Multiple Sequence Alignment Errors and Phylogenetic Reconstruction

THESIS SUBMITTED FOR THE DEGREE "DOCTOR OF PHILOSOPHY"

BY

Giddy Landan

SUBMITTED TO THE SENATE OF TEL-AVIV UNIVERSITY

August 2005

This work was carried out under the supervision of

Professor Dan Graur

Acknowledgments

I would like to thank Dan for more than a decade of guidance in the fields of molecular evolution and esoteric arts.

This study would not have come to fruition without the help, encouragement and moral support of Tal Dagan and Ron Ophir. To them, my deepest gratitude.

Time flies like an arrow

Fruit flies like a banana

- Groucho Marx

Table of Contents

Abstract

This research aims at developing a methodology for identifying and accounting for multiple sequence alignment (MSA) uncertainties in phylogenetic reconstruction. The research consists of two parts: (a) characterization of alignment errors and their effect on subsequent phylogenetic reconstruction, and (b) development of methods to identify alignment errors and reduce their detrimental effects on phylogenetic reconstruction. Phylogenetic reconstruction is but one alignment-dependent analysis that may benefit from the identification and management of alignment errors. Therefore, the methods and results of this study have methodological implications in other alignment-dependent sequence-analysis problems.

We describe and characterize multiple sequence alignment errors by comparing true native alignments from simulations of sequence evolution, with reconstructed alignments from ClustalW (Thompson *et al.*, 1994a), which is the most widely used multiple sequence alignment reconstruction program. Reconstructed alignments are found to contain many errors. Error rates increase with sequence divergence, and rapidly span very large portions of reconstructed MSAs, so that even for intermediate sequence divergence more than half of the columns of reconstructed alignments can be expected to be erroneous.

In closely related sequences, most errors consist of the erroneous positioning of a single indel event, and their extent is local. As sequences diverge, errors are the result of the simultaneous mis-reconstruction of many indel events, and the length of the affected MSA segments increase dramatically. We also found a systematic bias towards

underestimation of the number of gap characters, which lead to the shortening of reconstructed MSAs relative to their true MSA.

Alignment errors are unavoidable even when the evolutionary parameters are known in advance. Correct reconstruction can be guaranteed only when the true alignment is uniquely optimal in terms of its likelihood. However, true alignment features are very frequently sub-optimal or co-optimal, with the result that optimal but erroneous features are incorporated into the reconstructed MSA.

Progressive MSA utilizes an approximate phylogeny, or guide-tree, in the reconstruction of MSAs. We found that the quality of the guide-tree affects MSA error level only marginally, but that the guide-tree topology introduces a bias in the phylogenetic signal apparent in erroneous MSA columns.

Exploring the effects of alignment errors on subsequent phylogenetic reconstruction, we show that when presented with high-quality alignments, current phylogenetic reconstruction methods, such as BioNJ (Gascuel, 1997), are quite adequate. However, phylogenetic reconstruction rates deteriorate rapidly as alignments become more ambiguous. We clear consciously lay the blame at the feet of the reconstructed alignments.

To address the issue of MSA errors in real-life biological settings, we adopt a methodology that replaces the single reconstructed alignment with a set of alternative alignments for the same sequences. We propose that such a set should consist of equally likely alignments, and that its variability should reflect common types of reconstruction errors. A secondary requirement is that the alignment set should be of a moderate size to

render its analysis feasible. The alignment set we develop reflect two sources of MSA reconstruction errors: the addition order of sequences and the arbitrary choices from among co-optimal alignments.

One use of the alignment set is to derive local reliability measures for candidate MSAs. Elements of a candidate MSA that are reproduced in many MSAs within the set, are considered reliable, whereas parts of the candidate MSA that are poorly supported by the set are down-scored. We define a family of reliability measures with four levels of resolution: residues-pairs, residues, columns and the entire MSA. The local reliability measures are found to be excellent estimators and classifiers of MSA errors, and to be superior to currently used MSA quality scores.

We have tested the utility of the local reliability measures in phylogenetics by weighting and filtering a ClustalW MSA prior to phylogenetic reconstruction. Unfortunately, what we have found is that identification of alignment errors is not enough to boost the quality of MSA-dependent phylogenetic reconstruction. We explain this result by the observations that such filtering significantly reduce the sample size, and that the high-quality portions of the alignment are also less informative from the phylogenetic perspective. We conclude that poor-quality MSAs can not be transformed into high-quality ones merely by the identification of possible errors.

An alternative to filtering a single MSA is to derive a phylogeny directly from the set of MSAs. We reconstruct a phylogeny from each member of the alignment set, producing a set of alternative phylogenies. The consensus of these alternative phylogenies is than taken as the final reconstructed phylogeny. We note that this type of analysis utilizes

much more of the information contained in the alignment set than the scoring of a single MSA.

The utility of the methods we developed is demonstrated on a database of biological sequence alignments, BaliBase (Bahr *et al.*, 2001), which is routinely used for benchmarking alignment methods. We find that phylogenetic reconstruction based on alignment sets is significantly more accurate than the corresponding phylogeny derived from a single ClustalW MSA.

My final conclusion is that only very closely related, long sequences, with few indels to be reconstructed, and long between-gap anchors, are amenable to meaningful alignment reconstruction. I propose, than, that the prudent approach is never to use a single reconstructed MSA as the basis for further analysis, but to rely on simultaneous analysis of sets of equally likely MSAs.

Chapter 1: Introduction

Sequence alignment is the most basic analysis used in the comparative study of molecular sequences (nucleic acids and proteins). Prior to alignment, sequences can only be analyzed in isolation. Multiple sequence alignment relates sequence residues from several sequences, which enables analysis of a set of sequences as an ensemble.

Sequence alignment is the first step in many biological analyses, such as derivation of sequence similarity measures, identification of homologous sites, phylogenetic reconstruction, identification of functional domains, homology-based structure prediction and primer design. In short, it is the starting point of almost every analysis that involves the comparison of molecular data (Mullan, 2002).

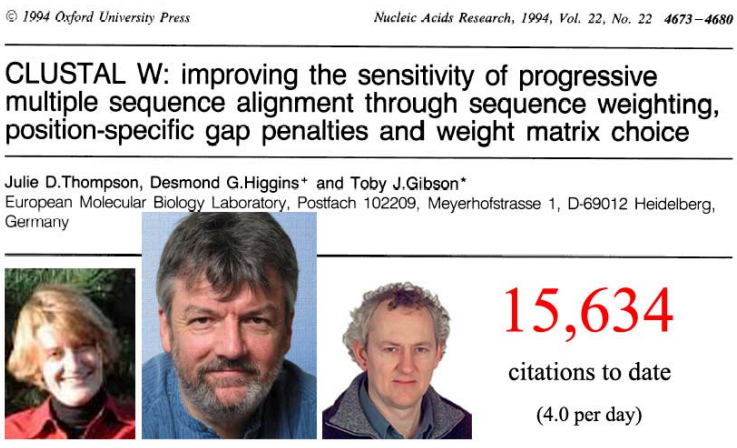


Figure 1: ***The ClustalW paper and people.***
Number of citations sampled on June 1, 2005.

The fundamental role of multiple sequence alignment is best demonstrated by noting that the paper describing the standard multiple-alignment reconstruction method,

ClustalW (Thompson *et al.*, 1994a), is the most cited paper in biology over the eleven years since its publication (Figure 1, and see http://www.in-cites.com/scientists/DesHiggins.htm).

Being fundamental ingredients in a wide variety of analyses, an issue of utmost importance is their reliability and accuracy: analyses based on erroneously reconstructed alignments are bound to be heavily handicapped (e.g., Morrison and Ellis, 1997, O'Brien and Higgins, 1998, Hickson *et al.*, 2000).

Sequence evolution

In the evolutionary context, sequence alignment is always coupled with a phylogeny. Together, the phylogeny-alignment pair provides a concise description of the evolution of a set of homologous sequences, as in Figure 2:

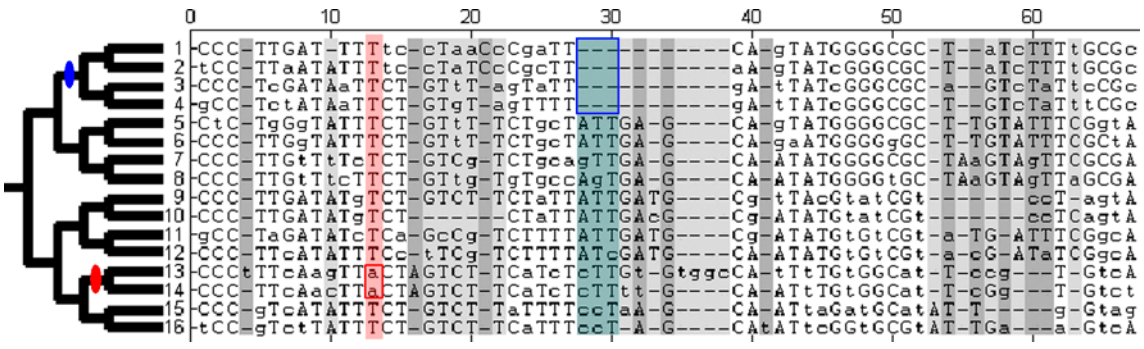


Figure 2: ***A phylogeny and a section of a 16-OTU multiple sequence alignment.*** *The sequences are related by the phylogeny (left). Each row is a sequence while columns are sets of homologous residues.*

The phylogeny summarizes the branching events that led from a single ancestral sequence and produced the several extant sequences. In the alignment, homologous residues in the several sequences are related to each other by the introduction of gaps into the sequence of actual extant residues. The introduced gaps represent insertion and

deletion events (collectively termed *indels*, blue). All residues in a column are homologs, and may have experienced substitution events (red). Since sequences change along the branches of a phylogeny, its structure is reflected in the alignment (ellipses), though with some noise resulting from the accumulation of multiple changes.

Were the detailed history of the evolution of a set of sequences known, it could have been represented in an MSA and a phylogeny, which we term the "*True*" MSA and phylogeny. For real sequences the true alignment-phylogeny pair is never known. Rather, those are the unknowns we set out to reconstruct, starting from the observed extant sequences. Thus, all empirical MSAs and phylogenies are "*Reconstructed*" ones. If one is exceptionally lucky, the reconstructed alignments and phylogenies will be identical to the true ones. The odds for that, as we shall see, are slim.

Alignment Reconstruction

The reconstruction of alignments of molecular sequences was first described by Needleman & Wunsch (1970). Since then the theory and art of sequence alignment reconstruction has flourished (see Appendix - A brief history of MSA). There has been a proliferation of alignment algorithms, aiming at the improvement of two aspects: (a) the computational feasibility and performances of alignment algorithms, and (b) the biological relevance and quality of deduced alignments. (for reviews of alignment methods, see Feng et al. 1984; Chan et al. 1992; McClure et al. 1994; Hirosawa et al. 1995; Taylor 1996, Thompson *et al.*, 1999b, Nicholas *et al.*, 2002, Notredame, 2002; for textbook treatment, see Waterman, 1995; Gusfield, 1997).

The most basic type of alignment is the pairwise alignment (PWA) of two sequences. Needleman & Wunsch (1970) first used dynamic programming for the reconstruction of global pairwise alignments. Global alignments were rendered more realistic biologically with the introduction of affine gap penalties (Altschul and Erickson, 1986), and the use of more accurate substitution matrices (Altschul, 1991, Gonnet *et al.*, 1992, Henikoff and Henikoff, 1992).

Global pairwise alignment is best described by a dot-matrix plot (Figure 3), where the two sequences are listed along the two dimensions of the matrix, and matrix entries gives the type of substitution (if any) for all pairs of residues, one from each sequence. Permissible PWAs are then all monotonically increasing paths through the matrix, with homolog residue pairs traced by diagonals, and gaps implied by horizontal or vertical segments of the path.

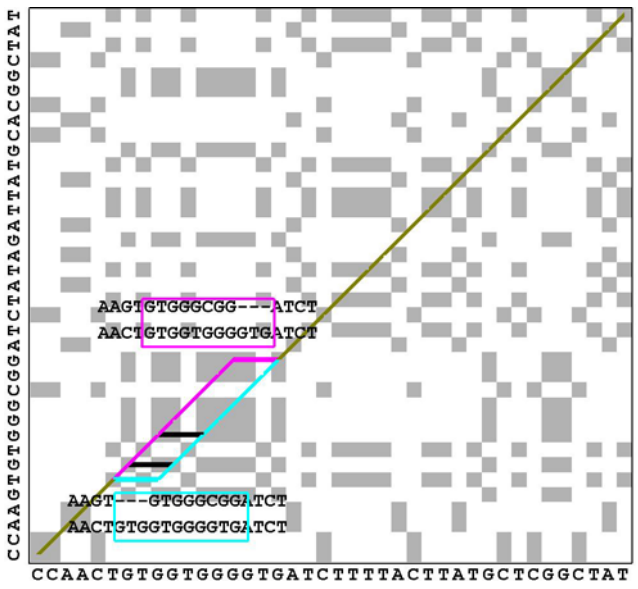


Figure 3: ***Dot-matrix representation of pairwise sequence alignment.*** *Permissible Alignments are monotonically increasing paths through the matrix, with diagonals traversing homolog residue pairs, and vertical and horizontal segments spell gaps in the relevant sequence.*

The alternative alignments through the dot-matrix are scored by assigning relative penalties to the different types of alignment columns: identity, substitution and gaps. To produce a biologically adequate PWA, the objective function used to score alignments must have penalty values that correspond to the evolutionary parameters (substitution and indel rates and distributions) that govern the sequence evolution. Given the penalties, one of the best-scoring alignments is selected arbitrarily, and retained as the reconstructed PWA. In practice, a dynamic programming algorithm can find an optimal path efficiently, and such algorithms are common to most alignment programs (Pearson and Miller, 1992). An important feature of global pairwise alignment is that any sub-alignment of an optimal alignment is optimal in itself.

Apart from solving the two-sequence problem, pairwise alignment is also a basic ingredient in multiple sequence alignment reconstruction. Pairwise alignments play two roles in MSA reconstruction: (a) all pairwise alignments of the several sequences are used to estimate preliminary sequence distances, and (b) partial MSAs are aligned to each other using a variant of the standard pairwise alignment algorithm.

Over the last twenty years, scores of MSA reconstruction methods have been developed. The most widely used method is ClustalW (Thompson *et al.*, 1994a.) ClustalW produce an MSA by progressive alignment (Feng and Doolittle, 1987) along a guide-tree, and includes internal estimation of evolutionary rates, as well as various refinements of the reconstruction process. In this study we used ClustalW as the standard in MSA reconstruction.

