DIMACS Technical Report 96-53
January 1997

On the Design of Optimization Criteria for Multiple Sequence Alignment

by

Dannie Durand[1,2]
Computational Biology Group
University of Pennsylvania
Philadelphia, PA 19104-6228
and
DIMACS
Rutgers University
Piscataway, NJ 08855-1179
durand@saul.cis.upenn.edu

Martin Farach[3,4]
Department of Computer Science
Rutgers University
Piscataway, NJ 08855
farach@cs.rutgers.edu, http://www.cs.rutgers.edu/~farach

[1]Permanent Member
[2]Supported by NSF Grants BIR-94-13215 A01 and BIR-94-12594.
[3]Permanent Member
[4]Supported by NSF Career Development Award CCR-95-01942, NSF Grant BIR-94-12594, an Alfred P. Sloan Research Fellowship and NATO Grant 96-0215.

ABSTRACT

Multiple sequence alignment (MSA) is important in functional, structural and evolutionary studies of sequence data. Much research has focussed on posing MSA as an optimization problem, and several optimization criteria have been explored. In this paper, we discuss biological and mathematical problems that arise in cost function design for the multiple sequence alignment problem. In particular, we focus on tree alignment, which is often viewed as the most "biological" of the rigorous approaches to MSA. We point out several important pitfalls in current optimization approaches to MSA and identify characteristics for good cost function design. We address some extra design issues specific to approximation algorithms. We hope these ideas will lead to future research on a biologically realistic and mathematically rigorous approach to MSA.

1 Introduction

One of the basic ways to extract shared information from a set of biopolymers is to compute their *multiple sequence alignment (MSA)*. MSAs are used to build evolutionary trees, extract information about the function and structure of proteins, identify patterns for data base searching and motif recognition and to design probes and primers for wet lab experiments. Suppose each polymer, P_1, $P_2, \ldots, P_k$, is represented as 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. Then an MSA is a set of sequences $\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_k$ drawn from $\Sigma \cup \{\_\}$, where _ is called a *blank*, and such that P_i is the string derived by removing all blanks from $\hat{P}_i$. Additionally, each $\hat{P}_i$ should be of the same length. We will call $\hat{P}_1[j], \ldots, \hat{P}_k[j]$ the *jth column* of the MSA. The standard interpretation of blanks is that they represent mutations in the form of insertions or deletions, and therefore they are sometimes called *indels*. However, the sequences $P_1, \ldots, P_n$ need not be evolutionarily related, a fact which we will explore further below.

Columns in an MSA should share a target biological feature. The feature sought depends on the application of the alignment. For example, if the multiple alignment is used to illustrate evolutionary relationships, then the residues in each column should 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. It is exactly this kind of feature extraction that is the goal of MSA computation, the underlying assumption of which is that the sequences of a molecule contain enough information to extract the feature. The question, then, is how to compute an MSA.

One of the most popular approaches to MSA computation is to design a *cost function* over the set of all possible MSAs and then to seek the MSA which maximizes[1] the cost. The advantage of this *optimization* approach is that the cost function makes explicit the assumptions upon which the optimization is based. Because they are explicit, these assumptions are open to scrutiny and falsification. Our focus on optimization should be contrasted with the approach typically taken in production software for MSA. Most programs use greedy incremental approaches to build some MSA, rather than having a predefined notion of when one MSA is better than another. At best, they rely solely on empirical validation against well-established alignments.

In Section 2, we examine several popular cost functions that have received widespread attention in the literature, including sum-of-pairs, star and tree alignment. We focus primarily on tree alignment, the most biologically motivated of the three. Important biological considerations that must be addressed by any sound optimization criterion are elucidated in Section 3. Here we also specifically address the biological reasonableness of tree alignment for MSA computation. We explore mathematical aspects of cost function design in Section 4. In particular, design issues related to approximation algorithms are discussed. We conclude that while optimization and approximation offers the best hope for well founded MSA computation, the current crop of optimization criteria are seriuosly wanting. We suggest some

[1] We will interchangeably minimize and maximize the function in our examples, as needed.

avenues of attack in correcting the situation.

