A New Look at Tree Models for Multiple Sequence Alignment[*]

Dannie Durand[†]

September 9, 1997

Abstract

Evolutionary trees are frequently used as the underlying model in the design of algorithms, optimization criteria and software packages for multiple sequence alignment (MSA). In this paper, we reexamine the suitability of trees as a universal model for MSA in light of the broad range of biological questions that MSA's are used to address. A tree model consists of a tree topology and a model of accepted mutations along the branches. After surveying the major applications of MSA, examples from the molecular biology literature are used to illustrate situations in which this tree model fails. This occurs when the relationship between residues in a column cannot be described by a tree; for example, in some structural and functional applications of MSA. It also occurs in situations, such as lateral gene transfer, where an entire gene cannot be modeled by a unique tree. In cases of nonparsimonous data or convergent evolution, it may be difficult to find a consistent mutational model. We hope that this survey will promote dialogue between biologists and computer scientists, leading to more biologically realistic research on MSA.

1 Introduction

Multiple sequence alignment (MSA) is important in functional, structural and evolutionary studies of sequence data. Much research has focussed on the formal study of MSA as an optimization problem and several optimization criteria have been discussed at length in the literature [8, 34, 42, 44, 58, 65, 75, 85]. In addition, many software tools for constructing MSA's are available, mostly based on heuristics although some use exact or branch-and-bound techniques (see [16, 54] for surveys.) The concept of an evolutionary tree is a widely used model for MSA, where the tree encodes the historical relationships between the modern sequences in the alignment. Tree models have been used to construct column scoring functions for optimization criteria. Trees have also been used as implicit or explicit structures in the design of algorithms and heuristics.

The design of MSA algorithms is based on assumptions about the application and the nature of the sequence data. While these assumptions must, of necessity, be abstractions of reality, a better understanding of the true nature of MSA applications and data will lead to better algorithm design. As the amount and diversity of sequence data grows, MSA is being applied to a wide variety of

[*] to appear in DIMACS Special Year volume "Mathematical Support for Molecular Biology", M. Farach, F. Roberts and M. Waterman, Eds, AMS, 1997.

[†] Computational Biology Group, University of Pennsylvania, 3401 Walnut Street, Suite 300C, Philadelphia, PA 19104-6228 (*durand@saul.cis.upenn.edu*) and DIMACS, Rutgers University, Piscataway, NJ 08855-1179. Supported by NSF Grants BIR-94-13215 A01 and BIR-94-12594.

biological questions. Although the tree model is frequently viewed as a universal model for MSA, for some applications it is not compelling. In addition, some data sets cannot be modeled by a unique tree.

In this paper, the biological applications of MSA are reviewed. Examples from the biology literature are used to illustrate cases where the tree model is not appropriate for multiple alignments. First, multiple sequence alignment is introduced as an optimization problem in Section 2 and optimization criteria and complexity results are reviewed. Next, in Section 3, the major applications of MSA are surveyed. A discussion of data sets where the tree model fails is presented in Section 4.

2 Multiple Sequence Alignment

First we provide a brief introduction to the multiple sequence alignment problem and its complexity, optimization criteria used to evaluate alignments and algorithmic approaches to solving it. This is not a comprehensive survey of multiple sequence alignment algorithms. Surveys of multiple sequence alignment algorithms can be found in [16] and in the introduction to [58]. An experimental comparison of multiple sequence algorithms is described in [54].

Multiple sequence alignment involves lining up a group of sequences to reveal similarities shared across the group. A sequence is a string of symbols chosen from an alphabet, Σ, where $\Sigma = \{A, C, G, T\}$ for nucleic acid sequences and Σ contains the twenty amino acids for protein sequences. In a *pairwise alignment* of two sequences, S_i and S_j, the sequences are lined up one on top of the other so that each symbol in S_i is paired with a symbol in S_j in a series of columns of height two. Blanks may be inserted into either sequence or at the ends of either sequence, so that a symbol may be paired with another symbol or with a blank. These blanks represent mutations in the form of insertions or deletions. Since it is impossible to tell whether a symbol was inserted in one sequence or deleted from the other, blanks are also referred to as "indels". Each column in the alignment contains a match, a mismatch or an indel. Metrics for associating a cost or similarity score with two paired symbols are discussed below. We will designate the cost of two symbols, x and y, as $\delta(x, y)$, where x and $y \in \Sigma \cup \{\_\}$, where "_" is the blank character.

A *global* alignment seeks to line up the two sequences so that the similarity in the alignment as a whole is maximized. In a *local* alignment, we seek a subsequence from each sequence such that when the two are aligned they yield the highest scoring region. For local alignments the average score must be negative, otherwise the longer regions would score above shorter regions by virtue of their length alone, regardless of the degree of similarity between the residues [3].

Multiple sequence alignment is an extension of pairwise alignment to more than two sequences. In this case, k sequences are lined up, inserting indels as necessary into any of the k sequences, to obtain a sequence of columns of k symbols. Again, either a global alignment, maximizing the similarity over all columns together, or a local alignment, k aligned subsequences, one from each sequence, that yield the lowest cost region, may be sought.

Each column in a multiple sequence alignment captures a shared relationship between the residues in the column. The relationship sought depends on the application of alignment. For example, if the multiple alignment is used to illustrate evolutionary relationships, then the residues in each column are assumed to have a shared evolutionary history. If the multiple alignment is used to determine structure or function, then the residues in each column should have a shared structural or functional rôle. Trying to find the best alignment is equivalent to finding the alignment that correctly captures this relationship in each column.

We attempt to express this by associating a score with each column that expresses how well matched the residues in the column are. We then seek the multiple alignment with the best score, where the score of the MSA is the sum of the scores of its columns. Formally, we can express multiple sequence alignment as an optimization problem. We are given sequences $S_1, \cdots S_k$ of length $N_1, \cdots N_k$, where $S_i = s_{i_1} \ldots s_{i_{N_i}}$, $s_{i_j} \in \Sigma$. We seek a matrix $A = \{a_{ij}\}$ where $a_{ij} \in \Sigma \cup \{\_\}$ and eliminating the indels from any row A_j gives the sequence S_j, such that the cost of the alignment

$$D(A) = \sum_j d(a_{1j} \ldots a_{kj})$$

is minimum. The choice of the column scoring function, d, should reflect the application of the alignment. Below, we review commonly used scoring functions and consider for which applications each is most appropriate.

Column Scores: Three metrics commonly used to evaluate each column, $d(a_{1j} \ldots a_{kj})$, are sum-of-pairs (SP), tree alignment (TA) and star alignment (SA). The sum-of-pairs cost [7, 56] of a column $a_{1j} \ldots a_{kj}$ is the sum of the costs of all unordered pairs in the column

$$d_{SP}(a_{1j} \ldots a_{kj}) = \sum_{p<q} \delta(a_{pj}, a_{qj}),$$

where $\delta(x, y)$ is the cost associated with aligning two symbols x and y in a pairwise alignment. This definition is mathematically natural but not biologically intuitive.