Progressive alignment along a tree proceeds in the following steps:

a. Estimation of a guide-tree:

    1. Estimation of all pairwise sequence distances based on all pairwise sequence alignments.

    2. Reconstruction of an approximate guide-tree, using some distance-matrix phylogenetic reconstruction method.

b. A series of pairwise profile alignments:

    1. Traversing the guide tree in a nearest neighbor order, sequences are added to a growing set of partial alignments termed profiles.

    2. At each step, standard global pairwise alignment is used to align two profiles or sequences to produce a partial MSA of the combined OTU set.

Note that when the phylogeny is known in advance, the guide tree estimation step can be skipped.

Errors in reconstructed MSAs

Many researchers routinely relay on reconstructed MSAs implicitly. This is so even though deduced sequence alignments are known to raise grave reliability and accuracy issues (Henikoff, 1991, Ellis and Morrison, 1995). Alignment reliability issues were first addressed from a theoretical, mainly mathematical, perspective (Gotoh, 1990, Goldstein and Waterman, 1992, Waterman and Vingron, 1994, Waterman, 1994, Yu and Smith, 1999, Frommlet *et al.*, 2004). Lately, several alignment algorithms were compared in terms of alignment quality, focused on the ability to reconstruct large-scale

features of reference alignments (McClure *et al.*, 1994, Thompson *et al.*, 1999b, Lassmann and Sonnhammer, 2002). In contrast, little attention has yet been given to the fine-detail quality of multiple sequence alignment (but see Thorne and Kishino, 1992, Thorne *et al.*, 1992b, Wheeler, 1995, Holmes and Durbin, 1998, Hickson *et al.*, 2000.)

A first example of errors in reconstructed MSA is presented in Figure 4. A simulated process of sequence evolution provides us with a "true" MSA (top), which is the target against which we compare a reconstructed MSA of the simulated sequences (bottom). Fully reconstructed columns are identical in both alignments (shaded). Other columns of the alignment are erroneously reconstructed, and span a sizable portion of the alignment length.

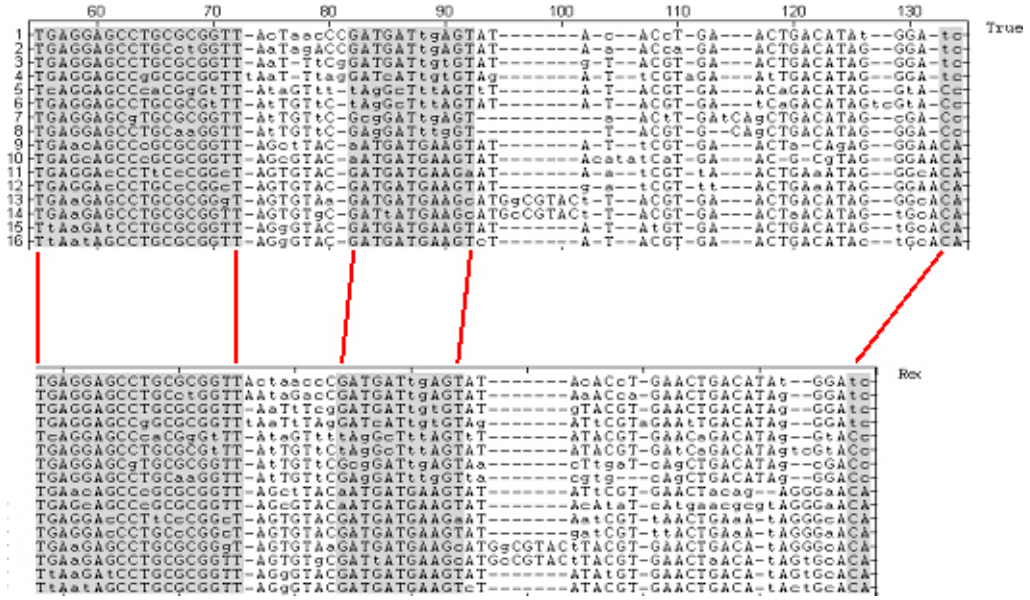


Figure 4: ***Comparison of true (top) and reconstructed (bottom) MSAs.*** *The alignments are decomposed into correctly (shaded) and erroneously (unshaded) reconstructed segments.*

The failure to correctly reconstruct the MSA stems from erroneous positioning of gaps during reconstruction. Therefore, most reconstruction errors occur near gapped columns of the true MSA, as is already evident in figure 4.

Some of the erroneously reconstructed columns in Figure 4 are in fact partially correct, and the column-based comparison is clearly too conservative. A more adequate description can be produced by comparing the two alignments at the residue and residue-pair level (see Chapter 2). The alignments of figure 4 are reproduced in Figure 5, using the more complex residue based comparison:

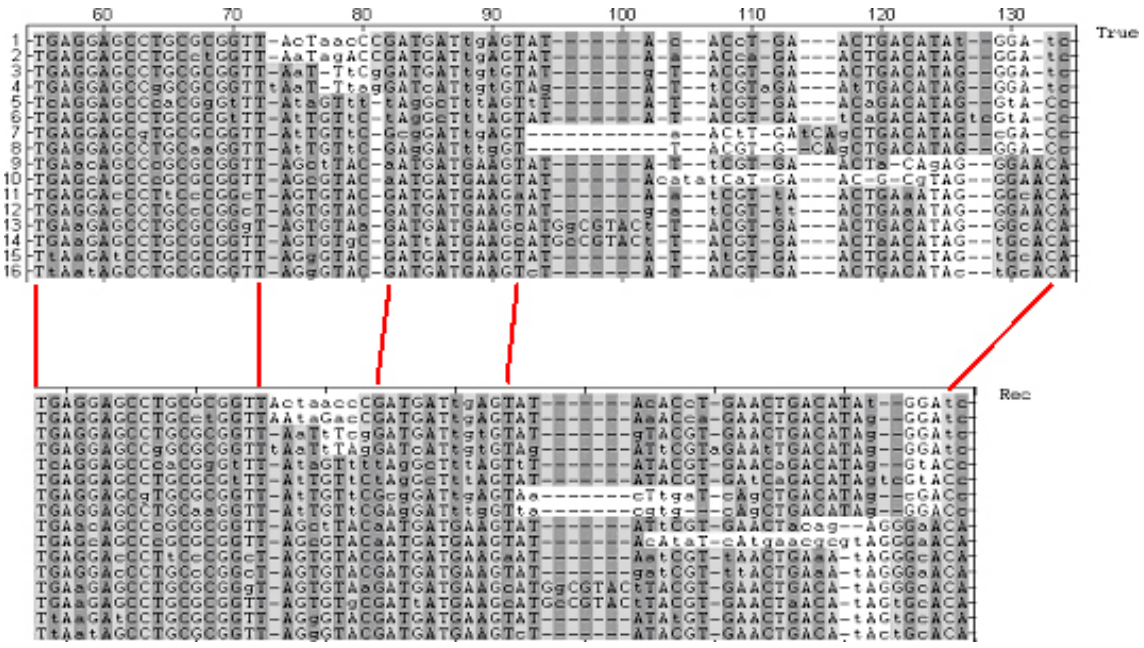


Figure 5: ***Residue based comparison of true (top) and reconstructed (bottom) MSAs.***
The alignments are decomposed into correctly (shaded) and erroneously (unshaded) reconstructed elements.

In certain cases such a comparison may help in the interpretation of the difference between the alignments. For example, the first error segment of Figure 4 can be interpreted, in light of Figure 5, as a removal of a single gap character from all sequences, albeit in a staggered fashion. The second error segment of Figure 4 is somewhat beyond easy interpretation, yet the residue-based comparison reveals that even here, while whole columns are but partially reconstructed, some subsets of OTUs (e.g., OTUS 1 through 6) are correctly reconstructed. Figures 4 and 5 provide us with a first glimpse of the complexities of MSA errors.

Motivation and aims

Although MSAs may be used for other purposes, in this study we focus on their use in the reconstruction of phylogenies. Concerned with the poor quality of many reconstructed phylogenies, we first posed the question: "Can reconstructed multiple sequence alignments be relied on implicitly when reconstructing phylogenies?" The answer was a clear and resounding "***NO!*** "

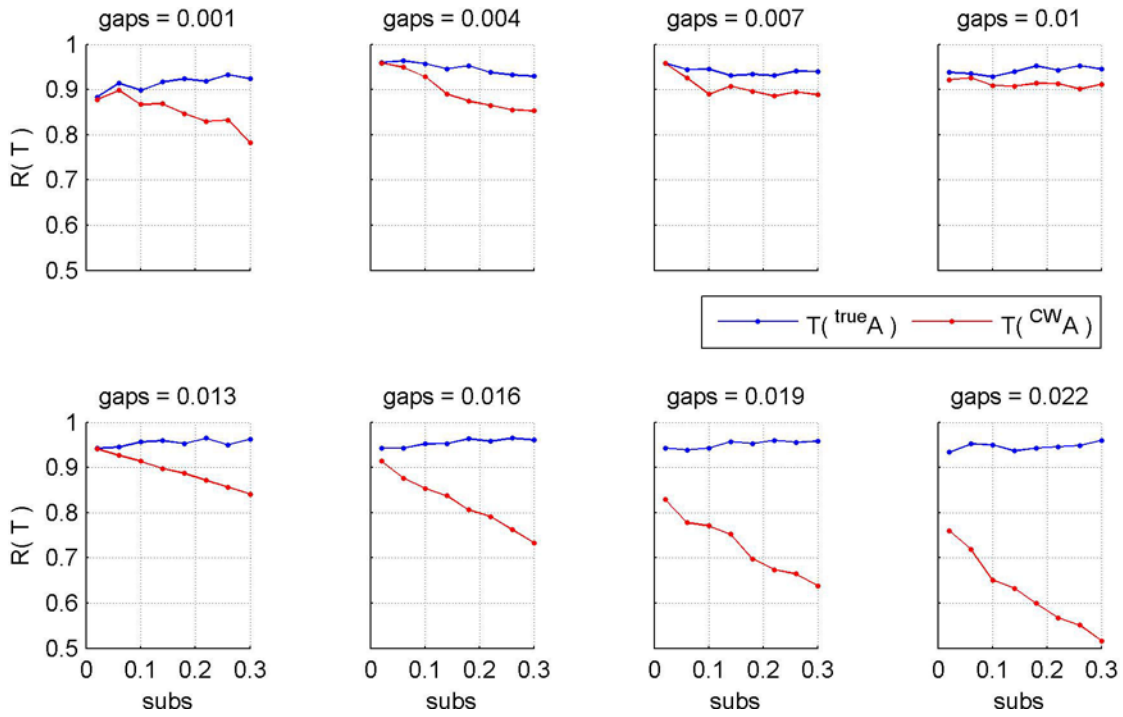


Figure 6: ***Phylogenetic reconstruction accuracy:*** *(a) BioNJ based on true alignments (blue), and (b) BioNJ based on reconstructed alignments (red). x-axis and panes are increasing sequence divergence. (See Figure 8 for details of graphs layout.)*

Figure 6 presents an example of the accuracy of phylogenetic reconstruction as a function of sequence divergence (for details Chapter 3.) As the sequence divergence increases, the reconstructed phylogenies quality is rapidly deteriorating, but only when using reconstructed, mostly erroneous, MSAs (red lines). Given the true evolutionary

alignment, phylogenetic reconstruction withstands divergence quite admirably (blue lines).

It seems, than, that current phylogenetic reconstruction methods are adequate, and that the poor quality of reconstructed phylogenies can be traced back to the poor quality of the reconstructed MSAs presented to them. It may be reasonably expected that similar conclusions apply to other MSA-dependent analyses as well.

In the first part of the study, we set out to obtain a better understanding of the sources and characteristics of MSA errors. To this end, we compare simulated true-MSAs to reconstructed MSAs, and provide a quantification of error levels encountered in the reconstructions. The characterization of MSA errors enables us to identify the major sources of alignment errors. In addition, we quantify the contribution of MSA errors to the erroneous reconstruction of phylogenetic trees. In a nutshell, MSA errors are shown to be very frequent, and their effects substantial.

We first set our attention on pairwise sequence alignment. Pairwise alignment is both the simplest case of sequence alignment, and a building block of multiple sequence alignment algorithms. Three main sources of errors are already apparent at the pairwise alignment level: (a) inadequate estimation of evolutionary parameters, (b) over-fitting due to strict optimization, and (c) arbitrary choice among co-optimal alternatives.

The major difference between pairwise comparisons and multiple sequence alignments is that in a MSA the several sequences are related by a phylogeny. The best MSA algorithms take this into account to provide better alignments. Thus, a fourth source of alignment errors is the uncertainty in alignment-guiding phylogenies.

Those sources of errors, compounded in a multiple sequence alignment, produce a plethora of error structures. We provide a characterization of the major types of alignment errors and their distribution.

The second part of the study aims at developing methods to identify MSA reconstruction errors, and devise tools through which alignment errors and uncertainties can be accounted for and managed in the context of phylogenetic reconstruction. One possible strategy may be to shift our attention from a single reconstructed MSA, to a larger set of equally likely MSAs. The construction of the alignment set is designed to produce fine-detail variability, which reflects some of the major sources of MSA reconstruction errors.

For error identification purposes, we shall use the variability within a set of alternative alignments to derive local, fine-detail, reliability measures for any candidate MSA. In simulation settings, we find that our quality measures are very accurate, and are superior to existing methods of MSA quality scoring. Although our reliability measures prove to be good predictors of MSA errors, their utility in boosting the performance of subsequent phylogenetic reconstruction is found to be marginal. We conclude that for phylogenetic reconstruction purposes, the identification of errors cannot enhance the utility of a single, poor-quality, MSA.

Another approach to account for MSA errors in phylogenetic reconstruction is to conduct a simultaneous analysis of the entire alignment set. We convert the alignment set to a phylogeny set by the use of standard methods for phylogenetic reconstruction. The consensus phylogeny derived from this phylogeny set is shown to be significantly more accurate than an analysis based on a single reconstructed MSA.

In the third and last part of the study we apply our methods to a database of real-life test cases, the BaliBase database (Bahr *et al.*, 2001). We find that our methods significantly enhance the accuracy of phylogenetic analysis. We conclude that the strategy of abandoning the single-MSA approach and replacing it by a variable MSA set is of great utility in realistic biological settings.