2 Optimization Criteria

As noted above, it is the column relationships which is the desired output of an MSA computation. Thus, it is natural to focus on optimization criteria which associate a score with each column, and most criteria in the literature take this approach[2]. This score expresses how well matched the residues in the column are. The score of the MSA is then the sum of the scores of its columns. We seek a set of $\hat{P}$ maximizing

$$D(A) = \sum_j d(\hat{P}_1[j], \ldots, \hat{P}_k[j]),$$

for column scoring function d. The choice of the scoring function, $d(\hat{P}_1[j], \ldots, \hat{P}_k[j])$, should reflect the application of the alignment. Residues that maximize $d(\hat{P}_1[j], \ldots, \hat{P}_k[j])$ should belong together.

Three cost functions commonly used to evaluate each column are sum-of-pairs (SP), tree alignment (TA) and star alignment (SA). The sum-of-pairs cost [24, 3] of a column $\hat{P}_1[j], \ldots, \hat{P}_k[j]$ is the sum of the costs of all unordered pairs in the column

$$d_{SP}(\hat{P}_1[j], \ldots, \hat{P}_k[j]) = \sum_{p<q} \delta(\hat{P}_p[j], \hat{P}_q[j]),$$

for some binary cost function $\delta(x, y)$. This definition is mathematically natural but not biologically intuitive.

Tree alignment [30, 31] 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 single tree for all columns. Under this model, a column is scored by computing the cost of the underlying tree. The score of the tree expresses the strength of our belief, 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.

[2] We will not give a comprehensive survey of multiple sequence alignment algorithms here. Such surveys can be found in [8] and in the introduction to [25]. An experimental comparison of multiple sequence algorithms is described in [22].

Let us consider how to infer the ancestral sequences and compute the branch costs when the topology is known. The k extant sequences are associated with the k leaves of the tree. Sequences for the internal nodes are selected so that the tree is the best estimate of the true tree under some model of evolutionary change. The model in use 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 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 minimized. This assumption that evolution is parsimonious, is a subject of much debate[3].

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 the tree which yields the best score must be found. Since the number of trees with k leaves grows exponentially with k, this approach rapidly becomes prohibitively expensive if exhaustive enumeration is used. In general, no fast algorithm is known.

A variant on tree alignment is star alignment (SA), in which it is assumed that the underlying tree is a star. 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., [2, 15]).

Altschul and Lipman [2] pointed out that SP, TA and SA costs for the same column can be very different, as shown in Figure 1. SP overcounts mutations in a column because

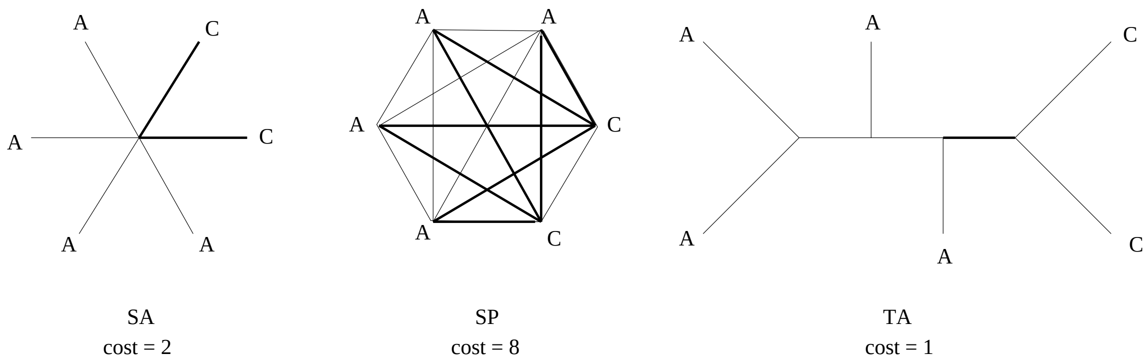

Figure 1: Comparing alignment costs.

it considers that a separate mutation occurred between each pair in the column. In some sense, it is a "least parsimonious," or profligate, 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.

[3]Indeed, the parsimony criterion is anti-Darwinian, since it assumes that either mutation or selection favors parsimonious trees. In the first case, this implies that mutations are guided along shortest paths rather than being random. In the second, we must conclude that the fitness of a living species is related to its mutational history, rather than to its sequence.

Biological Validation of MSA The methods described above use only primary sequence information to align sequences. However, structural information has also been used 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 [20], 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, 29]). This type of structural alignment is often done manually. Second, structural alignments may be performed by minimizing the root mean square distance between the α-carbon molecules in the backbone (see [38, 39], for an example).

