, LNAI 3303, 33-45, @ Springer-Verlag Berlin Hidelberg 2004.

Redelings, B. D. and Suchard, M. A. (2005) Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol*. **54,** 401–418.

Reese, J. T. and Pearson, W. R. (2002) Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics* **18,** 1500–1507.

Roshan, U. and Livesay, D. R. (2006) Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* **22,** 2715–2721.

Sauder, J. M., Arthur, J. W., and Dunbrack, R. L., Jr. (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* **40,** 6–22.

Sabath, N., Landan, G., Graur, D (2008). A Method for the Simultaneous Estimation of Selection Intensities in Overlapping Genes. PLoS ONE.

Samudrala R, K (2006): Incorporating background frequency improves entropy-based residue conservation measures**.** BMC bioinformatics, 7:385.

Sahraeian, S. M. E et al (2011). PicXAA-R: Efficient structural alignment of multiple RNA sequences using a greedy approach. BMC Bioinformatics. 12, (1). http://www.biomedcentral.com/1471-2105/12/S1/S38

Schwartz, A. S., Myers, E., and Pachter, L. (2006) Alignment metric accuracy. *arXiv 2006:q-bio.QM/0510052*.

Singh S, Kumer, G. S., Anuradaha, N. et al. (2010) Comparative Modeling Study of the 3-D Structure of Small Delta Antigen Protein of Hepatitis Delta Virus. J Comput Sci Syst Biol 3: 001-004.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., *et al.* (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci*. **12,** 327–345.

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*. **26,** 320–322.

Stoye, J., Evers, D., and Meyer, F. (1998) Rose: generating sequence families. *Bioinformatics* **14,** 157–163.

Steel, M. and Hein, J. (2001) Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star-shaped tree. *Appl. Math. Lett*. **14,** 679–684.

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins* **61,** 127–136.

Thompson, J. D., Plewniak, F., and Poch, O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15,** 87–88.

Thompson, J. D., Plewniak, F., and Poch, O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*. **27,** 2682–2690.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1992) Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol*. **34,** 3–16.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005) Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res*. **6,** 1453–1484.

Van Walle, I., Lasters, I., and Wyns, L. (2005) SABmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics* **21,** 1267–1268.

Vingron, M. and von Haeseler, A. (1997) Towards integration of multiple alignment and phylogenetic tree construction. *J. Comput. Biol*. **4,** 23–34.

Viterbi, A. J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory* **It13,** 260.

Whelan, S. and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol*. **18,** 691–699.

Yu, C.-N., Joachims, T., Elber, R., and Pillardy, J. (2007) Support vector training of protein alignment models. RECOMB.

Zhu, J., Liu, J. S., and Lawrence, C. E. (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14,** 25–39.