Chapter 2: Methods

Symbols and Acronyms

A	Alignment	${}^{true}A$, ${}^{assisted}A$, ${}^{cw}A$
AS	Alignment Set	${}^{pw}AS$, ${}^{gt}AS$
E	Error rate	${}^{pairs}E$, ${}^{res}E$, ${}^{col}E$
LRM	Local Reliability Measure	
MSA	Multiple Sequence Alignment	
M	Reliability measure	${}^{pairs}M$, ${}^{res}M$, ${}^{col}M$
N^{otu}	Number of OTUs	N^{otu}
OTU	Operative Taxonomic Unit	
PWA	Pairwise Alignment	
Q	Enropy based Quality measure	${}^{col}Q$
R	Reconstruction rate	${}^{pairs}R$, ${}^{res}R$, ${}^{col}R$, ${}^{phy}R$
S, s	Sequence set and sequences	$S = \{ s_i \}$
T	Tree (phylogeny)	${}^{true}T$, ${}^{guide}T$, ${}^{true\text{-}A}T$, ${}^{assisted\text{-}A}T$, ${}^{cw}T$, ${}^{as}T$, ${}^{itr\text{-}as}T$, ${}^{ref}T$
TS	Tree (phylogeny) Set	${}^{gt}TS$

Statistical methods

We have employed standard statistical methods, as can be found in Sokal and Rohlf (1995) and Zar (1999). Receiver-operating characteristic (ROC) analysis is described in Zweig and Campbell (1993).

Standard analysis software

Throughout this study we used ClustalW (Thompson *et al.*, 1994a) as the MSA reconstruction tool. Pairwise alignments studied in Chapter 3 were reconstructed using the ALIGN program (Pearson and Lipman, 1988). Phylogenies were reconstructed using the BioNJ program (Saitou and Nei, 1987, Gascuel, 1997), operating on pairwise distances corrected for multiple substitutions (Jukes and Cantor, 1969, Felsenstein, 1993.) Apart from these methods, I have implemented all other algorithms and analyses in the Matlab® environment.

Alignment databases

Three multiple sequence alignment databases were used in this study: EMBL-Align (Lombard *et al.*, 2002), PIR-Align (Srinivasarao *et al.*, 1999) and BaliBase version 2 (Bahr *et al.*, 2001). EMBL-Align and PIR-Align were analyzed to define the range of MSA problems that are of biological relevance. Simulation studies were limited to problems that span 80% of the empirical alignments deposited in those databases.

The test cases of Chapter 5 were derived from the BaliBase database. We used only Datasets 1-5 of BaliBase, since the other datasets focus on sequence rearrangements phenomena that are outside the scope of this study. We limited the number of OTUs analyzed to 25, by randomly drawing OTUs from BaliBase alignments with more OTUs. In addition, we have not used sequences shorter than 25 residues.

Evolutionary simulations

In a manner similar to that of the ROSE program (Stoye *et al.*, 1998), we simulated sequences, phylogenies and native alignments with the following simulation process:

An ancestral nucleic acid or protein sequence of length ℓ_0 was randomly generated. The ancestral sequence was evolved along a binary tree by duplication at tree nodes and accumulation of changes along branches. The process was repeated iteratively, producing N^{otu} extant sequences.

Changes to the sequences along branches consisted of substitutions, insertions and deletions. The simulation parameters were chosen so as to produce alignments that are comparable to real alignments by drawing them from an empirical distribution derived from real world alignments, where each database alignment provided estimates of substitution and indel probabilities, as well as indel length distribution, number of OTUs and alignment lengths. The descriptive statistics of simulated alignments correspond to 80% of the alignments in the databases, thus ensuring that the simulated alignments are of biological and practical interest.

Each simulation record provides the full evolution history, including the ancestral sequence and the ordered series of changes. For the purposes of the current study, only part of this information was used: the extant (or OTU) sequences along with the phylogeny and the native alignment, where all residues in a given column are true homologs.

We conducted a number of simulation runs exploring different aspect of the alignment problem. A typical simulation run consisted of about 100 replications for 8 levels of substitution rates and 8 levels of indel rates, for a total of 6400 cases per run. In chapter 3 and 4 we used the value of N^{otu}=16.

Comparison of MSAs

Throughout this study we relay heavily on the comparison of alternative MSAs of the same sequence set. Our measures are based on the comparison of residue-pairs, as in Thompson *et al.* (1999).

Given a set of N^{otu} extant sequences $S = \{s_{1..N^{otu}}\}$, and an MSA A of length ℓ^a, we recode A by the sequence position of the residues:

$$A = \{a_i^k\}, \qquad a_i^k = \begin{cases} \text{index of residue in } s_i \\ 0 \text{ for gaps} \end{cases}$$

Where $i \in [1..N^{otu}]$ is the OTU index, and $k \in [1..\ell^a]$ is the MSA column index.

Next, we construct the set of residue-pairs indices: for each MSA column k and OTU pair $\{i,j\}$ the index pair is:

$$p_{i,j}^k = \{a_i^k, a_j^k\}$$

and the set of all index pairs is:

$$P(A) = \{p_{i,j}^k\}_A$$

Given two MSAs, a reconstructed MSA $\boldsymbol{A}$, and a reference MSA ${}^{ref}\boldsymbol{A}$, we score each index pair of $\boldsymbol{A}$ by its occurrence in ${}^{ref}\boldsymbol{A}$. We define the residue-pair reconstruction score as:

$$ {}^{pairs}\boldsymbol{R}^{k}_{i,j} = \boldsymbol{R}(\boldsymbol{p}^{k}_{i,j}) = \begin{cases} 0: & \boldsymbol{p}^{k}_{i,j} \notin \boldsymbol{P}({}^{ref}\boldsymbol{A}) \\ 1: & \boldsymbol{p}^{k}_{i,j} \in \boldsymbol{P}({}^{ref}\boldsymbol{A}) \end{cases} $$

The binary residue-pair score can be averaged to yield the residue reconstruction rate:

$$ {}^{res}\boldsymbol{R}^{k}_{i} = \frac{\sum_{j \neq i} {}^{pairs}\boldsymbol{R}^{k}_{i,j}}{\boldsymbol{N}^{otu} - 1} $$

The most useful score is the column reconstruction rate, which is obtained by averaging the residue score:

$$ {}^{col}\boldsymbol{R}^{k} = \overline{{}^{res}\boldsymbol{R}^{k}_{i}} $$

Note that for the case of pairwise alignment (PWA), the three levels of comparison are identical.

Finally, the overall alignment score, relative to the reference ${}^{ref}\boldsymbol{A}$, is:

$$ {}^{ali}\boldsymbol{R} = \overline{{}^{col}\boldsymbol{R}^{k}} $$

All the above scores take values on the interval [0..1], with 1 for full agreement between the two MSAs.

When the reference MSA is the true alignment from simulations, we interpret the above series of scores as reconstruction rates. We also define a series of error rates, which records the presence of any error in specific alignment elements:

$$*\boldsymbol{E} = \begin{cases} 0: & *\boldsymbol{R} = 1 \\ 1: & *\boldsymbol{R} < 1 \end{cases}, \qquad \text{for } * \in \{\text{pairs, res, col}\}$$

For the residue-pairs level the error rate is simply the complement of the reconstruction rate, and both are analogs of Thompson *et al.* (1999) measure SPS. For the residue and column resolutions, the error rates are more strict measures of accuracy than the reconstruction rates. The column error rate, ${}^{col}\boldsymbol{E}$, is the analog of Thompson *et al.* (1999) measure CS.

Another use of the comparison scores is when a reconstructed MSA $\boldsymbol{A}$ is compared to a set of alternative alignments:

$$\boldsymbol{AS} = \left\{ {}^{alt}\boldsymbol{A}_i \right\}.$$

In this context we average the MSA-pair scores over the set alignments, to produce our series of local reliability measures:

$$*\boldsymbol{M}(\boldsymbol{A} \mid \boldsymbol{AS}) = \overline{*\boldsymbol{R}(\boldsymbol{A} \mid {}^{ref}\boldsymbol{A} \in \boldsymbol{AS})}, \qquad \text{for } * \in \{\text{pairs, res, col}\}$$

The local reliability measures, $*\boldsymbol{M}$, take values on the interval [0..1], with 1 for full support.

Comparison of phylogenies

An N^{otu} fully resolved unrooted phylogeny is composed of ($2 \cdot N^{otu} - 3$) branches. For our purposes we focus on the tree topology, and ignore the branch lengths. In such settings, the N^{otu} terminal branches are trivial and non-informative, since they appear in any phylogeny of the OTUs. Thus, the remaining ($N^{otu} - 3$) internal branches uniquely define the tree topology.

Each internal branch divides the set of OTUs into two complementing subsets:

$$b_i = \left(\{o\}_i , \overline{\{o\}_i} \right),$$

where $i \in \left[1..N^{otu-3} \right]$ is the branch index, and o are the OTU indices. The phylogenetic tree topology is than defined by:

$$T = \{b_i\}$$

Given two phylogenies, the symmetric tree distance is defined as the number of partitions that differ between the two trees (Felsenstein, 2004):

$$D(T_1, T_2) = \left| T_1 \setminus T_2 \right| = N^{otu} - 3 - \left| T_1 \cap T_2 \right|$$

When one of the trees is a reference against which we compare a reconstructed phylogeny, we define the phylogenetic reconstruction rate as the normalized symmetric similarity:

$$^{phy}\boldsymbol{R}(\boldsymbol{T} \mid {}^{ref}\boldsymbol{T}) = \frac{\left|\boldsymbol{T} \cap {}^{ref}\boldsymbol{T}\right|}{\boldsymbol{N}^{otu} - 3}.$$

The phylogenetic reconstruction rate, $^{phy}\boldsymbol{R}$, take values on the interval [0..1], with 1 for full reconstruction.

Consensus phylogeny

To derive a consensus phylogeny from a set of alternative trees, we used a variant of the majority-rule consensus method (e.g. Felsenstein, 2004), which we term the "*member consensus phylogeny*". Given a set of alternative trees over the same OTUs,

$$\boldsymbol{TS} = \left\{ {}^{alt}\boldsymbol{T}_i \right\},$$

we score each tree by its mean support in reference to all the other trees in the set:

$$\boldsymbol{S}({}^{alt}\boldsymbol{T}_i \mid \boldsymbol{TS}) = \overline{{}^{phy}\boldsymbol{R}({}^{alt}\boldsymbol{T}_i \mid {}^{ref}\boldsymbol{T} \in \boldsymbol{TS})}$$

Our member consensus phylogeny, $^{as}\boldsymbol{T}$, is than randomly chosen from among the $\boldsymbol{TS}$ trees with maximal support. Note that our method differs from standard majority-rule consensus in that the consensus must be a member of the set. This enable us to by-pass the issue of partially resolved consensus trees (Felsenstein, 2004).

Chapter 3: Alignment errors and their effects

In this part of the study we use simulations to provide us with true phylogenies and true MSAs that serve as a reference against which to compare reconstructed MSAs and phylogenies. The simulations are generated in a fashion similar to that of the ROSE program (Stoye *et al.*, 1998, and see Chapter 2) Each simulation replicate produces three datasets. For N^{otu} OTUs:

a. The extant OTU sequences, without gaps, $\boldsymbol{S} = \left\{ \boldsymbol{s}_{1..N^{otu}} \right\}$.

b. The true MSA of the OTU sequences, ${}^{true}\boldsymbol{A}$.

c. The true phylogeny of the OTUs, ${}^{true}\boldsymbol{T}$.

In the second step, the extant OTU sequences are used as input to the ClustalW program, to produce reconstructed MSAs. We reconstruct two MSAs from each sequence set, once using the true phylogeny as a guide tree, and a second time using the ClustalW internal estimation of a guide tree. Thus, at this stage of the analysis we have an additional phylogenetic tree, the ClustalW guide tree, ${}^{guide}\boldsymbol{T}$, and three MSAs:

a. The true MSA from the simulation step, ${}^{true}\boldsymbol{A}$.

b. A reconstructed MSA based on the true phylogeny,
${}^{assisted}\boldsymbol{A}$ = ClustalW ($\boldsymbol{S}$ | ${}^{true}\boldsymbol{T}$).

c. A reconstructed MSA based on the ClustalW guide tree,
${}^{cw}\boldsymbol{A}$ = ClustalW ($\boldsymbol{S}$ | ${}^{guide}\boldsymbol{T}$).

In the third step, we use the three MSAs to derive distance matrices that are then analyzed by the BioNJ method to produce three reconstructed phylogenies. We end with five phylogenies:

a. The true phylogeny from the simulation, ${}^{true}\boldsymbol{T}$.

b. The guide-tree estimated by ClustalW, ${}^{guide}\boldsymbol{T}$.

c. The BioNJ phylogeny based on the true MSA, ${}^{true\text{-}A}\boldsymbol{T}$ = BioNJ (${}^{true}\boldsymbol{A}$).

d. The BioNJ phylogeny based on the assisted ClustalW MSA, with the true phylogeny as guide tree, ${}^{assisted\text{-}A}\boldsymbol{T}$ = BioNJ (${}^{assisted}\boldsymbol{A}$).

e. The BioNJ phylogeny based on the standard ClustalW MSA, with the guide tree derived from all pairwise alignments, ${}^{cw}\boldsymbol{T}$ = BioNJ (${}^{cw}\boldsymbol{A}$).

	Sequences	MSAs	Trees
Simulation	$\boldsymbol{S} = \{ \boldsymbol{s_i} \}$ Extant OTUs	1. ${}^{true}\boldsymbol{A}$	1. ${}^{true}\boldsymbol{T}$
ClustalW	-	2. ${}^{assisted}\boldsymbol{A}$ = ClustalW ($\boldsymbol{S}$ | ${}^{true}\boldsymbol{T}$). 3. ${}^{cw}\boldsymbol{A}$ = ClustalW ($\boldsymbol{S}$ | ${}^{guide}\boldsymbol{T}$), standard ClustalW.	2. ${}^{guide}\boldsymbol{T}$
BioNJ	-	-	3. ${}^{true\text{-}A}\boldsymbol{T}$ = BioNJ (${}^{true}\boldsymbol{A}$) 4. ${}^{assisted\text{-}A}\boldsymbol{T}$ = BioNJ (${}^{assisted}\boldsymbol{A}$) 5. ${}^{cw}\boldsymbol{T}$ = BioNJ (${}^{cw}\boldsymbol{A}$), standard BioNJ.

Table 1: ***The data structures used in this chapter and their dependencies.***
Rows are analyses types and columns are output data types.

Table 1 summarizes the relationships among the models used in this chapter. Note that in real world sequence analysis problems the true phylogeny and MSA are not available.

Overall error levels in reconstructed MSAs and phylogenies

First we present the overall reconstruction rates encountered in MSA reconstruction. The reconstruction rate we use take values in the range [0..1], with 1 for full success (see Chapter 2). Figure 7 summarizes the mean reconstruction rates for ClustalW MSAs, ${}^{cw}\boldsymbol{A}$, as a function of the sequence divergence. The residue-pairs reconstruction rate, ${}^{pairs}\boldsymbol{R}$, range from ~95±2% for very closely related sequences to 10±7% for very distantly related sequences, with a monotonic dependency on the evolutionary rates.