Tree alignment [65, 66] is based on the assumption that the residues in the columns of the multiple sequence alignment share an evolutionary history and that this history can be expressed as a tree. Under this model, a column is scored by computing the cost of the underlying tree. The score of the tree expresses the likelihood, under some model of evolution, that these residues are related in the manner described by this tree.

In order to use this approach, several issues must be resolved. First, a tree topology is needed. In general, the underlying tree is not known. In fact, multiple sequence alignments are generally used to estimate evolutionary trees and not vice versa. Second, it is generally assumed that every column in the alignment has the same underlying tree topology. As we shall see below, this may not always be the case. Third, in order to compute the branch costs of the tree, we need to know the ancestral sequences associated with the internal nodes. Fourth, given a tree topology with ancestral sequences at the nodes, a cost must be associated with each branch of the tree. This implies an underlying model of evolutionary change. The appropriate model will vary with the data.

Let us consider how to infer the ancestral sequences and compute the branch costs when the topology is known. The k modern sequences are associated with the k leaves of the tree. Sequences for the internal nodes will be selected so that the tree is the best estimate of the true tree under some model of evolutionary change. The most common model is *maximum parsimony*, in which it is assumed that the true tree required a minimum number of evolutionary steps. Under this model, the cost of an edge (X_i, X_j) in the tree is the minimum number of mutations required to get transform sequence X_i into sequence X_j. That is, internal nodes are selected such that the sum of edge costs, i.e. the total number of mutations required along the branches of the tree, is minimimized. This assumption that evolution is parsimonious, is a subject of much debate.

The above approach assumes that the topology of the evolutionary tree is known, e.g. from morphological data. If the tree topology is not known, then all possible topologies of trees with k leaves must be considered to find the optimal topology. For each topology, the optimal ancestral sequences must be inferred. Since the number of unrooted, bifurcating trees with k leaves grows as

$$\frac{(2k-5)!}{2^{k-3}(k-3)!},$$

this approach rapidly becomes prohibitively expensive.

A variant on tree alignment is star alignment (SA), in which it is assumed that the underlying tree is a star. This implies that all sequences share a common ancestor. Restricting the topology makes this approach much more tractable and it has been used as an underlying structure in many algorithms and analyses (e.g., [5, 34]).

Altschul and Lipman [5] have shown that SP, TA and SA costs for the same column can be very different. The SP, TA and SA alignment costs for the column (AAAACC) are shown in Figure 1. SP overcounts mutations in a column because it considers that a separate mutation occurred between each pair in the column. In some sense, it is a “least parsimonious” model of mutation. In contrast, TA has been traditionally used with the maximum parsimony criterion. For most data sets the true answer is somewhere in between.

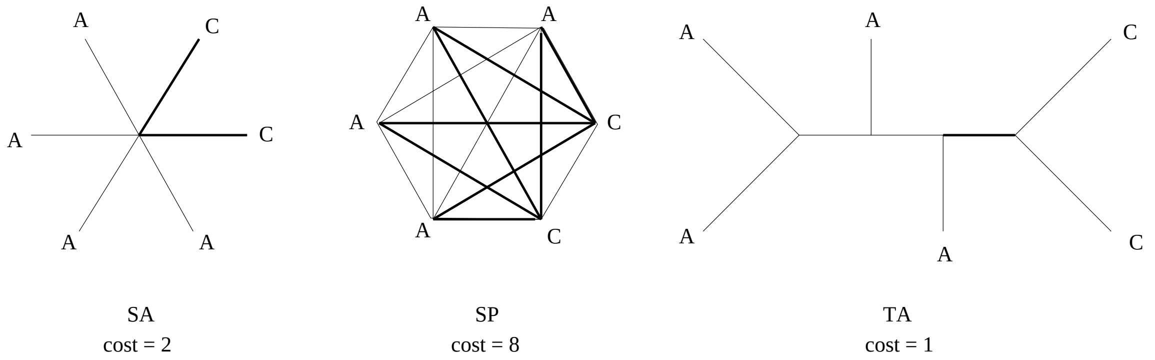

Figure 1: Comparing alignment costs.

Pairwise Costs: Regardless of which metric you choose, an underlying pairwise cost, $\delta(x,y)$ is needed. The simplest cost function for DNA is the edit distance,

$$\delta(x,y)=\begin{cases}0 & \text{if } x=y\\ s & \text{if } x\neq y\\ f & \text{if } x=\_ \text{ or } y=\_\end{cases}$$

where s and f can be chosen to tune the relative importance of substitutions and deletions. This model assumes that all possible nucleotide replacements are equally probable. More complex cost models (e.g., the Jukes-Cantor and Kimura two-parameter models [43, 45]) take variations in nucleotide frequencies into account, as well as the fact that the probability of substitution differs for each of the six possible pairs of nucleotides.

For proteins, substitution matrices are used. Substitution matrices can be based on observed substitution frequencies in sequences known to be related or on the chemical properties of the amino acid residues. The most widely used substitution matrices are the PAM matrices and their log-odds counterpart MDM_{78} [19, 68]. PAM matrices were constructed using observed amino acid substitutions in homologous sequences and are parameterized according to the degree of evolutionary divergence. One PAM (or accepted point mutation) corresponds to an average of one point mutation per 100 amino acid residues. MDM_{78} is a log-odds version of the PAM-250 matrix in which the probability of an amino acid substitution is normalized by the probability that the two residues could be aligned by chance. The BLOSSUM matrices, also based on observed substitution frequencies, were computed more recently from a larger set of homologous sequences [39]. It has been argued that when the objective of alignment is to obtain structural rather than evolutionary information, matrices based on biochemical properties such as hydrophobicity will give better alignments [6, 25, 48, 60].

A major issue in multiple sequence alignment algorithms is the ability to handle gaps; that is, sequences of one or more contiguous indels. The simplest model of insertions/deletions is to assign a fixed penalty to every indel. However, there is evidence [27] that better alignments can be obtained when the cost of beginning a gap is greater than the cost of continuing a gap. The simplest such cost is the affine gap function where the cost of a gap of length x, is $g + lx$, where g is the gap creation penalty and l is the gap extension penalty. More complex parameterized gap functions have been considered as well [35, 83]. In addition, gaps at either end of either sequence may or may not be penalized, depending on the application. In general, end gap penalties are inappropriate for local alignments, when the sequence lengths differ substantially or when the sequences may not be related [32]. Lesk *et al.* [52] and Barton and Sternberg [10] have suggested that the gap cost should be much higher in regions of known structural importance. Unfortunately, this approach can only be used if structural information is available.

Biological Metrics for MSA The methods described above use only primary sequence information to align sequences. However, structural information has also been used to guide the computation of MSA's and to validate them.

Some of the earliest work in structure-based alignments was presented by Lesk and Chothia [51], who used superposition of secondary structures to align globin sequences. Today there are two common approaches to structural alignment. First, one may associate a secondary structure type (α-helix, β-sheet or random coil) with each residue and then impose the additional constraint that only residues associated with the same type of secondary structure may be aligned (for example, [1, 62]). Second, structural alignments may be performed by minimizing the root mean square distance between the α-carbon molecules in the backbone (see [84, 81], for an example).