Structural and functional information has been used to validate multiple sequence alignments. In a comprehensive study of twelve different multiple alignment programs, McClure, Vasi and Fitch [22] measured algorithm performance by computing a numerical score based on the ability to find known motifs in four different data sets, for which supporting evidence from structural or mutational studies was available. Barton and Sternberg [5] 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. While the fraction of biopolymers for which structural information is available is small, the majority of newly sequenced biopolymers turn out to be members of families for which some structure is known. Another consideration is whether a structural approach to MSA is useful in all cases. For example, as discussed in Section 3, residues that share a structural or functional rôle do not always share an evolutionary history.

To our knowledge, there has been no research that attempts to unify biological and mathematical notions of what is a good alignment in a formal context.

Complexity of Multiple Sequence Alignment. Complexity results have been published for MSA for sum-of-pairs and tree alignment. Hardness results include those in [36, 40]. Exact methods for both SP and tree alignment 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 the obvious dynamic program in $O(2^k N^k)$ evaluations of the cost function d using $O(N^k)$ space. An exact tree alignment algorithm for a given tree topology has been presented by Sankoff [30]. 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. Approximation algorithms for both SP and tree alignment include those in [4, 15, 25, 26]. Notably, a polynomial time approximation scheme, yielding a $(1+\epsilon)$-approximation for arbitrarily small ϵ, has been presented by Jiang, Lawler and Wang [18] for tree alignment on a fixed topology.

3 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 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. In general, these assumptions do not hold for biological sequence data.

The residues in a column to 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 [29, 9, 10]. 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)[17, 37]. 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 [29]. 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[4]. 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, Vasi and Fitch [22],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 [27]. 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*) [28]. 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 [14]. 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 [35, 34]. The LDL receptor gene was shown to share 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 [11] 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 [6]. 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 [6], 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

[4]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.

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 $\rightarrow$ C $\rightarrow$ T), coincidental substitution (e.g., A $\rightarrow$ C vs. A $\rightarrow$ G), parallel substitution (e.g., A $\rightarrow$ C vs. A $\rightarrow$ C) and back substitution (e.g., A $\rightarrow$ C $\rightarrow$ 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 [33]. 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.

Tetraloops in rRNA provide another example of convergent evolution [16, 41]. 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 non-adjacent 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 [21, 23].

Additional variations in selective pressure occur at the DNA level. In protein coding regions, substitutions can be non-synonymous (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 non-synonymous changes. Originally, it was thought that sequence positions could be classified as replacement sites, synonymous sites and non-coding 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 [19, 32]

4 Approximations and Convergence

Since any reasonable cost function for MSA will probably be NP-hard to optimize [36, 40], it is natural for researchers to look for approximation algorithms which find near optimal solutions to the MSA problem. These solutions will not be meaningful unless they satisfy (at least) the following two properties.

Smoothness The success of the optimization approach to MSA depends on finding a cost function such that the optimal solution with respect to that cost function is close enought to the true solution. In the previous sections, we have seen that the definition of "true" solution depends on the application (e.g., reconstructing phylogeny, determining structure or function). So the cost function should reflect the application in some way. Unlike problems such as the Traveling Salesman Problem, where the cost function (total distance traveled) is directly related to the problem under consideration (finding the shortest tour), mathematical cost functions used in MSA often have little relationship to the solution sought.

For approximation, stricter constraints on the properties of MSA cost functions are required. Intuitively, an MSA that is close in score to the optimal MSA should also be similar biologically. We call this notion *smoothness*. Unfortunately, "biologically similar" is a vague notion. It would be useful to make this notion concrete, for approximation algorithms are only meaningful if we understand the behavior of the cost function in the neighborhood of the optimal solution. What mathematical properties does a cost function require in order to exhibit smoothness? Under tree alignment, for example, it is not generally true that MSAs with similar scores have similar biological interpretation, since it is possible for many different trees to have the same parsimony score.

Consistency Biological wisdom holds that MSA computation should get easier as more sequences are introduced, and harder as the sequences get longer. Roos [29] and McClure, Vasi and Fitch [22] have described specific instances of this behavior in their research. By contrast, phylogeny computation is held to get harder as the number of species increases but easier as the length of the sequences describing the species increases. Our optimization criteria should capture this varied dependence of problem instance difficulty on the size of the data set if they have any hope of capturing the underlying biology.