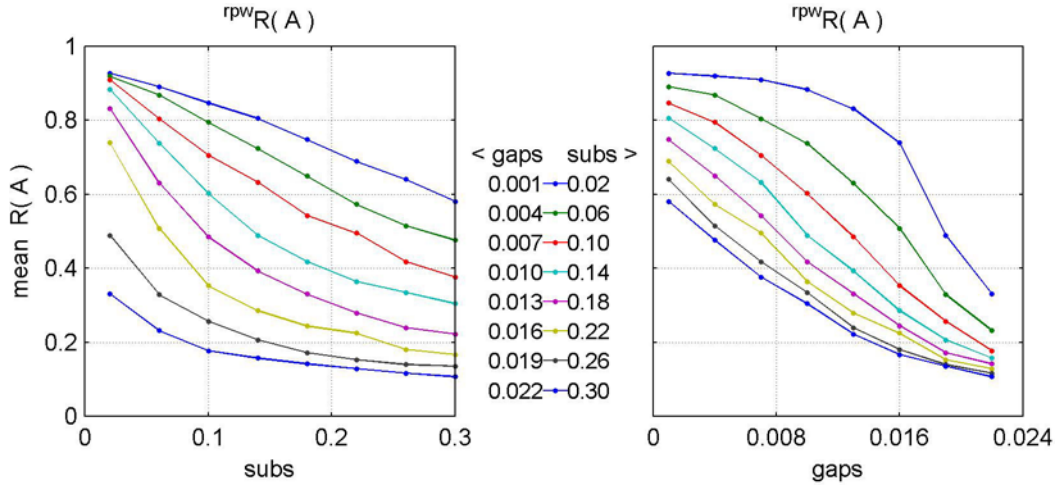


Figure 7: ***Mean ClustalW reconstruction rate as a function of sequence divergence.***

Graphs layout:

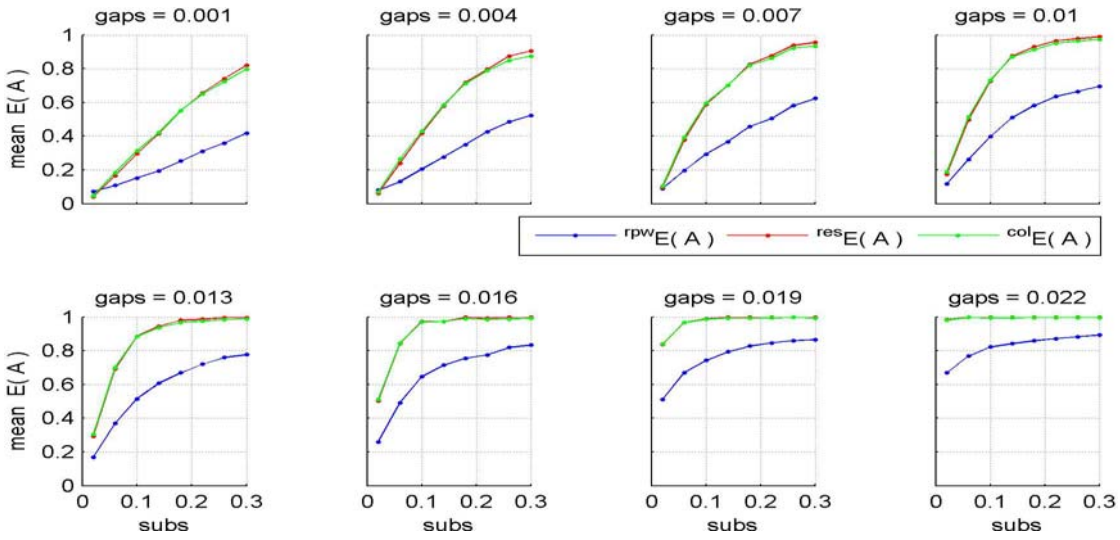
We report various metrics as a function of two sequence divergence parameters, substitution rate and indel rate. To visualize the surface traced by the response metrics, we provide two orthogonal projections of the surface. The left pane presents the metric as a function of the substitution rate (abscissa) for several values of the gaps parameter (lines). In the right pane the roles of the two parameters as abscissa and lines are switched, while the ordinate retains its role as the metric value. Each dot is an average over 100 simulation replications at one combination of 8 substitution levels and 8 indel levels, for 16 OTUs. Standard errors are reported in the text where appropriate.

I have opted not to include standard errors in the graphs, since this would clutter them beyond comprehension. I have conducted only those tests that I deemed interesting and important. For example, there is no point in providing tests for all scores of the 8x8=64 combinations of the simulation parameters, since what is of interest here is only the overall trend of dependence and the extreme values attained.

In terms of error rates (Figure 8), the residue-pair error rate, ${}^{pairs}\boldsymbol{E}$ (blue), is simply the complement of the reconstruction rate, and both are analogs of Thompson *et al.* (1999) measure SPS. Requiring that all pairs for a specific residue be correctly reconstructed yields the residue error rate, ${}^{res}\boldsymbol{E}$ (red), while requiring that the entire column will be correctly reconstructed yields the column error rate, ${}^{col}\boldsymbol{E}$ (green).

The residue and column error rates, ${}^{res}\boldsymbol{E}$ and ${}^{col}\boldsymbol{E}$, are almost identical, and ${}^{col}\boldsymbol{E}$ is the analog of Thompson *et al.* (1999) measure CS. Apart from very closely related sequences, The column error rate is higher than 50%, and rapidly reaches 100%, that is, mis-reconstruction of all MSA columns. Since this measure may be too harsh, in what follows we will mainly refer to the residue-pair reconstruction rate.

a. Abscissa: substitution rate; panes: indel rate.



b. Abscissa: indel rate; panes: substitution rate.

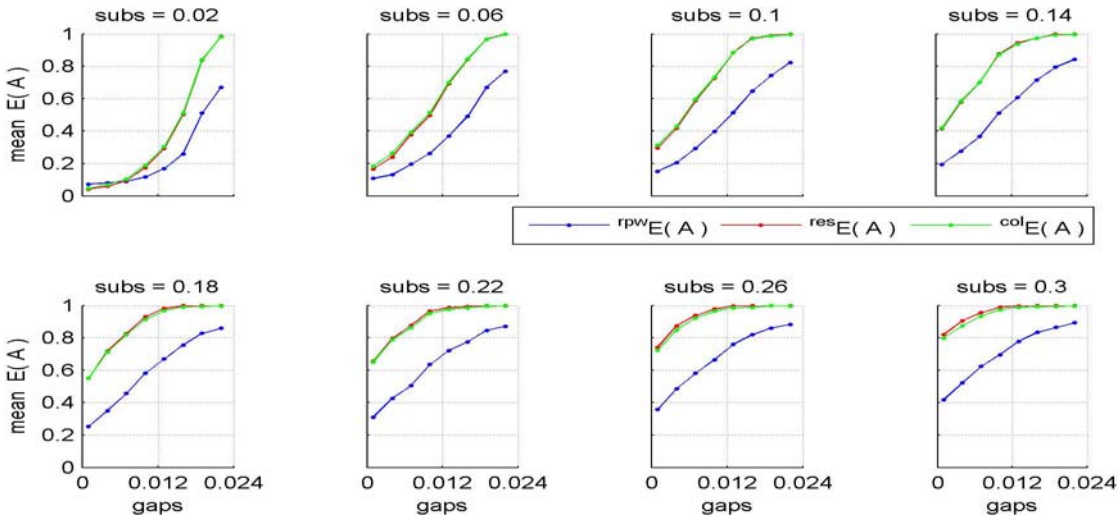


Figure 8: ***Mean ClustalW error rates as a function of sequence divergence.*** *Three errors rates are reported: residue-pairs error rate (blue), residue error rate (red), and column error rate (green).*

Graphs layout:

When comparing several metrics as a function of the two sequence divergence parameters, we report either of two projections: (a, top) the several responses (lines) as a function of substitution rate (abscissa), in a series of panes for constant values of indel rate (value indicated above panes), and (b, bottom) the roles of the two parameters as abscissa and panes are switched, while the ordinate retains its role as the metrics value (lines).

When using reconstructed MSAs to estimate phylogenies, MSA errors may cause errors in the reconstructed phylogenies (Figure 9). As sequence divergence increases, the phylogenetic reconstruction rate, $^{phy}R(^{cw}T)$, drops dramatically (Figure 9, red lines). In contrast, phylogenies reconstructed from true alignments retain high reconstruction rates, $^{phy}R(^{true\text{-}A}T)$, even with very high sequences divergence (green lines). There is a high correlation (r=0.56, p-value<10^{-12}) between MSA reconstruction rates and phylogenetic reconstruction rates, but the deterioration in phylogenetic reconstruction rates is less sharp than the deterioration in MSA reconstruction rates $^{pairs}R(^{cw}A)$ (blue lines).

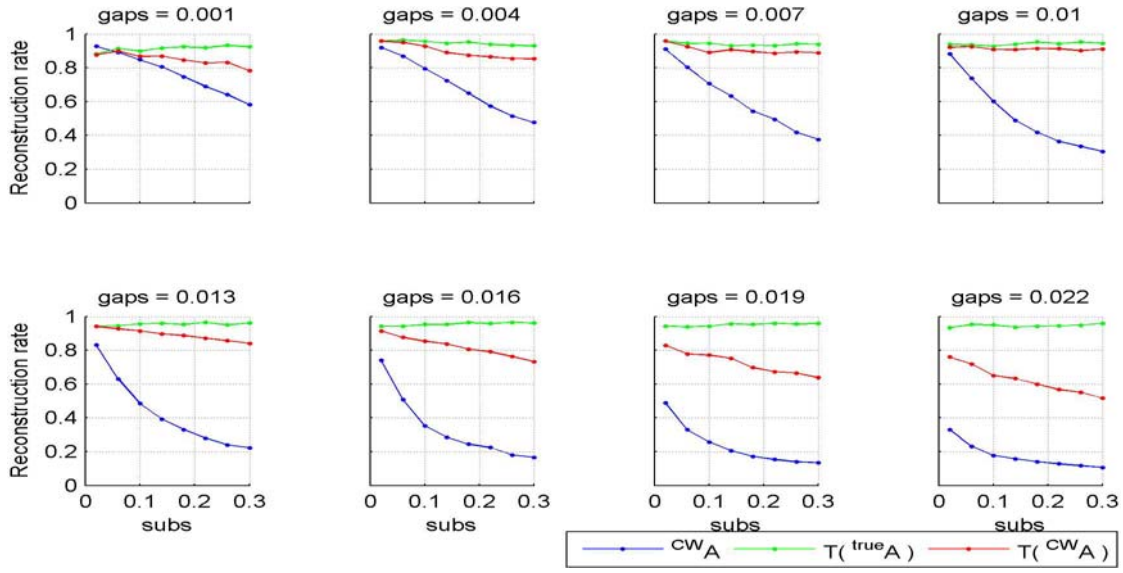


Figure 9: ***Phylogenetic reconstruction rates based on true and reconstructed MSAs.***
(a) Phylogenies derived from trueMSAs (green), (b) phylogenies derived from reconstructed MSAs (red), and (c) comparison to MSA reconstruction rate (blue). (See Figure 8 for details of graphs layout.)

Pairwise alignment errors

We start our characterization of alignment errors by considering the simplest case of sequence alignment – pairwise alignment (PWA) of two sequences. In addition to being a special case of MSA, pairwise alignments are also used as building blocks in MSA reconstruction algorithms. In this part, pairwise alignments were reconstructed using the "ALIGN" program (Pearson and Lipman, 1988), which is the standard implementation of the affine gap cost algorithm.

The most common use of alignment algorithms is that which employs the program's default penalty scores (ALIGN's DNA defaults are: *match=5; mismatch=-4; gap-open=-16; gap-extent=-4*). The default parameters are thought to be adequate for a wide range of practical problems, and are indeed a reasonable choice when no prior knowledge of evolutionary parameters is available. It is expected, however, that using penalty scores that corresponds to the true evolutionary parameters, will produce better quality alignments.

In the pairwise context, the phylogeny is reduced to a single branch, whose length is the divergence between the two sequences. The topology of this phylogeny is unique, and therefore trivial. For pairwise alignment, than, the entities we compare reduce to those in Table 2.

	Sequences	PWAs	Divergence
Simulation	$\boldsymbol{S} = \{ s_1, s_2 \}$ Two extant OTUs	1. ${}^{true}\boldsymbol{A}$	1. ${}^{true}\boldsymbol{D}_{1,2}$ and 2. ${}^{true}\boldsymbol{P} = \{native\ penalties\}$
ALIGN	-	2. ${}^{assisted}\boldsymbol{A}$ $=$ ALIGN ($\boldsymbol{S} \mid {}^{true}\boldsymbol{P}$). 3. ${}^{def}\boldsymbol{A}$ $=$ ALIGN ($\boldsymbol{S} \mid {}^{default}\boldsymbol{P}$), standard ALIGN.	

Table 2: ***The PWA data structures used in this chapter and their dependencies.***
Rows are analyses types and columns are output data types.

Distribution of pairwise alignment errors

The overall reconstruction rate $\boldsymbol{R}({}^{def}\boldsymbol{A})$, is dependent on the actual divergence of the sequences, with reconstruction rates that rapidly deteriorate with increasing sequence divergence (Figure 10, red lines). Using the default penalty values, although a widespread practice, may introduce a bias that will result in reconstruction errors. To quantify the level of errors resulting from inadequate penalties, we repeat the analysis using the exact penalty scores corresponding to the true alignment (green lines).

The PWA reconstruction rates achieved when the true evolutionary parameters are known in advance, $\boldsymbol{R}({}^{assisted}\boldsymbol{A})$, are only marginally higher than reconstruction rates achieved when utilizing default values, with average improvement of ~3% and peaking at ~10%. It follows that although appropriate penalties are desirable, using the default values is by no means the foremost source of errors.