Structural information has also been used to set gap penalties. In separate studies, Lesk *et al.* [52] and Barton and Sternberg [10] observed that, due to selection constraints, indels seldom occur in structurally important regions. If structural information is available for at least one sequence in the alignment, gap penalties can be increased in regions encoding helices and strands of β sheets. Both groups showed that this approach led to substantially better alignments.

Structural and functional information has been used to validate multiple sequence alignments. In a comprehensive study of twelve different multiple alignment programs, McClure *et al.* [54] measured algorithm performance by computing a numerical score based on the ability to find known motifs in four different data sets. For all four data sets, supporting evidence for the motifs from structural or mutational studies was available. Barton and Sternberg [10] have also used structural alignments to validate sequence-based alignments.

The use of structural information to guide or validate alignments is only possible for data sets where structural information is available. Another consideration is whether a structural approach to MSA is useful in all cases. For example, as discussed in Section 4, residues that share a structural or functional rôle, do not always share an evolutionary history. Finally, in an experimental comparison of a number of structural alignment methods, Godzik [30] has pointed out that, in general, there is no unique structural alignment: different methods for comparing structures result in different alignments.

Complexity of Multiple Sequence Alignment. Complexity results have been published for MSA using both the sum-of-pairs and the TA metrics. Wang and Jiang [85] has shown that SP alignment is NP-complete and Sweedyk and Warnow [75] showed that general TA is NP-complete. TA for a fixed binary tree topology is also NP-complete [85] and a polynomial time approximation scheme has been presented [42]. However, when the fixed topology is a star, TA is max-SNP hard [85]. This means that no polynomial time approximation scheme can deliver an arbitrarily good approximation.

Exact methods for both SP and TA have been presented, although they are prohibitively expensive for more than a few sequences on the order of 100 base pairs in length. An optimal alignment of k species can be obtained using dynamic programming in $O(2^k N^k)$ evaluations of the cost function d using $O(N^k)$ space. An exact TA algorithm for a given tree topology has been presented by Sankoff [65]. This requires $O(M(2N)^k)$ steps where M is the number of internal nodes, k is the number sequences and N is the maximum length over all sequences. Algorithms that reduce the time to find an exact solution by pruning the search space using an upper bound on the cost of the alignment have been presented by Carillo and Lipman [13] and Kececioglu [44].

Approximation algorithms for both SP and TA have been presented. Gusfield [34] presented a $(2-2/k)$-approximation algorithm for SP alignment, while Pevsner [58] reduced the approximation factor to $(2-3/k)$. Subsequently, Bafna *et al.* [8] presented a $(2-l/k)$-approximation algorithm, for any fixed constant l. For tree alignment, Gusfield [34] showed that the cost of the minimum spanning tree of the sequences is at most twice the optimal cost, yielding a 2-approximation. A polynomial time approximation scheme, yielding a $(1+\epsilon)$-approximation for arbitrarily small ϵ has been presented by Jiang *et al.* [42] for a fixed topology.

The number of heuristics for multiple sequence alignment is too numerous to mention them all here. Most multiple sequence alignments fall into a few common approaches. Many heuristics compute pairwise alignments for all $\binom{k}{2}$ sequence pairs and then progressively merge the pairwise alignments. The methods differ in the strategies they use for merging, which include the clustering and consensus approaches and merging according to a tree. The resulting multiple alignment will depend on the order in which the pairwise alignments were merged. McClure *et al.* [54] also report that methods are sensitive to the size and composition of the data set. A second common heuristic is the use of consensus patterns. These are methods that look for regions that are highly similar in all sequences and then extend the alignment between the regions. Some methods that use statistical pattern matching in multiple sequence alignment include the Gibbs sampler [50] and hidden Markov models [47]. Multiple sequence alignment algorithms have been surveyed in Chan *et al.* [16] and the introduction to [58]. McClure *et al.* [54] compared the performance of twelve multiple sequence analysis tools experimentally.

3 Biological Applications of Multiple Sequence Alignment

Multiple sequence alignments are used alone to show relationships between sets of sequence data. However, they are also often used as input to other algorithms which further exploit the similarity relationships that MSA's encode. Multiple sequence alignment is intended to capture relationships between sequence elements, but the nature of the relationship depends on the application. "Related" can refer to evolutionary, structural or functional relationships and within these categories there are many further nuances. What constitutes a good alignment depends on how it is used.

3.1 Characterizing Patterns in Biopolymers

Multiple sequence alignments are frequently used to identify and characterize patterns in proteins and nucleic acids. Once a conserved region is found in an MSA, a unifying representation of that region is sought that can be used for human perusal or as input to another program, often a database search.

The simplest representation for a pattern describing a conserved region in a multiple alignment is a consensus sequence. In this case, a single character is used to describe each position in the alignment. This character can be either the most common residue found at that position or a more generic encoding such as 'R' for purines (A and G) and 'Y' for pyrimidines (C, T and U). Characters in bold face can be used to indicate that the residue is invariant and upper and lower case can be used to indicate whether the residue is strongly or weakly conserved at this position. The advantage of the consensus sequence representation is that it allows a complex pattern to be represented as a "one liner". The disadvantage is that a great deal of information is lost, since only one character per site is retained.

A more informative representation is the position weight matrix, which not only shows which residue is most common at a given position, but it also reveals how common it is and what other residues are present at that position. A position weight matrix is a $|\Sigma|$ by N matrix, M, in which each column represents a position in the alignment. The r^{th} entry in each column represents the relative probability of finding the r^{th} residue at that position. One possible formulation of a position weight matrix is

$$M_{r,p} = \frac{\sum_{s=1}^{k} w_s \delta_{r,s_p}}{\sum_{s=1}^{k} w_s},$$

where w_s is the weight of sequence s, s_p is the residue at position p in sequence s and δ is the Kronecker delta,

$$\delta_{ij} = \begin{cases} 1 & i = j, \\ 0 & i \neq j \end{cases} .$$

Logarithmic probabilities can also be used (see, for example, [33]). The weights, $\{w_s\}$, are assigned to the sequences to compensate for the fact that the sample of sequences in the MSA is often biased.

Because most biological research focuses on a few model organisms (e.g., humans, mice, *drosophila*, *C. elegans* and yeast), sequences in the data base are not uniformly distributed in a taxonomic sense. For a discussion of the importance of sequence weighting and a variety of ways to compute such weights see [4, 69, 80, 82].

The relative probabilities in the position weight matrix can be combined with a substitution matrix to obtain a profile, a scoring matrix that is specific to the pattern under consideration called a profile. A profile can be computed by multiplying a substitution matrix with the position weight matrix for the desired pattern:

$$P_{r,p} = \sum_{i=1}^{|\Sigma|} S_{r,i} \cdot M_{i,p},$$

where $S_{i,j}$ is a substitution matrix such as the Dayhoff [19, 68] or Blossom [39] matrices. In addition to a row for each residue, profiles generally have two rows for gaps, one for gap initiation and one for gap extension. This allows the fact that some columns may have many gaps to be reflected in the profile.