Ideally, the maximum of our optimization criterion should coïncide with the "true" or biological alignment. A weaker notion, and one inspired by the considerations above, is that as the number of sequences being considered increases, the maximum of our optimization criterion should *converge* on the true alignment. This idea of converging on the true answer as we observe more data is basically the notion of *consistency* [7, 13]. For conctretness, we define an algorithm to be *consistent* with respect to some parameter k of data size (e.g., number of species for MSA or sequence length for phylogeny) if it converges to the true solution as k goes to infinity. Similarly, a cost function is *consistent* if an algorithm that outputs its maximum is consistent. Notice that for MSA we would not expect a cost function to be consistent with respect to sequence length.

Suppose we have a cost function which is consistent. Would an approximation algorithm for the function be consistent? We consider a couple of examples. Suppose

$$g(\phi, d_k) = \begin{cases} 2 & \text{if } \phi \text{ is the true alignment for data } d_k \text{ on } k \text{ sequences;} \\ 1 & \text{otherwise.} \end{cases}$$

Now trivially, g is a consistent cost function. But a 1/2-approximation algorithm for g would be inconsistent, because *any* alignment yields a score at least half the maximum, so a 1/2 approximation algorithm need not converge on the true answer. On the other hand, suppose

$$h(\phi, d_k) = \begin{cases} k & \text{if } \phi \text{ is the true alignment for data } d_k \text{ on } k \text{ sequences;} \\ 1 & \text{otherwise.} \end{cases}$$

Then any constant factor approximation algorithm for h will be consistent. This is because a $1–\epsilon$ approximation algorithm for h will be forced to output the true alignment for $k > 1/(1–\epsilon)$.

So we see that consistency for approximation algorithms requires more than an understanding of the biology of MSA, but also an understanding of the entire solution space for the cost function of choice. In particular, we need to know what happens to the set of alignments that are within a factor of $1-\epsilon$ of the best alignment, and how this set changes with k. We will say that a cost function f is *ϵ-consistent* if the set of all solutions within $1-\epsilon$ of the optimal for f converges to the true solution as k goes to infinity. We say f is *strongly consistent* if f is ϵ-consistent for every constant ϵ. In our example above, function g is ϵ consistent, for any $\epsilon < 1/2$, while function h is strongly consistent.

For phylogeny, a strongly consistent cost function was proposed in [12]. It was possible to do so in the case of phylogeny because the stochastic process relating phylogenies to DNA is well understood. We must seek such an understanding of MSAs before we can propose meaningful optimization criteria for this problem.

5 Conclusion

Optimization is an extremely effective tool for attacking many computational problems. In the case of problems from computational biology, one must be careful that the optimization criterion used closely follows some specific biological model of how the data relate to the underlying structure sought, e.g. a phylogeny or an MSA. In this vein, tree alignment has been touted as a "biological" approach to the MSA problem. However, as argued in Section 3, the definition of tree alignment (i.e. a single tree under a parsimonious scoring scheme) renders it meaningless in many, if not most, biological settings. Furthermore, we argue in

Section 4 that approximation algorithms for tree and other alignment criteria do not exploit biologically intuitive notions of convergence.

None of this need be the case. We are not pointing out fundamental shortcomings of the optimization/approximation approach in computational biology. Indeed, we believe that this approach offers the best hope for a principled attack on MSA and other problems. However, the strength of this approach, most notably a transparent association between underlying assumptions and computational methods, is rendered meaningless if the underlying assumptions are not biological. We therefore call for a more focused search for a biologically grounded model for MSA computation. Our paper points out many pitfalls of any such search, and suggests new directions for improvement. It should be considered a call to action.

6 Acknowledgements

The authors wish to thank Sergei Agulnik, Maja Bucan, Dan Gusfield, Laura Landweber, Eugene Myers, Dalit Naor, Chris Overton, R. Ravi, David Roos, Ilya Ruvinsky and Mona Singh for helpful discussions.

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

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

[4] 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.

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

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

[7] J. Cavender. Taxomony with confidence. *Mathematical Biosciences*, 40:271–280, 1978.

[8] 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.

[9] 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.

[10] 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.

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

[12] M. Farach and S. Kannan. Efficient algorithms for inverting evolution. In *Proceedings of the Symposium on the Theoretical Aspects of Computer Science*, 1996.

[13] J. Felsenstein. Cases in which parsimony or compatibility methods will be postively misleading. *Syst. Zool.*, 22:240–249, 1978.

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

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

[26] R. Ravi and J. D. Kececioglu. Approximation algorithms for multiple sequence alignment under a fixed evolutionary tree. In *6th Ann. Symp. On Pattern Combinatorial Matching, Springer Verlag Lecture notes in Computer Science*, 1995.

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

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

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

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

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

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

[41] 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.