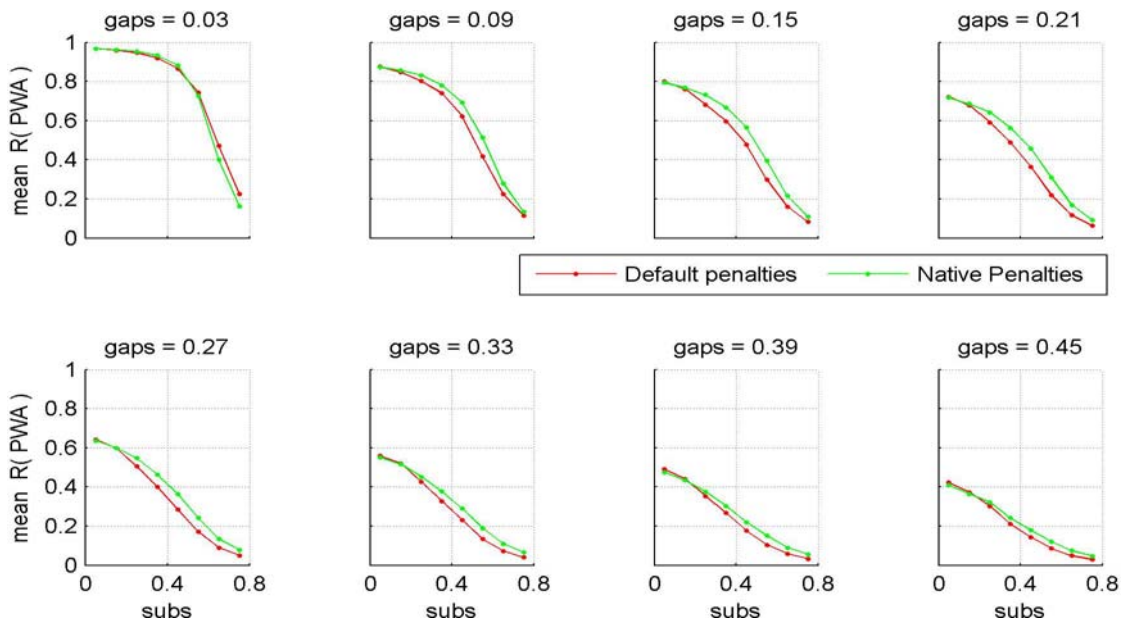
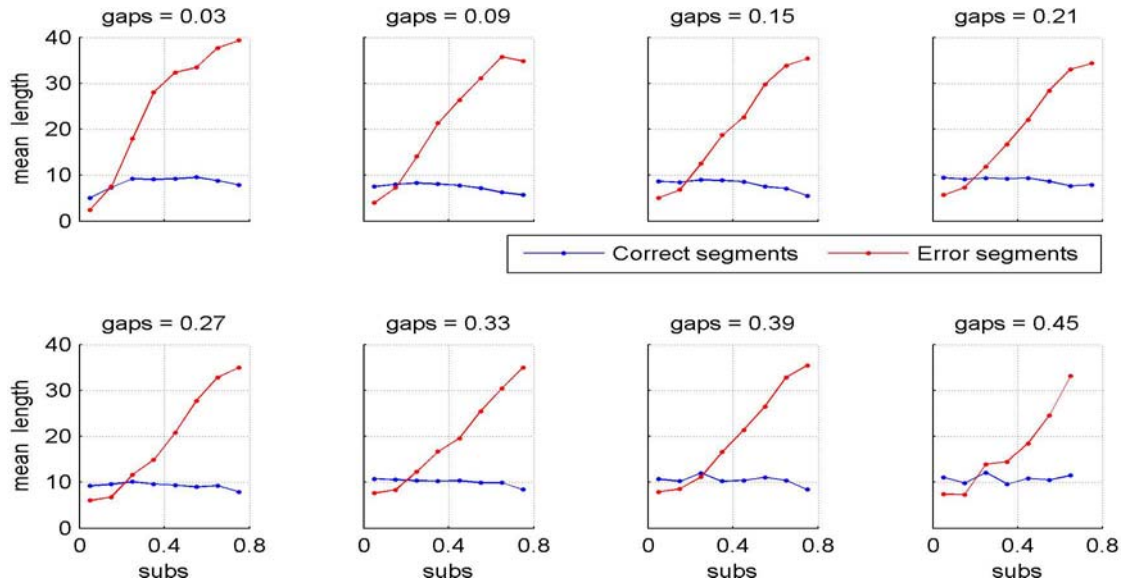


Figure 10: ***PWA reconstruction rates when using default (red) or true (green) penalties.*** *(See Figure 8 for details of graphs layout.)*

Since in providing the true parameters we utilized all the available prior knowledge, the resulting reconstruction rates represent the maximum reconstruction level that can be attained by PWA programs such as ALIGN. We must emphasize that the practice of providing true parameters is not applicable to real world problems, where the true alignment in not known in advance. Even under such favorable conditions, PWA programs are far from foolproof, and the level of errors can be quite high. We proceed to further characterize these unavoidable errors.

Given a reconstructed PWA and the corresponding true alignment, both alignments can be decomposed into alternating alignment segments where erroneously aligned subsequences are flanked by correctly aligned segments, and vice versa. Correctly reconstructed segments are identical in both alignments, while erroneous segments in the reconstructed PWA correspond to mis-reconstructed segments of the true alignment.

a. Mean number of residues in correctly and erroneously reconstructed segments



b. Mean number of indels and gap characters per error segment

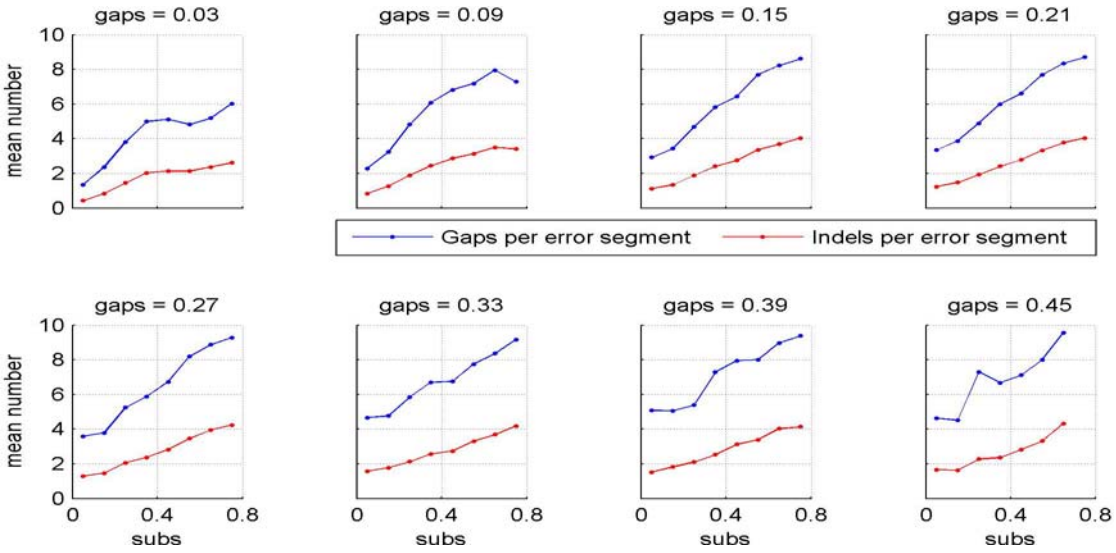


Figure 11: ***Comparison of reconstructed and true PWAs by segmentation into correctly and erroneously reconstructed segments***
(a) Error (red) and correct (blue) segment lengths as a function of sequence divergence. (b) Mean number of indels (red) and gap characters (blue) per error segment. (See Figure 8 for details of graphs layout.)

First we note that the mean length of error segments (Figure 11.a, red lines) increases dramatically with increasing substitution rate, while the mean length of correctly

reconstructed segments remains fairly stable (Figure 11.a, blue lines). We also note that the mean numbers of indel events and gap characters (Figure 11.b) increases with increasing substitution rate as well.

For the easy cases of closely related sequences, the error segments are short and are frequently the result of a single indel event erroneously positioned. As the two sequences are farther diverged errors multiply. At the same time, near-by indel events in the true alignment are interfering with one another to produce error segments where multiple indels are simultaneously misplaced. At yet higher divergence rates, the error segments get longer and longer, with relatively short intervening correct segments, until almost the whole reconstructed alignment consists of error segments.

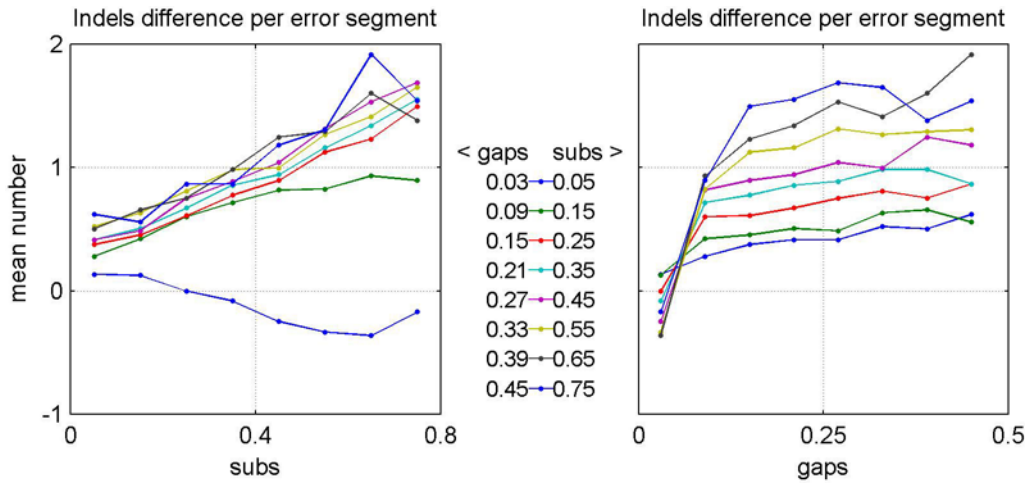


Figure 12: ***Difference of the number of indels between error segments and the corresponding true alignment segments.*** *(See Figure 7 for details of graphs layout.)*

Examining error segments, we note that the reconstruction algorithm introduce a systematic bias towards shortening the alignment and reconstructing fewer indel events than are present in the true alignments (Figure 12).

Apart from cases with very few gaps, reconstructed error segments contain fewer indel events, and are shorter, then the corresponding true segments. This is a bias resulting from the strict optimization of the objective function, coupled with the fact that for the same number of matches, shorter alignments usually score better than longer ones.

Characterization of PWA errors

Considering the objective function scores, reconstruction errors can be classified into two types:

1. Co-optimal alignment segments. Under any scoring function, many different alignments may attain the maximal score. All these alignments are equivalent, and without outside knowledge there is no way to select one of them as the "best" alignment. The alignment produced by PWA programs is an arbitrary one from the set of co-optimal alignments.

2. Sub-optimal true alignment segments. The true alignment, being some concrete realization of a stochastic process, rarely reproduces the expected frequencies of column types. This leads to the situation where an erroneous alignment segments can be assigned a higher score then the true alignment segment even by an exact scoring function. In other words, true alignments are a-priory expected to be sub-optimal in many of their elements. In contrast, the reconstructed PWA is always, by definition, optimal by the current objective function, as is any segment of the reconstructed alignment.

To enumerate the effects of co- and sub-optimality, we compare the objective function scores of error segments in the reconstructed PWA to those of the corresponding mis-reconstructed true segments. Where the scores are the same, the error can be attributed to co-optimality. Otherwise, the score of the true segment is always lower, and the error is the result of sub-optimality (Figure 13).

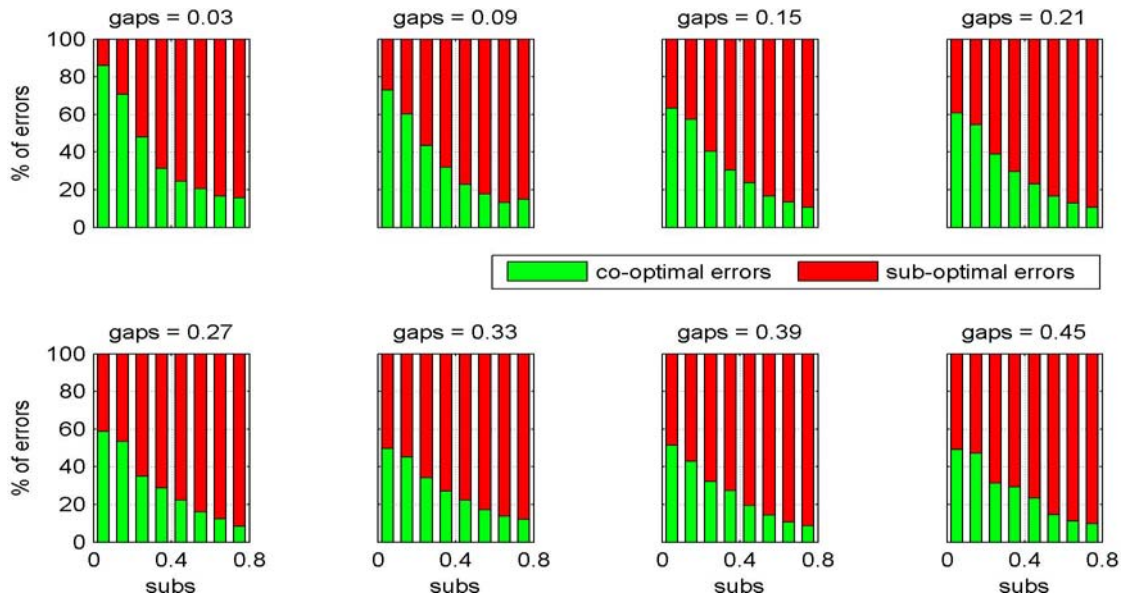


Figure 13: ***Relative frequency of error segments where the corresponding true segment is co-optimal (green) or sub-optimal (red).*** *(See Figure 8 for details of graphs layout.)*

We note that even under the most favorable conditions of close sequence relatedness, sub-optimality accounts for at least 50% of all errors. That is, the alignment is over-fitted spuriously to maximize the objective function score.

Among the simple, isolated, error segments, several types of frequent errors can be discerned (figure 14):

1. Shift error: a single indel event is erroneously positioned while its length is preserved. This is the simplest of all reconstruction errors, and the most frequent

in cases of closely related sequences. The length of the error segment is not determined by the length of the misplaced gap, but rather by the difference of the true and erroneous positions. The range of the error segment resulting from a single position error is increasing with higher substitution rates.

2. Split error: a single indel event is reconstructed as two indel events, either on the same sequence or one event on each sequence. The true indel length may not be preserved in any of the two erroneous indels, but the difference of gap content between the two sequences is the same. The true indel position is not necessarily preserved, but can be reproduced in one of the two erroneous indels.

3. Merge error: two indel event, wither on the same sequence or one on each sequence, are reconstructed as a single indel event. Again, the difference of gap content between the two sequences is preserved.

4. De-novo error: two indels of the same length are introduced, one into each of the sequences, where no indel was present in the true alignment. This type of error can be regarded as the extreme case of a split error.

5. Purge error: two equal length true indels, one on each of the sequences, are not reconstructed at all, and the resulting error segment is devoid of gaps. It as the extreme form of the merge error.