Statistical representations of patterns, such as position weight matrices and profiles, have been used to represent characteristic patterns found in biopolymers and as input to programs that search for instances of sequences containing these patterns. Some examples of the use of these representations in searching protein data bases include [33, 79, 80]. Statistical characterizations of sequence and structural motifs include Helix-Turn-Helix [49], the calcium-binding EF-hand [47] and libraries of motifs such as [9, 39, 70]. Statistical representations of patterns in nucleic acids have been used to characterize and search for regulatory regions in DNA (see, for example, [17, 26, 28, 72]).

MSA's are also used to identify subsequences for laboratory techniques, such as PCR and library screening, that isolate DNA fragments containing a particular pattern. In these methods, a short piece of DNA containing the desired pattern (the "probe" or "primer") is used to hybridize with fragments containing that pattern in the target DNA. A minimum degree of similarity, which depends on the stringency of the laboratory conditions, between the probe and target fragments is needed for hybridization to take place. For example, when new members of a known gene family are sought, an MSA is used to find a region that is conserved in all known members of the family. From the MSA, a consensus sequence is constructed that is close enough to all known members to hybridize with any of them. The hope is that this consensus sequence will also hybridize with the, as yet undiscovered, new members of the family. This is a use of MSA where homology is not of interest. Only similarity is important.

3.2 Phylogeny

A phylogeny describes the evolutionary history of taxa. Originally, a *taxon* referred to a species or group of species but now can also describe sub-organismal entities such as genes. Such histories are frequently expressed as *evolutionary trees*, branching processes that describe the inheritance relationships between the taxa. Traditionally, these relationships were determined by making morphological comparisons. Now that large quantities of molecular data are available, inheritance relationships can also be inferred from sequence information.

Traditionally, phylogenetic trees have been used to determine ancestral relationships between species. However, with the advent of molecular data, it is now possible to ask questions about the relationships between molecular entities including genes, proteins and regulatory elements. Because of gene duplications, exon shuffling and lateral displacements, the history of a gene may not be the same as the history of the species carrying it. With this in mind, both designers and users of MSA algorithms should focus carefully on what question they are trying to answer and which algorithm and data are most appropriate to answer it.

In an evolutionary tree, each leaf is associated with a modern day species. Branch points indicate points where species separated. These internal nodes are associated with ancestral species. If there is information available concerning the amount of evolutionary time that has passed between speciation events, then the tree can be rooted and lengths (e.g. times) associated with the branches.

A multiple sequence alignment captures the inheritance information needed to reconstruct the tree. In this case, the sequences in the MSA are the leaves of the tree and it is assumed that there is some underlying tree that generated this alignment. However, the inheritance information in the alignment is incomplete (since only modern day sequences are available), the data may be noisy and the MSA may be incorrect. With these limitations in mind, and given an assumption about how sequences evolve, we need an algorithm to reconstruct this tree. Most such algorithms fall into three categories: distance-based methods, character-based methods and maximum likelihood. A survey of tree reconstruction methods is given in Swofford and Olsen [76, 77]. Multiple sequence alignments are required as input for all of these methods. Biologists frequently use only a robust subset of the columns as input. In such cases, MSA's are used to determine which columns to represent unambiguous homology.

Distance-based phylogeny: Distance-based methods work by computing pairwise distances between taxa and fitting those distances to a tree. A distance matrix, M, is derived from an MSA by computing the distance between every pair of taxa in the alignment (not the minimum pairwise distance!). If these distances fit a tree, M is said to be additive and the tree is unique and can be reconstructed in $O(k^2)$ time [86]. In general, however, M is not additive. In that case, a tree is sought that "best approximates" the observed distances. Mathematically, we seek a tree T such that

$$\sum_{ij} \|M[ij] - d_{ij}^T\|^l \tag{1}$$

is minimized, where d_{ij}^T is the distance between taxa i and j and the norm, l, is usually 1, 2 or ∞. It has been shown that finding the optimal T is NP-hard for L_1, L_2 [18] and L_∞ [2]. In addition, approximation algorithms for this problem have been developed [2]. However, it is not clear that a tree that minimizes Equation 1 is a good approximation to the true tree. There are many reasons why distance matrices are not additive including noisy sequence data, poor multiple sequence alignments and the possibility that the underlying process that generated the data was not a tree to begin with.

Character-based phylogeny: In character-based approaches, a data set is treated as a set of characters. Each character can be in one of several states. Each species is specified by a state vector. Traditionally, characters were derived from morphological, behavioral or chemical data so states might include things like "Has wings?", "Breeds in water?" or the strength of an immunological reaction. With molecular data, each column in an MSA is a character. The possible states are the four bases for DNA data and the twenty amino acids for protein data. Thus, a species is characterized by the sequence elements found in each column in the alignment.

A tree can be associated with each character. The state associated with each species is assigned to a leaf in the tree. Ancestral character states are inferred from leaf nodes and associated with internal nodes. By associating a cost with every state change, we can associate a cost with the tree by summing, over all branches, the cost of the state changes along each branch. This per character cost can be summed over all characters to obtain a character-based cost for the tree.

Within this framework, it is possible to find the best estimate of the true tree given a model of evolutionary change. The most common model is *maximum parsimony*, in which it is assumed that the true tree required a minimum number of evolutionary steps. The most parsimonious tree is the lowest cost tree over all possible tree topologies for k species and all possible inferred ancestral sequences for a given topology. One problem with maximum parsimony is that it does not take convergent or parallel evolution, multiple state changes or reversals in character state into account. Thus it is not a suitable model for all data sets.

Maximum Likelihood: In the maximum likelihood method, a commonly used statistical technique (see, for example, [23]) is used to find the tree most likely to have generated the current data. This requires an underlying model of sequence evolution specifying both residue frequencies and rates of evolution such as the Jukes-Cantor [43] or Kimura two-parameter [45] models in the case of DNA data or the PAM matrices in the case of protein data. The likelihood of seeing a transition at a given sequence position across a single branch of length d is the probability that a transition from residue i to residue j occurred during time d according to the model of sequence evolution chosen. If the data at each position are independent, then the likelihood of the branch over all characters is the product of the likelihoods for each position.

The likelihood for the entire tree can be computed by observing that the likelihood of seeing a residue at a given internal node is the product of the likelihoods that each daughter tree of that node gave rise to this nucleotide. As in parsimony, all possible tree topologies for k sequences must be considered.

3.3 Structure Prediction

Since the structure of biopolymers reveals much about their function, there has been great interest in determining the structures of proteins and RNA molecules. Protein structures can be determined experimentally using X-ray crystallography and nuclear magnetic resonance (NMR) techniques. However, these methods are difficult and time consuming, often requiring three to five years to determine a single structure experimentally. X-ray crystallography requires that a crystalline form of the protein be obtained. Reliable methods for protein crystallization do not exist. Furthermore, proteins *in vivo* are not in crystal form, so that protein structures determined through X-ray crystallography may not be the same as those which occur in the body. NMR does not require that proteins be crystalline but so far NMR has only been effective in determining the structure of small proteins.