6. All other errors are designated "*Complex errors*"

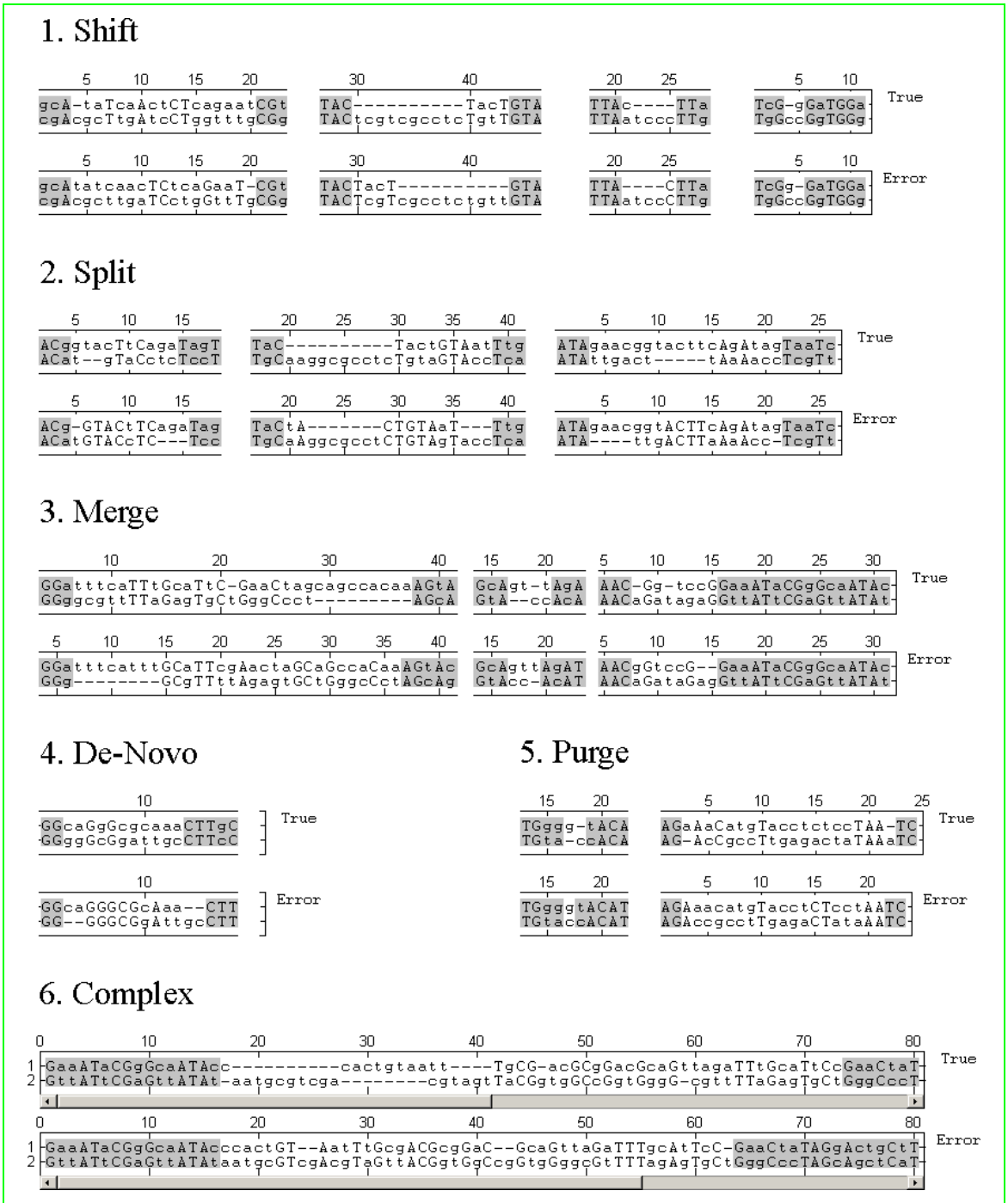


Figure 14: Examples of the five simple error types and a complex error.

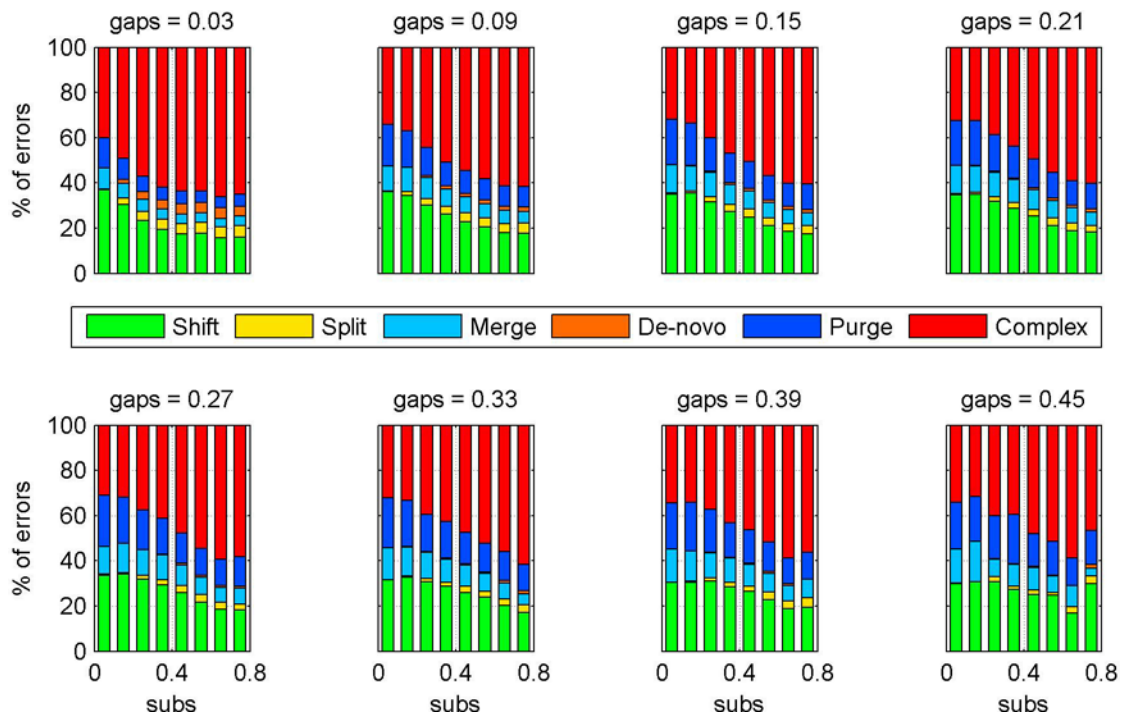


Figure 15: ***Relative frequencies of the six error types.*** *(See Figure 8 for details of graphs layout.)*

As sequence divergence increases, the simple errors types 1-5 account for fewer cases of the overall errors (Figure 15). Among the errors affecting two indel events, the errors that result in fewer indels, merge and purge (blues), are much more frequent than the errors resulting in more indel events, split and de-novo (oranges). This is another demonstration of the bias towards the minimization of inferred indel events. Note that the shift error (green), the simplest of all, is also the commonest among the simple error types, and may therefore deserve special attention.

Multiple sequence alignment errors

To study the errors in MSA reconstruction, we compared true MSAs from simulations to reconstructed MSAs produced by the ClustalW program (Thompson et al. 1994), at its default values. Note that ClustalW employs internal estimation of evolutionary parameters to derive penalty values, so the default values are even less critical than the PWA defaults used by ALIGN. On the other hand, progressive alignments use approximate phylogenies as guide trees, which may be critical to their performances (Lake, 1991).

Comparing the reconstructed alignments, ${}^{cw}\boldsymbol{A}$, to the true alignments from simulation, ${}^{true}\boldsymbol{A}$, we first note that reconstruction errors occur much more frequently in columns with gaps than in "anchor" columns (i.e., columns with no gaps).

	Anchor	Gapped	Total
Correct	419703 (0.319)	112621 (0.086)	532324 (0.405)
Error	369167 (0.281)	413944 (0.315)	783111 (0.595)
Total	788870 (0.600)	526565 (0.400)	1315435 (1.000)

Table 3: ***Number (and frequency) of columns in error and correct segments, classified as anchor vs. gapped columns.***
Substitution rate=0.1; Indel rate 0.007;

Table 3 presents the frequencies of errors for anchor and gapped columns, for one combination of simulation parameters (Substitution rate=0.1; Indel rate 0.007;). Only 40% of the columns are correctly reconstructed, and the vast majority of those are anchor columns. The error rate in anchor columns is 47%, whereas in gapped columns the error rate reaches 79%.

The difference of error rates between anchor and gapped columns reflects the nature of the problem: after all, alignment reconstruction proceeds through the positioning of gaps, and where there are few gaps to misplace, there are few errors. Yet, this does not mean that anchor columns are immune to error. In fact, misplaced gaps can have quite a long range, affecting anchor as well as gapped columns.

In order to classify reconstruction errors, we divide the length of the alignment into segments of consecutive columns, where correctly aligned segments delimit error segments. For each error segment we can then compare the true indel structure to the erroneously deduced one.

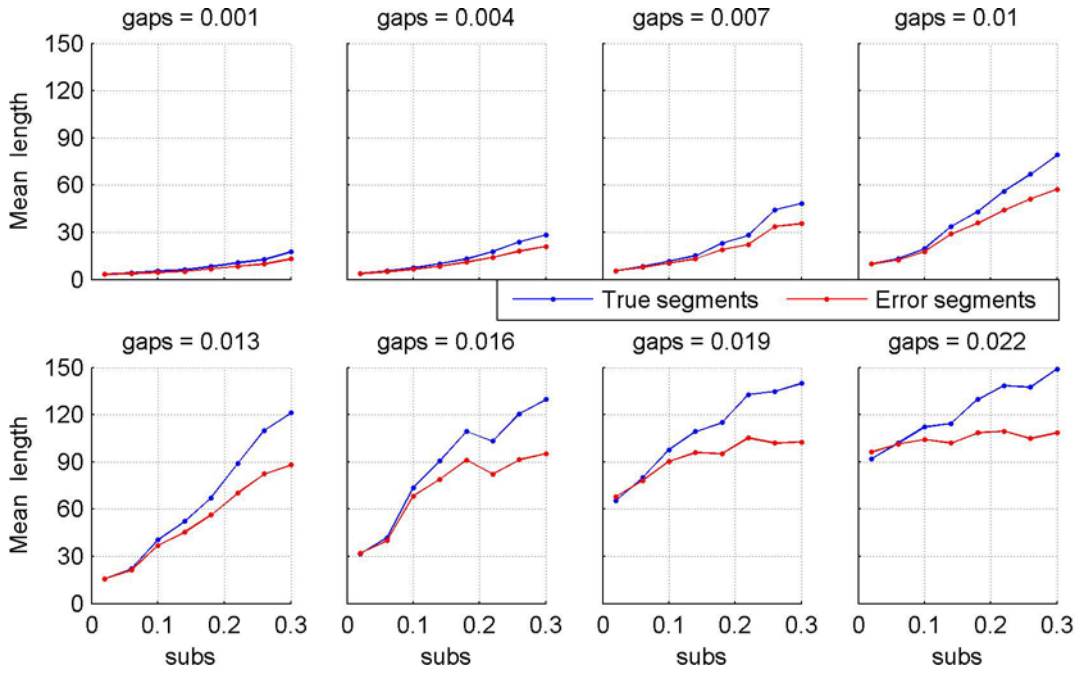


Figure 16: ***Error segments sizes as a function of sequence divergence.*** *(a) Mean length of error segments (red), and (b) Mean length of the corresponding true segments (blue). (See Figure 8 for details of graphs layout.)*

In high quality reconstructions, error segments are short and wide apart, and encompass few indels. As the overall error rate increases, so does the length of error segments (Figure 16). An erroneously reconstructed segment of an MSA can contain any number

of anchor and gapped columns that are different in the native and reconstructed alignment. Note that the true MSA segments that were erroneously reconstructed (blue), are longer than the reconstructed segments (red), and that the discrepancy increases with sequence divergence. Since the number of residues in both segments is identical, the shortening of reconstructed segments is wholly due to a lower gap character content in those segments.

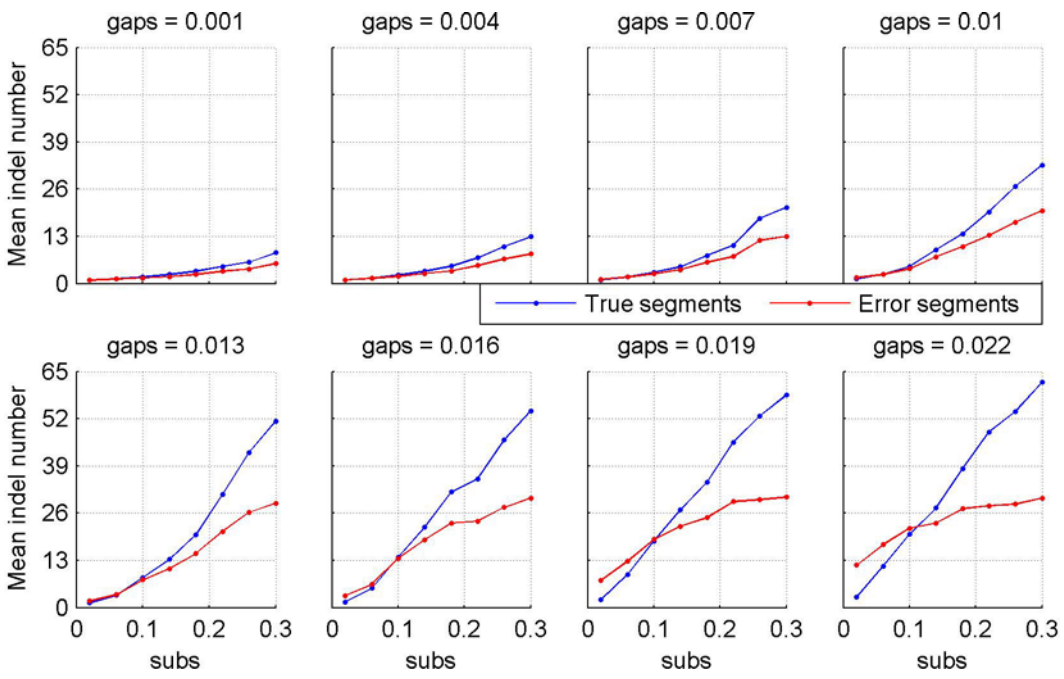


Figure 17: ***Mean number of indel events per error segment.***
(a) Indel content in the reconstructed error segments (red), and (b) in the corresponding true segments (blue). (See Figure 8 for details of graphs layout.)

This bias is even more pronounced when comparing the number of indel events in error segments (figure 17, red) to the true number of indel events that should have been reconstructed (blue). Errors consisting of misconstruction of very few indel events are prevalent when the number of substitutions is small and where indel events are rare, coupled with long intervening anchor stretches. The presence of conserved anchor stretches isolates and limits the range of the erroneous segment. As evolutionary rates

increase, the density of gapped columns rises, and errors at near positions are merged to produce longer error segments, comprising many simultaneously misplaced indel events. In such cases, the overall result is of a combinatorial nature, and is very hard to interpret.

To probe the fine details of this phenomenon, we categorized errors by the number of indel events involved in the true and erroneous segments. Table 4 presents the relative frequency of error structures for two divergence levels.

a. Closely related sequences, overall error rate 10%

^{true}A \ ^{cw}A	0	1	2	3	>3
0	-	-	0.000	-	-
1	-	0.679	0.004	0.054	0.000
2	0.007	0.051	0.103	0.041	0.030
3	0.000	0.006	0.004	0.011	0.000
>3	-	0.000	0.000	0.000	0.008

b. Intermediate sequence divergence, overall error rate 40%

^{true}A \ ^{cw}A	0	1	2	3	>3
0	-	-	0.000	0.000	-
1	-	0.024	0.001	0.003	0
2	0.001	0.008	0.019	0.006	0.005
3	0.000	0.001	0.008	0.015	0.015
>3	-	0.002	0.006	0.018	0.865

Table 4: ***Frequency of error segments categorized by the number of indel events in the true*** *(rows)* ***and reconstructed*** *(columns)* ***alignments.***
(a) Closely related sequences; (b) Intermediate sequence divergence

For closely related sequences (table 4.a), the most frequent type of error is the simplest of all: one indel event is erroneously reconstructed, as a single indel event of the same extent but at a different position. We termed such an error a "*Shift*" error. For the

example of closely related sequences, shift errors account for ⅔ of all reconstruction errors. For comparison, in more distantly related sequences (table 4.b) the vast majority (87%) of errors result from the simultaneous mis-reconstruction of more than 3 indel events, while simple shift errors account for only 2.4% of all errors.

Figure 18 presents the relative abundance of shift errors (green), error involving only pairs of indels (blue), and complex indel misalignment errors involving three or more indels (red).

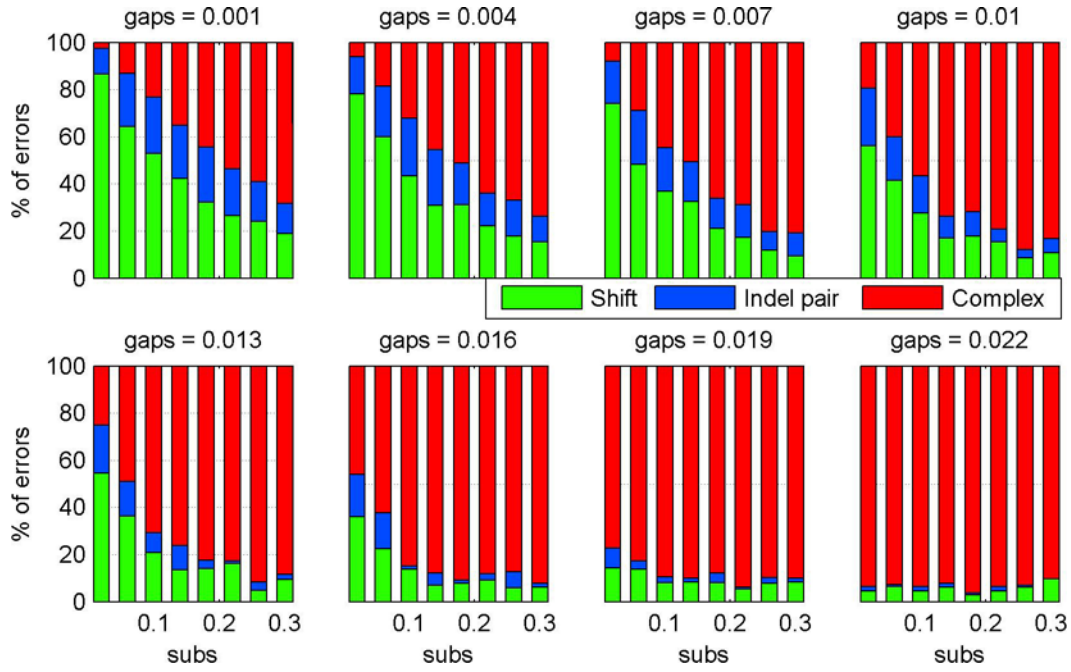


Figure 18: ***Relative frequencies of error types.***
(a) Shift errors (green), (b) indel pairs errors (blue), and (c) complex errors with more than three indel events (red). (See Figure8 for details of graphs layout.)

We note that as sequences diverge, the transition from simple errors to complex ones is much sharper than that we observed earlier for PWA errors (Figure 15). This can be understood by noting that MSA are reconstructed by a series of pairwise profile alignments, so that even if at each PWA step the errors are strictly shift errors, compounding them will produce complex errors in the MSA.

MSA errors and the guide tree

Progressive MSA reconstruction methods proceed by first estimating a phylogeny from all pairwise distances. This phylogeny is then used as a "guide-tree", which determines the sequential addition order of sequences to the growing reconstructed alignment, and the penalties for the profile pairwise alignment steps. The guide-tree, however, may be erroneous.

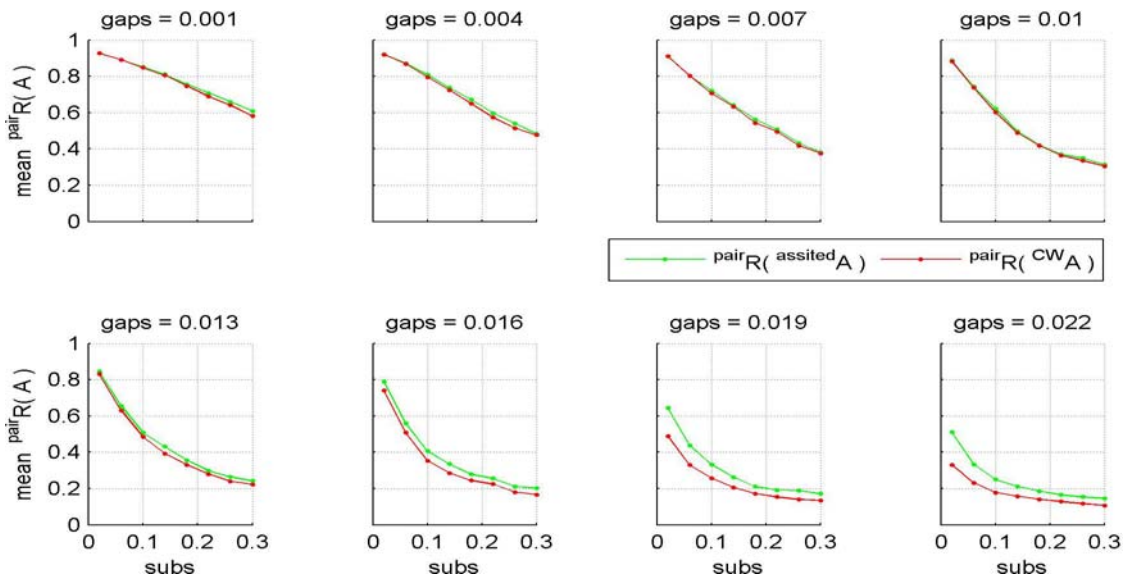


Figure 19: ***MSA reconstruction rates.***
(a) Assisted MSAs (green) utilizing the true phylogeny as the guide-tree, and (b) Standard ClustalW MSAs (red) with guide-tree derived from all pairwise alignments. (See Figure 8 for details of graphs layout.)

To assess the contribution of guide-tree inaccuracies to the MSA error rates, we consider MSAs that are guided by the true underlying phylogeny, $^{assisted}A$ (Figure 19, green). We find that the assisted MSAs are only marginally better than the ClustalW MSAs, ^{cw}A, which employ approximate guide-tree. The relative contribution of guide-tree errors to the overall MSA reconstruction error rate peaks at about 10%. Thus, inaccuracies in the reconstruction of guide-trees cannot be deemed the major source of errors in MSA reconstruction. Nevertheless, better guide-trees are always desirable.

Although the additional errors introduced by approximate guide-trees are relatively few, the effects of guide-tree errors on subsequent phylogenetic reconstruction are quite substantial. The phylogenies derived from assisted MSAs, ${}^{assisted\text{-}A}\boldsymbol{T}$, display success rates that are relatively stable even as sequence divergence increase (Figure 20, green lines).

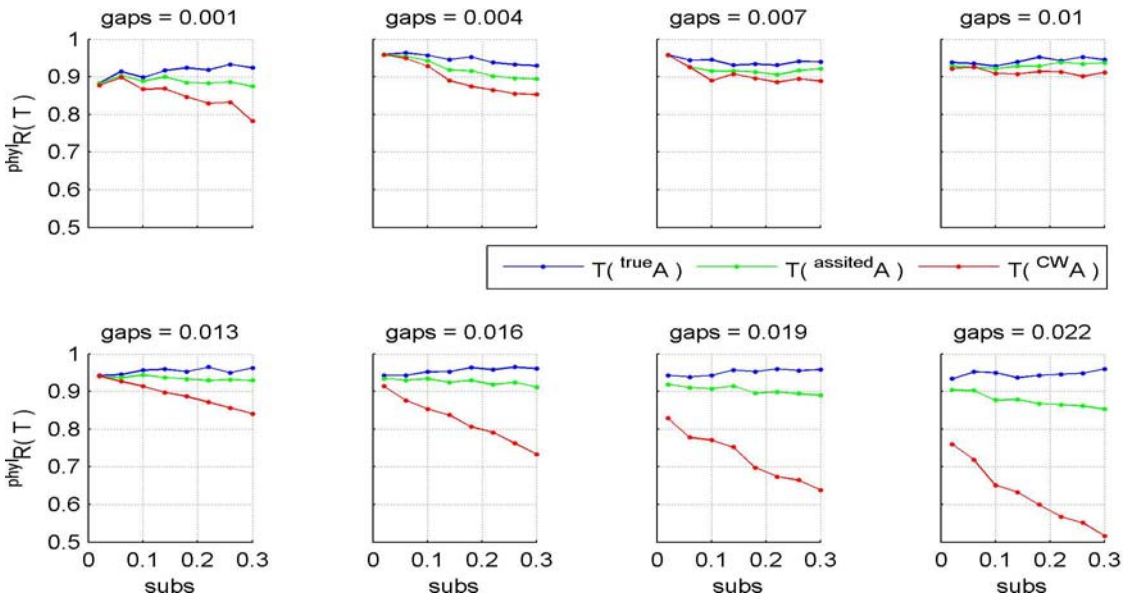


Figure 20: ***Phylogenetic reconstruction rates.***
(a) BioNJ based on true MSAs (blue), (b) BioNJ based on Assisted MSAs (green) utilizing the true phylogeny as the guide-tree, and (c) BioNJ based on standard ClustalW MSA (red) with guide-tree derived from all pairwise alignments. (See Figure 8 for details of graphs layout.)

It seems, than, that most of the increase in phylogenetic error rate which characterize phylogenies based on ClustalW MSAs, ${}^{cw}\boldsymbol{T}$ (red lines), is attributable to inaccuracies in the guide tree.

In previous sections we analyzed MSA errors mainly in the spatial dimension, that is, along the length of the alignment. We now turn to the characterization of errors in relation to the phylogenetic, or temporal dimension of MSAs. Sequence positions that have undergone changes, substitutions or indels, along an internal branch of their

phylogeny, result in informative MSA columns that reflect the partition defined by that internal branch (see Figure 2, red and blue). This phylogenetic signature will reflect the true phylogeny when changes occurred only on a single branch of the phylogeny. When multiple changes occur on different branches of the phylogeny, the partitioning apparent in a column may be misleading. In addition, several partitioning may be deduced from a single column.

We compare the phylogenetic signal of correct and error segments of reconstructed MSAs by enumerating all implied phylogenetic partitioning, and use each partition frequency over all informative columns as a support score for that particular branch. We further classify each branch as true or false in reference to both the true phylogeny and the guide tree (Table 5).

	Correct segments	Error segments	
		^{true}A	^{cw}A
True guide tree branches	0.0165	0.0443	0.0464
True branches not in guide tree	0.0015	0.0160	0.0094
False guide tree branches	0.0001	0.0032	0.0127

Table 5: ***Mean branch support per column in error and correct segments.*** *Correct segments are identical in the true and reconstructed MSAs.*

We find that error segments have higher proportions of implied partitions supporting erroneous guide tree internal branches, in expense of support of poorly supported true branches (yellow). Clearly, this bias is the result of overfitting the reconstructed MSA columns to conform to the guide tree. Also note that correct segments have lower phylogenetic signal than error segments of both the true and reconstructed MSAs (green). In other words, less variable elements of the true MSA are more easily reconstructed, but are also less informative from the phylogenetic perspective.

Sources of multiple sequence alignment errors

We summarize our understanding of the sources of errors from two perspectives, the theoretical and the specific.

From the fine details perspective, we can ascribe specific errors to several sources:

a. Positioning errors

b. Simultaneous errors

c. Biased underestimation of gaps

Positioning, or "shift", errors are the major class of errors for closely related sequences, and are in some cases the result of an arbitrary choice between co-optimal alternatives. Another common source of errors is the splitting or union of indel pairs, where the resulting gain or loss of local sequence similarity offsets, or compensate, the added or saved gap costs. For more distantly related sequences, where the anchor segments that intervene between gapped columns are less preserved, the majority of errors are the results of the simultaneous misplacement of many indel events. In most cases such errors can be classified as sub-optimal errors.

Moreover, although the objective function is assumed to balance between substitutions and indel events, there is a marked bias towards the overall minimization of gaps and a corresponding shortening of the whole alignment, thereby producing higher score alignments. A related bias results from the fact that the number of gap characters in an alignment segment depends not only on the indel length, but also on the number of

sequences sharing these gap characters. Thus, an insertion and a deletion of the same length, occurring on the same phylogeny branch, may produce very different gap content in the alignment columns. The overall bias for fewer gaps is thus translated into a bias in the reconstruction rates for insertions versus deletions.

From the underling logic of reconstruction algorithms, three aspects can be viewed as sources of reconstruction errors:

a. The guide-tree

b. Co-optimality

c. Sub-optimality

Providing the best starting point to the reconstruction algorithms, that is, the true phylogeny and the true substitution and indel rates, improves the resulting MSAs on the order of 10%. Although such improvement is desired, it does not turn poor quality alignments into good ones. On the other hand, “easy” sections of MSAs are correctly reconstructed even when using very rough approximations of the evolutionary parameters, while the “hard” parts of alignments are erroneously reconstructed even under the best external information.

Categorizing the error segments into those that are co-optimal to the true alignment and those where the corresponding true segment is sub-optimal under the optimized objective function, we find that the vast majority of MSA errors are sub-optimal ones. Only for very closely related sequences, that is, easy problems, errors are the result of the arbitrary choice among co-optimal alignments.

We conclude, therefore, that the primary culprit for the low reconstruction rates resides in the stochastic nature of the problem. Faced with situations where the realized events form a set of far from maximal likelihood, the strict optimization of an objective function leads to over-fitting. Even when the realized events approximate the expected distribution, multiple maxima of the objective function, producing co-optimal alternatives, translate into a substantial level of mis-reconstruction. Put another way, correct reconstruction can be guaranteed only when the true MSA segment is uniquely optimal, which is a relatively rare occurrence as sequences diverge.