Because of the difficulty in determining protein structures experimentally, there has been great interest in computational methods to predict a protein's structure from its sequence. Since proteins need to maintain their structure in order to function properly, regions of structural importance are usually highly conserved under selective pressure. Multiple sequence alignment can be used to identify these regions and, hence, underlies many structure prediction methods. Below we briefly review the use of multiple sequence alignment in determining structure. A detailed survey of protein structure prediction methods can be found in [24].

To facilitate reasoning about proteins, protein structure has been described hierarchically. The sequence itself is referred to as the primary structure. Secondary structure refers to α-helices and β-sheets, regular structures that are encoded by short subsequences and combine in myriad ways to form more complex structures. Regions encoding secondary structures are constrained by selection and therefore exhibit a high degree of conservation in related proteins. Variable length regions connecting α-helices and β-sheets, referred to as random coils, are much less constrained. Combinations of several secondary structures that are found in many, unrelated proteins are referred to as motifs (see [12] for an example). The complete structure of the folded protein is called the tertiary structure.

Prediction methods for tertiary structure range from the highly detailed energy minimization approaches, in which the physical interactions of all amino acids are modeled, to abstract approaches, such as lattice-based models and models where only the hydrophobic/hydrophilic properties of each amino acid are considered. In cases where the structure of several members of a protein family have been determined experimentally, this information can also be used to predict how other proteins in the family will fold. Comparative modeling methods use this approach by constructing a multiple sequence alignment and then aligning the new protein to this MSA. From that alignment, an initial estimate of the structure of the new protein is obtained from the structure of the known protein. This estimate is refined by perturbing the structure to accommodate the physical properties of the substituted amino acid in the new protein (see [31] for an example).

Secondary structure prediction methods have generally been statistical or learning-based approaches. Using a data base of known structures, these methods associate similar structures with patterns found in the primary sequence. In recent years, these methods have been substantially enhanced by the use of multiple sequence alignments. Multiple sequence alignments provide a statistical characterization of the patterns that encode the structure. An MSA is more informative than statistics of sequences considered individually since the information is grouped by position. Researchers have reported improvements in the accuracy of secondary structure prediction methods from roughly 60% without the use of MSA's up to 70% with the use of MSA's [55, 63, 64]. Statistical approaches to multiple sequence alignment have also been used to characterize motifs such as the EF-hand [47] and the Helix-Loop-Helix [49, 50] motifs. However, Krogh *et al.* [47] point out that while these methods find patterns associated with these motifs in the primary sequence, it has not been verified whether all of the patterns found actually encode functional motifs.

Multiple sequence alignments are also used in RNA structure prediction. Unlike DNA, in which two strands attract to form the well-known, regular helical structure, the single strand of an RNA molecule may be attracted to itself. Bonds form between different parts of the RNA strand causing the molecule to fold up into a three-dimensional structure. Like proteins, the tertiary structure of RNA is composed of a set of regular secondary structures. In RNA, these secondary structures include helical regions resulting from Watson-Crick base pairing (that is, the same bonding that unites the double helix of DNA molecules) and a variety of loops and bulges. A detailed survey of RNA structure is given in [36].

In order to maintain the structure of the RNA molecule, distant residues that bind to form a structural interaction are constrained to mutate together. Thus covariance in a multiple sequence alignment of related RNA molecules is a major source of predictive information. This approach to

RNA structure prediction is often called comparative analysis. These methods search for pairs of residues in the MSA that are linked: a change in one column is (almost) always accompanied by a complementary change in the other. This approach is also used to identify canonical structural building blocks. Some examples of RNA structure prediction algorithms that exploit evidence of correlation in multiple sequence alignments include [15, 37, 38].

3.4 Function

As discussed above, residues that contribute to the structure of a protein are constrained by selective pressure since a protein must hold its shape in order to function properly. Residues that are involved the biochemical function of the protein are also highly conserved. Casari *et al.* [14] argue that the evolutionary constraints on functional residues are even greater than on residues of structural importance and cite the *Ser-His-Asp* triplet in protein kinases as an example of functional conservation. This triplet has been shown to have a crucial rôle in catalysis. Such functional residues may be close in the 3-D structure but distant in the sequence. In a protein super family, residues that contribute to the specificity of function of the protein tend to be conserved within each subfamily and to vary from one subfamily to another. Multiple sequence alignments can be used to identify residues of functional importance and to discriminate between subfamilies. For example, Casari *et al.* [14] developed an interactive program that takes an MSA as input and, using principal component analysis, identifies the most highly conserved residues that characterize the family as a whole and as well as those residues that distinguish one subfamily from another.

4 Evaluating Tree Alignment as a Model for MSA

There is an unstated assumption that tree alignment is the gold standard, the ideal cost function, for multiple sequence alignment because the residues in each column are thought to be related by an evolutionary tree. The implication is that tree alignment is not used only because it is intractable. In addition, many MSA programs assume that a tree is the underlying model. Although they do not use a tree to score the aligment, a tree is used to construct the alignment (e.g., by guiding the order in which pairwise alignments are merged.)

In this section, we present some examples that demonstrate that a tree is not always the appropriate model for multiple sequence alignment. In considering whether tree alignment is appropriate for a given set of sequences, two issues must be addressed:

- Is a tree the correct model for describing the relationship between residues in each column?

  A tree may not be a suitable model because the relationship between residues is functional or structural rather than historical. Even if the relationship is historical, for some data sets, no single tree will describe all columns in the alignment.

- What is the correct mutational model for scoring the branches of the tree?

  Tree alignment has historically been based exclusively upon the parsimony criterion. Data that does not happen to be parsimonious can favor the wrong tree model. In addition, column-oriented optimization approaches to MSA usually assume that sequence positions are independent and identically distributed. These assumptions do not, in general, hold for biological sequence data.

The residues in a column do not share a common ancestor: When alignment is used to study function or structure, residues in a column do not always share a common ancestor. The goal is to align residues that share the same rôle. Although functional or structural residues usually share an evolutionary history, sometimes functional or structural rôles can migrate to neighboring residues.

A possible example of shifting function occurs in dihydrofolate reductase (DHFR) gene – an important chemotherapeutic target in treating cancer and various infectious diseases. In their studies on protozoan parasites, Roos and colleagues have sought to design drugs that inhibit a metabolic protein in the parasite without affecting the infected host [62, 20, 21]. This requires identifying regions of structural or functional importance that differ substantially between protozoan and human versions of the DHFR protein.

Early sequence alignments placed the malaria parasite DHFR residue Phe^{223}downstream of structurally conserved regions of the protein (within the linker region which joins DHFR to thymidylate synthase, forming a bifunctional protein in protozoa and plants) [41, 78]. This result was puzzling because mutational studies had suggested that Phe^{223}plays an important rôle in drug sensitivity. Realignment using sequences from the related parasite *Toxoplasma gondii* indicated that Phe^{223}is more likely homologous to a portion of the β-sheet which comprises the enzyme backbone [62]. The *T. gondii* sequence thus provided additional information that suggested an alternative alignment. Other protozoan sequences in the alignment have substantial, and different, nucleotide biases[1]. The *T. gondii* gene, which has relatively equal nucleotide distribution links the other protozoan sequences, facilitating alignment. This is another example of the observation, also made by McClure *et al.* [54],that MSA's are very sensitive to sequence choice.

Phe^{223}is thought to play an indirect rôle in enzyme activity, interacting with His^{34}, within the active site [59]. The residue that plays this stabilizing rôle may have changed over time: a residue in a different position may provide this stabilizing effect in certain taxa (e.g. kinetoplastid parasites such as *Leishmania* and *Trypanosoma*) [61]. In this scenario, a random mutation allows a previously inactive residue to take on the functional rôle played by Phe^{223}. In Roos' alignment, the residues currently thought to provide this stabilizing effect appear in the same column, but these residues in this column may not all share a common ancestor.

The tree is not unique: Another case where a tree is not an appropriate model occurs when the residues in any particular column share a common ancestor but the columns themselves have different evolutionary histories. A tree may describe any given column, but the columns taken together cannot be modeled by a single tree.

One situation where this occurs is exon shuffling. Most vertebrate genes consist of coding regions (exons) separated by DNA segments that are not translated into proteins (introns). The discovery of this intron/exon structure lead to the theory of exon shuffling, first proposed by Gilbert in 1978 [29]. This theory posits that exons represent functionally and/or structurally important subunits of proteins, that introns occur at the boundaries of these modules and that proteins share and reüse the modules that exons encode.

The first evidence that the same exons appear in more than one gene was found in the human low-density lipoprotein (LDL) receptor gene [74, 73]. The LDL receptor gene was shown to share

[1] The statistical profile of the primary sequence of genes can vary substantially, resulting in variations in, for example, the percentage of *GC* nucleotides. Such differences tend to obscure similarities between related sequences.

exons with genes for epidermal growth factor, blood clotting factor IX and complementation factor C9. Since then, many such "mosaic" genes have been discovered (see [22] for a survey). In aligning mosaic genes, a different tree may be needed for each exon. If the exon boundaries are known, each exon could be aligned separately. However for many sequences, the splice sites have not been determined. Detecting such boundaries would require that the alignment already be known.

Residues with different evolutionary histories within a single gene can also occur due to horizontal gene transfer, the transfer of genetic material between species. For example, sequences similar to the Fn3 module in fibronectin have been found in bacterial proteins [11]. Fibronectin is a protein found in animals. Since Fn3 is found in both bacterial and animal proteins, one would expect to find Fn3 modules throughout the tree of life. However, Fn3 sequences have not been found in simpler eukaryotes, plants or fungi [11], suggesting a direct transfer of genetic material between bacteria and animals. Thus, in an alignment of genes containing the Fn3 sequence, one would not expect the residues in the Fn3 module to share an evolutionary history with other residues in the alignment. More than one tree is needed to model the sequence and, as in the case of exon shuffling, it may not be possible to know where the module boundaries occur. Other examples of mixing of genetic material possibly requiring more than one tree include transposition and gene conversion.

The tree is not parsimonious Sequence data are not generally parsimonious, especially between distantly related sequences. Multiple substitution (e.g., A → C → T), coincidental substitution (e.g., A → C vs. A → G), parallel substitution (e.g., A → C vs. A → C) and back substitution (e.g., A → C → A), can all obscure the evolutionary history of a sequence.

Convergent evolution results in a situation where the parsimony criterion will lead to the wrong tree model. Residues that are similar under selective pressure because they perform the same function can appear to be closely related. An example of this occurs in cows and colobine monkeys, species that independently evolved foregut fermentation [71]. In order to maximize the nutritional benefits of foregut fermentation, the enzyme lysozyme had to evolve in these species to function in the acidic, pepsin-rich environment of the stomach. In other species, lysozyme is only needed in the intestines. As a result, lysozymes in cows and colobine monkeys exhibit amino acid substitutions not found in other species, suggesting, wrongly, that the two species are closely related. This would result in the wrong tree for those residues.

Sequence level convergence has also been observed by Hillis and his colleagues in viral strains grown in the laboratory [40]. Phage T7 strains were grown in different environmental conditions characterized by different bacterial hosts and different temperature regimes. Several separate viral populations were grown for each set of environmental conditions. Phylogenetic analysis of these populations showed convergent nucleotide substitutions at specific sites in the viral lineages in response to environmental factors (both host and temperature.) In an MSA of this viral DNA, parsimony analysis would infer a very different tree at those convergent sites than at other sites.

Tetraloops in rRNA provide another example of convergent evolution [36, 87]. Tetraloops are strings of six bases that form loops at the end of helices in rRNA structures. The two end bases bond, allowing the internal four bases to form a loop. Although there are 256 possible inner loop sequences, only a small number actually occur in nature. Tetraloop sequences will tend to appear to be closely related, even when they are not.

Sequence data is not i.i.d. Structural constraints prevent sequence data from being independent. Structural integrity depends on interactions between nonadjacent residues in the sequence. For example, α-helices are characterized by a heptad repeat, so that there are chemical interactions between every seventh residue in helical regions. Similarly, in order to maintain the structure of the RNA molecule, distant residues bind to form a structural interaction. Compensatory mutations between distant residues that form structural bonds are selectively favored.

Sequence data are not identically distributed either. Structural constraints on protein sequences result in variations in selective pressure at different positions, depending on whether they are located in α-helices, β-sheets or random coils and whether they have a rôle in tertiary structure or biochemical function. This fact has been recognized and exploited by researchers who developed structure-specific substitution matrices for recognizing specific secondary structures and motifs [53, 55, 57].

Additional variations in selective pressure occur at the DNA level. In protein coding regions, substitutions can be nonsynonymous (resulting in an amino acid substitution in the protein coded for) or synonymous (resulting in a different codon for the same amino acid). Due to differences in selective presure, synonymous changes are seen with more frequency than nonsynonymous changes. Originally, it was thought that sequence positions could be classified as replacement sites, synonymous sites and noncoding sites and that mutation rates within each class would be relatively constant. More recently, evidence has emerged that suggests that selective pressure can vary within each class, even within a single gene or intron [46, 67]

5 Conclusion

Evolutionary trees have been used as an abstract model for multiple sequence alignment in designing algorithms and optimization criteria and in building software tools. Although frequently viewed as the fundamental model of MSA, the appropriateness of the tree model depends both on the application and the data set. After surveying the major biological applications, problems with the tree model were illustrated using examples from the biological literature. If the alignment is constructed to reveal functional or structural similarities, a tree may not correctly describe the relationship between the residues. Lateral transfers of DNA fragments between genes may result in situations where no single tree can model the entire gene. Nonparsimonious data and convergent evolution can lead to the wrong tree model.

This paper is intended to give computer scientists a better understanding of the biological uses of multiple sequence alignment and how real data sets differ from the abstract assumptions made about sequence data. We hope that these ideas will provide a basis for dialogue between biologists and computer scientists and mathematicians, leading to better algorithm design and software development for multiple sequence alignment.

6 Acknowledgements

The author wishes to thank Sergei Agulnik, Maja Bucan, Martin Farach, Dan Gusfield, John Kececioglu, Laura Landweber, Jason Miller, Eugene Myers, Gary Olsen, Chris Overton, R. Ravi, David Roos, Ilya Ruvinsky, Chris Sander, Mona Singh, Temple Smith amd E. Trifonov for helpful discussions and Harold Stone for invaluable editorial advice.

References

[1] A. Aevarsson. Structure-based sequence alignment of elongation factors TU and G with related GTPases involved in translation. *Journal of Molecular Evolution*, 41:1096 – 1104, 1995.

[2] R. Agarwala, V. Bafna, M. Farach, B. Narayanan, M. Paterson, and M. Thorup. On the Approximability of Numerical Taxonomy (fitting distances by tree metrics). In *Proceedings of the Symposium on Discrete Algorithms*, 1996.

[3] S.F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology*, 219:555–565, 1991.

[4] S.F. Altschul, R.J. Carroll, and D.J. Lipman. Weights for data related by a tree. *Journal of Molecular Biology*, 207:647–653, 1989.

[5] S.F. Altschul and D.J. Lipman. Trees, Stars and Multiple Sequence Alignment. *Journal of Applied Mathematics*, 49(1):197–209, 1989.

[6] P. Argos. A sensitive procedure to compare amino acid sequences. *Journal of Molecular Biology*, 193:385 – 396, 1987.

[7] D. J. Bacon and W. F. Anderson. Multiple sequence alignment. *Journal of Molecular Biology*, 191:153–161, 1986.

[8] V. Bafna, E. L. Lawler, and P. Pevzner. Approximation Algorithms for Multiple Sequence Alignment. In *5th Ann. Symp. On Pattern Combinatorial Matching*, volume 807, pages 43–53, 1994.

[9] A. Bairoch and P. Bucher. PROSITE: recent developments. *Nucleic Acids Res.*, 22(5):3583–3589, 1994.

[10] G.J. Barton and M.J.E. Sternberg. Evaluation and improvements in the automatic alignment of protein sequences. *Protein Engineering*, 1:89–94, 1987.

[11] P. Bork and R. F. Doolittle. Proposed acquisition of an animal protein domain by bacteria. *Proc.Natl.Acad.Sci. USA*, 89:8990–8994, 1992.

[12] C.-I. Branden. The TIM barrel – the most frequently occuring folding motif in proteins. *Current Opinion in Structural Biology*, 1:978–983, 1991.

[13] H. Carillo and D.J. Lipman. The Multiple Sequence Alignment problem in biology. *Journal of Applied Mathematics*, 48:1073–1082, 1988.

[14] G. Casari, C. Sander, and A. Valencia. A method to predict functional residues in proteins. *Structural Biology*, 2(2):171–178, 1995.

[15] L. Chan, M. Zuker, and A. B. Jacobson. A computer method for finding common base paired helices in aligned sequences: application to the analysis of random sequences. *Nucleic Acids Res.*, 19(2):353–358, 1991.

[16] S.C. Chan, A.K.C. Wong, and D.K.Y. Chiu. A Survey of Multiple Sequence Comparison Methods. *Bulletin of Mathematical Biology*, 54:563–598, 1992.

[17] Q. K. Chen, G. Z. Hertz, and G. D. Stormo. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Computer Applications in the Applied Sciences*, 11:563–66, 1995.

[18] W. H. E. Day. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49(4):461–467, 1987.

[19] M. O. Dayhoff. *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington, DC, 1978.

[20] R. G. Donald and D. S. Roos. Stable molecular transformation of *toxoplasma gondii*: a selectable dihydrofolate reductase-thymidylate synthase marker based on drug-resistance mutations in malaria. *Proc.Natl.Acad.Sci. USA*, 90:11703–11707, 1993.

[21] R. G. K. Donald and D. S. Roos. Homologous recombination and gene replacement at the dihydrofolate reductase-thymidylates synthase locus in *toxoplasma gondii*. *Molec. Biochem. Parasitol.*, 63:243–253, 1994.

[22] R. L. Dorit and W. Gilbert. The limited universe of exons. *Curr Opin Genet Dev*, 1:464–469, 1991.

[23] David Durand. *Stable Chaos. An introduction to statistical control.* General Learning Press, Morristown, NJ, 1971.

[24] F. Eisenhaber, B. Persson, and P. Argos. Protein structure prediction: Recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Critical Rev in Biochem and Mol Bio*, 30(1):1–94, 1995.

[25] D.F. Feng, M.S. Johnson, and R.F. Doolittle. Aligning amino acid sequences: comparison of commonly used methods. *Journal of Molecular Evolution*, 21:112–125, 1985.

[26] J. W. Fickett. Quantitative discrimination of mef2 sites. *Molecular and Cellular Biology*, 16(1):437–441, 1996.

[27] W.M. Fitch and T.F. Smith. Optimal sequence alignments. *PNAS*, 80:1382–1386, 1983.

[28] M. S. Gelfand. Prediction of function in DNA sequence analysis. *Journal of Computational Biology*, 2(1):87–115, 1995.

[29] W. Gilbert. Why genes in pieces? *Nature*, 271:501, 1978.

[30] A. Godzik. The structural alignment between two proteins: Is there a unique answer? *Protein Science*, 5:1325–1338, 1996.

[31] J. Greer. Comparative modeling methods: Application to the family of the mammalian serine proteases. *PROTEINS: Structure, Function and Genetics*, 7:317–334, 1990.

[32] M. Gribskov and J. Devereux. *Sequence Analysis Primer.* Stockton Press, New York, NY, 1991.

[33] M. Gribskov, R. Luthy, and D. Eisenberg. Profile analysis. In *Methods in Enzymology*, volume 183, pages 146–159. Academic Press, 1990.

[34] D. Gusfield. Efficient Methods for Multiple Sequence Alignment with Guaranteed Error Bounds. *Bulletin of Mathematical Biology*, 55:141–154, 1993.

[35] D. Gusfield, K. Balasubramanian, and D. Naor. Parametric Optimization of Sequence Alignment. In *Proceedings of the Symposium on Discrete Algorithms*, pages 432–439, 1992.

[36] R.E. Gutell, N. Larsen, and C.R. Woese. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiological Reviews*, 58(1):10–26, March 1994.

[37] R.E. Gutell, A. Power, G.Z. Hertz, E.J. Putz, and G.D. Stormo. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, 20(21):5785–5795, 1992.

[38] K. Han and H-J Kim. Prediction of common folding structures of homolgous RNAs. *Nucleic Acids Res.*, 21(5):1251–1257, 1993.

[39] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc.Natl.Acad.Sci. USA*, 89:10915–9, 1992.

[40] D. Hillis. Personal communication.

[41] J.E. Hyde. The dihydrofolate reductase-thymidylate synthetase gene in the drug resistance of malaria parasites. *Pharmacol Ther*, 48(1):45–59, 1990.

[42] T. Jiang, E.L. Lawler, and L. Wang. Aligning sequences via an evolutionary tree: Complexity and approximation. In *Proceedings of the Symposium on the Theoretical Aspects of Computer Science*, pages 760–769, 1994.

[43] T. H. Jukes and C. R. Cantor. Evolution of Protein Molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[44] J. Kececioglu. The maximum weight trace problem in multiple sequence alignment. In *4th Ann. Symp. On Pattern Combinatorial Matching, Springer Verlag Lecture notes in Computer Science*, volume 684, pages 106–119, 1993.

[45] M. Kimura. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[46] M. Kreitman and R. R. Hudson. Inferring the evolutionary histories of the Adh and Adh-dup loci in *drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics*, 127:565–582, 1991.

[47] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology: Applications to protein modeling. Technical Report UCSC-CRL-93-32, University of California, 1993.

[48] Y. Kubota, S. Takahashi, K. Nishikawa, and T. Ooi. Homology in protein sequences expressed by correlation coefficients. *Journal of Theoretical Biology*, 91:347–361, 1981.

[49] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liuand, A. F. Neuwald, and J. C. Wootton. Detecting Subtle Sequence Signals: a Gibbs Sampling Strategy for Multiple Alignment. *Science*, 262:208–214, 1993.

[50] C. E. Lawrence, S. F. Altschul, J. C. Wootton, M. S. Boguski, A. F. Neuwald, and J. S. Liu. A Gibbs Sampler for the Detection of Subtle Motifs in Multiple Sequences. In *Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences*, pages 245–254, 1994.

[51] A.M. Lesk and C. Chothia. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *Journal of Molecular Biology*, 136:225–270, 1980.

[52] A.M. Lesk, M. Levitt, and C. Chothia. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Protein Engineering*, 1:77–78, 1986.

[53] R. Luthy, A. D. McLachlan, and D. Eisenberg. Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins*, 10:229–239, 1991.

[54] M.A. McClure, T.K. Vasi, and W.M. Fitch. Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.*, 11:571 – 592, 1994.

[55] P.K. Mehta, J. Heringa, and P. Argos. A fast and simple approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Science*, 4:2517–2525, 1995.

[56] M. Murata, J. S. Richardson, and J. L. Sussman. Simultaneous comparison of three protein sequences. *Proc.Natl.Acad.Sci. USA*, 82:3073–3077, 1985.

[57] J. Overington, D. Donnelly, M. S. Johnson, A. Sali, and T. L. Blundell. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Science*, 1:216–226, 1992.

[58] P.A. Pevsner. Multiple alignment, communication cost and graph matching. *Journal of Applied Mathematics*, 52:1763–1779, 1992.

[59] M. Reynolds, D. Carter, M. Schumacher, and D. S. Roos. Personal communication.

[60] J. L. Risler, M. O. Delorme, H. Delacroix, and A. Henaut. Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *JMB*, 204:1019–1029, 1988.

[61] D. S. Roos. Personal communication.

[62] D.S. Roos. Primary structure of the dihydrofolate reductase-thymidylate synthase gene from *toxoplasma gondii*. *J. Biol. Chem.*, 268:6269–6280, 1993.

[63] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *JMB*, 232:584–599, 1993.

[64] A. A. Salamov and V. V. Solovyev. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of Molecular Biology*, 247:11–15, 1995.

[65] D. Sankoff. Minimal mutation trees of sequences. *Journal of Applied Mathematics*, 28:443–453, 1975.

[66] D. Sankoff and R. J. Cedergren. Simultaneous Comparison of Three or More Sequences Related by a Tree. In *Timewarps, Edits and Macromolecules: The Theory and Practise of Sequence Comparison*, pages 253–258. Addison-Wesley, Reading, MA, 1983.

[67] S. W. Schaeffer and C. F. Aquadro. Nucleotide sequence of the Adh gene region of *drosophila pseudoobscura*: evolutionary change and evidence for an ancient gene duplication. *Genetics*, 117:61–73, 1987.

[68] R. M. Schwartz and M. O. Dayhoff. Matrices for detecting distant relationships. In M. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, pages 353–358. National Biomedical Research Foundation, Washington, DC, 1979.

[69] P.R. Sibbald and P. Argos. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *Journal of Molecular Biology*, 216:813–818, 1990.

[70] E. L. Sonnhammer and D. Kahn. Modular arrangement of proteins as inferred from analysis of homology. *Protein Science*, 3:482–492, 1994.

[71] C.-B. Stewart and A. C. Wilson. Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters. *Cold Spring Harbor Symposium on Quantitative Biology*, 52:891–899, 1987.

[72] G. D. Stormo. Consensus patterns in DNA. In *Methods in Enzymology*, volume 183, pages 211 – 221. Academic Press, 1990.

[73] T. C. Sudhof, J. L. Goldstein, M. S. Brown, and D. W. Russell. The LDL receptor gene: a mosaic of exons shared with different proteins. *Science*, 228:815–822, 1985.

[74] T. C. Sudhof, D. W. Russell, J. L. Goldstein, M. S. Brown, R. Sanchez-Pescador, and G. I. Bell. Cassette of eight exons shared by genes for LDL receptor and EGF precursor. *Science*, 228:893–895, 1985.

[75] E. Sweedyk and T. Warnow. Manuscript. 1992.

[76] D. L. Swofford and G. J. Olsen. Phylogeny Reconstruction. In *Molecular Systematics*, pages 411–501. Sinauer Associates, Inc., Sunderland, MA, 1990.

[77] D. L. Swofford, G. J. Olsen, Waddell, and D. M. Hillis. Phylogenetic Inference. In *Molecular Systematics.* Sinauer Associates, Inc., Sunderland, MA, 1996.

[78] M. Tanaka, H. M. Gu, D. J. Bzik, W. B. Li, and J. W. Inselburg. Dihydrofolate reductase mutations and chromosomal changes associated with pyrimethamine resistance of *plasmodium falciparum. Mol Biochem Parasitol*, 39:127–134, 1990.

[79] R. L. Tatusov, S.F. Altschul, and E.V. Koonin. Detection of Conserved Segments in Proteins: Iterative Scanning of Sequence Databases with Alignment Blocks. *Proceedings of the National Academy of Sciences, USA*, 91:12091–12095, 1994.

[80] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *CAB*, 10:19–29, 1994.

[81] A. Valencia, M. Kjeldgaard, E. F. Pai, and C. Sander. GTPase domains of ras p21 oncogene protein and elongation factor Tu: analysis of three-dimensional structures, sequence families, and functional sites. *Proc.Natl.Acad.Sci. USA*, 88:5443–5447, 1991.

[82] M. Vingron and P.R. Sibbald. Weighting in sequence space: a comparison of methods in terms of generalized sequences. *Proc.Natl.Acad.Sci. USA*, 90:8777–8781, 1993.

[83] M. Vingron and M. S. Waterman. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *Journal of Molecular Biology*, 235:1–12, 1994.

[84] G. Vriend and C. Sander. Detection of common three-dimensional substructures in proteins. *Proteins*, 11(1):52–58, 1991.

[85] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337–348, 1994.

[86] M. S. Waterman, T. F. Smith, M. Singh, and W. A. Beyer. Additive Evolutionary Trees. *Journal of Theoretical Biology*, 64:199–213, 1977.

[87] C. R. Woese, S. Winker, and R. R. Gutell. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc.Natl.Acad.Sci. USA*, 87:8467–8471, 1990.