Chapter 4: Identification and management of MSA errors

In this chapter our goal was to develop methods to deal with MSA errors. First, we present a family of local reliability measures that efficiently identify alignment errors. Next we present phylogenetic reconstruction methods that take into account MSA errors. We evaluate our methods by analysis of simulated sequences, as in the previous chapter. Application of these methods to empirical data is presented in the next chapter.

The methods we develop operate on extant OTU sequences and assume no prior knowledge of either OTU phylogeny or residue homology. Moreover, we make use of standard methods for MSA and phylogenetic-tree reconstruction.

Alternative Alignments

Our approach rests on two observations regarding reconstructed MSAs:

a. For any set of sequences there are very many biologically similar MSAs, and any single reconstructed MSA can be viewed as an arbitrary choice from among those alternatives.

b. Good quality portions of the alignments are similar among the alternative MSAs, whereas every poor quality portion is biased in its own particular way.

Thus, we shift our attention from a single reconstructed MSA to a set of alternative alignments. The simultaneous analysis of several equally likely MSAs allows us to identify the high-quality parts of MSAs, and to average out the effects of non-systematic biases in the reconstructed MSAs.

Our goal, then, is to produce a set of alternative, equally likely alignments. The set should be moderately large so as to allow for meaningful statistics, while not too large to render the analysis impractical.

Pairwise alignment – the co-optimal envelope

For the simple case of pairwise sequence alignment, we note that any reconstructed PWA, although strictly optimal, may be an arbitrary choice from among numerous co-optimal alignments (see, for example, Waterman, 1995; Gusfield, 1997). Therefore, a natural PWA set to consider is the set of all co-optimal PWAs.

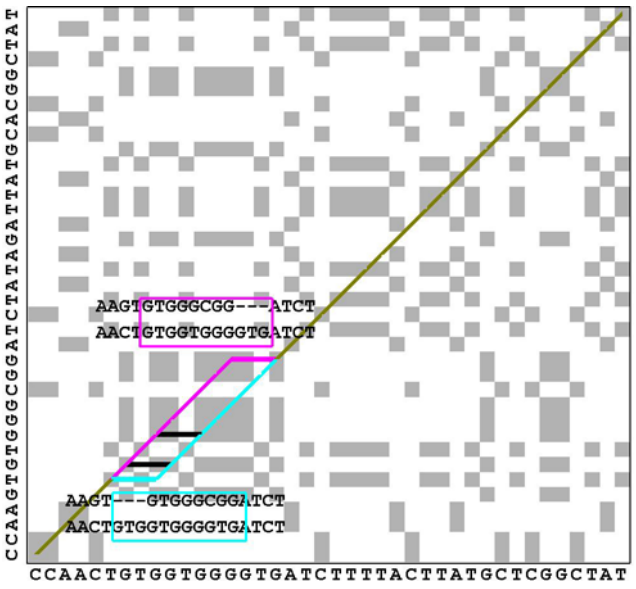


Figure 21: ***Co-Optimal PWAs as paths through a dot-matrix of the two sequences. Green segments are uniquely optimal and are common to all co-optimal alignments. Other segments are co-optimal alternatives. The co-optimality envelope is defined by the high-road (red), and low-road (blue) segments.***

For practical purposes, the full co-optimal set is far too large to enumerate explicitly (Naor and Brutlag, 1994). Yet, its main features can be summarized by considering its "envelope." Figure 21 presents an example of a co-optimal PWA set as paths through a

dot matrix view of the two sequences. Note that all co-optimal alignments share a unique path in some segments of the sequences (green), while in other regions they trace different paths (black, blue and red).

We can capture this information by considering the two extreme paths, the "high-road" alignment (red) and the "low-road" one (blue). From the consideration of figure 21, it is clear that the reliability of segments that differ between the high- and low-road alignments is at most half that of the identical segments. Some alignment programs, such as PileUp (Dolz, 1994, Womble, 2000), lets the user determine which road he likes to travel, while ALIGN and ClustalW arbitrarily chose to report the low-road alternative (Pearson and Lipman, 1988, Thompson et al. 1994). In such cases the other extreme alignment can be obtained easily by presenting the methods with the sequences in reversed residue order. Inversing the sequences amounts to reversing the direction of both axes of the dot matrix, thereby converting the high-road to low-road and vice versa (see, for example, the blue arrows in Figure 22). We term this pair of co-optimal PWAs the "Head-Tail" pair, and define it to be our basic alignment set for pairwise alignment, ***${}^{pw}AS$.***

Multiple alignment – sequence addition order

Here our goal is to produce a set of MSAs instead of the single ClustalW alignment ($^{cw}\boldsymbol{A}$) of the sequences. The alignment set should contain alignments of the sequences that are similar to $^{cw}\boldsymbol{A}$. The variation among alignments in the set should represent alternatives that are related to common sources of MSA errors.

Given an approximate $\boldsymbol{N^{otu}}$ guide-tree, we define the guide-tree alignment set, $^{gt}\boldsymbol{AS}$, as follows:

For each of the ($\boldsymbol{N^{otu}\text{-}3}$) internal branches of the guide tree, partition the sequences into two groups. Construct two sub-alignments for sequence group:

a. $^{sub}\boldsymbol{A}^{head}_{i,j}$, Which is the ClustalW alignment of the sequence group

b. $^{sub}\boldsymbol{A}^{tail}_{i,j}$, Which is the ClustalW alignment of the reversed sequences.

Where $\boldsymbol{i} \in \left[1..\boldsymbol{N}^{otu} - 3\right]$ is the branch index, and $\boldsymbol{j} \in \left[1,2\right]$ is the group index (figure 22, middle).

For each internal branch use ClustalW profile alignment to align the four combinations of the sub-alignments, aligning each combination in both the head and tail directions, to yield a total of 8 full MSAs (Figure 22):

$$\left\{ {}^{sub}\boldsymbol{A}^{head}_{i,1} , {}^{sub}\boldsymbol{A}^{tail}_{i,1} \right\} \underset{head-tail}{\mathbf{X}} \left\{ {}^{sub}\boldsymbol{A}^{head}_{i,2} , {}^{sub}\boldsymbol{A}^{tail}_{i,2} \right\}$$

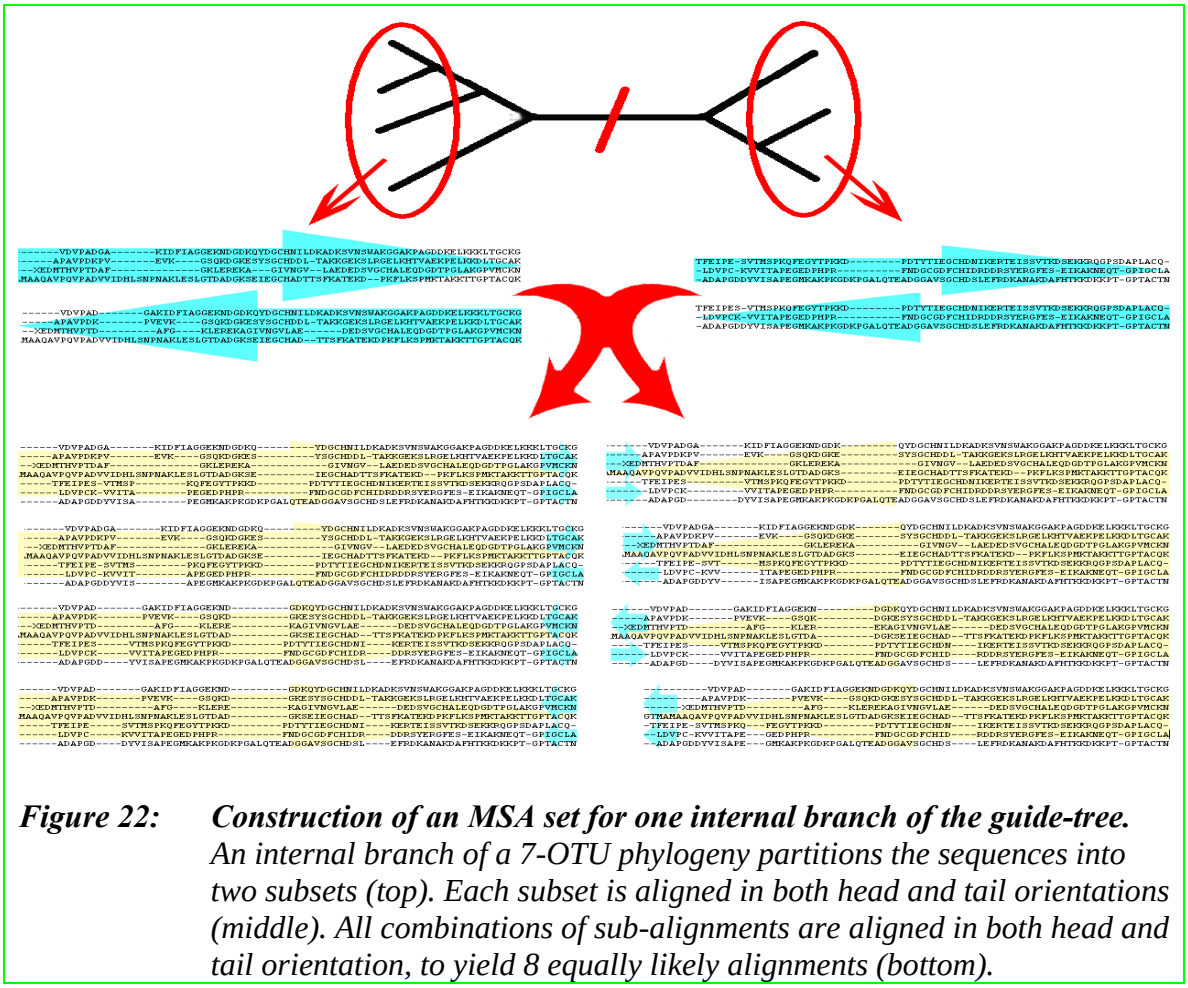


Figure 22: ***Construction of an MSA set for one internal branch of the guide-tree.*** *An internal branch of a 7-OTU phylogeny partitions the sequences into two subsets (top). Each subset is aligned in both head and tail orientations (middle). All combinations of sub-alignments are aligned in both head and tail orientation, to yield 8 equally likely alignments (bottom).*

The process is repeated for each internal branch of the guide-tree. All in all, than, ***^{gt}AS*** contains ***$8 \cdot (N^{otu} - 3)$*** alignments. These alignments differ from each other in two respects: (a) the addition order of sequences and profiles to create the final MSA, and (b) the high- or low-road selection of co-optimal sub-alignments. Any alignment in ***^{gt}AS*** could be qualified as a bona-fide progressive alignment. Thus, the alignments in ***^{gt}AS*** can be considered as equally likely alternatives.

Guide-tree alignment sets can be constructed for different choices of approximate guide trees, and combined to produce even larger sets of alternative alignments.

Local reliability measures for MSA

The local reliability of reconstructed MSAs is usually viewed as related to the local divergence of the sequences. Thus, current local reliability measures (LRMs) are based on the column entropy or variation (e.g., Thompson *et al.*, 1997). While it is true that low entropy, that is highly preserved, segments of an MSA are more easily reconstructed by MSA algorithms, column entropies are poor in identifying errors in an MSA. In this part of the study, we develop a class of LRMs that better identifies MSA errors.

Given a candidate reconstructed MSA $\boldsymbol{A}$, we construct the corresponding alignment set, ${}^{gt}\boldsymbol{AS}$, and score the elements of $\boldsymbol{A}$ by their reproduction in ${}^{gt}\boldsymbol{AS}$ (Figure 23). For each MSA column $\boldsymbol{k}$ and OTU pair $\{\boldsymbol{i,j}\}$, we define our basic local measure, the residue-pair reliability measure, ${}^{pairs}\boldsymbol{M}^{k}_{i,j}$, as the proportion of alignments in ${}^{gt}\boldsymbol{AS}$ that replicate this residue pair (see Chapter 2). The measure takes values on the interval [$\boldsymbol{0..1}$], with $\boldsymbol{1}$ for total support.

Averaging of the residue pair support gives rise to a series of reliability measures:

a. The mean residue reliability: $${}^{res}\boldsymbol{M}^{k}_{i} = \overline{{}^{pairs}\boldsymbol{M}^{k}_{i,*}}$$

b. The mean column reliability (Figure 23): ${}^{col}\boldsymbol{M}^{k} = \overline{{}^{res}\boldsymbol{M}^{k}_{*}}$

c. The overall mean alignment reliability: ${}^{ali}\boldsymbol{M} = \overline{{}^{col}\boldsymbol{M}^{*}}$

Figure 23: ***Column reliability measure*** *(red line)* ***for a candidate MSA.*** *Columns of a candidate MSA (bottom) are scored by their reproduction in the MSA set (top)*

The LRMs are intended to identify errors in reconstructed MSAs. We therefore test their performances by comparing them to the known error structure of reconstructed MSAs, in simulation settings as in chapter 3.

One use of our reliability measures is as binary classifiers of local MSA features as correct or erroneous. Figure 24 presents a receiver-operating characteristic (ROC) analysis (Zweig and Campbell, 1993) of $^{pairs}\boldsymbol{M}$ as a classifier of residue-pairs errors. Since the residue-pairs reconstruction, $^{pairs}\boldsymbol{R}$, is binary, the two populations – error (***H0***, red) or correct (***H1***, green) reconstructions - are strictly defined. Our measure $^{pairs}\boldsymbol{M}$ is seen admiringly to separate the two populations, with very high power (area under curve, AUC=0.95).

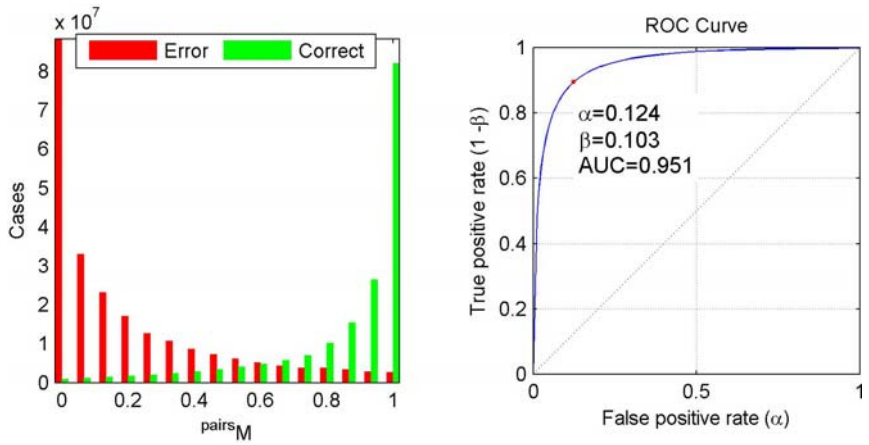
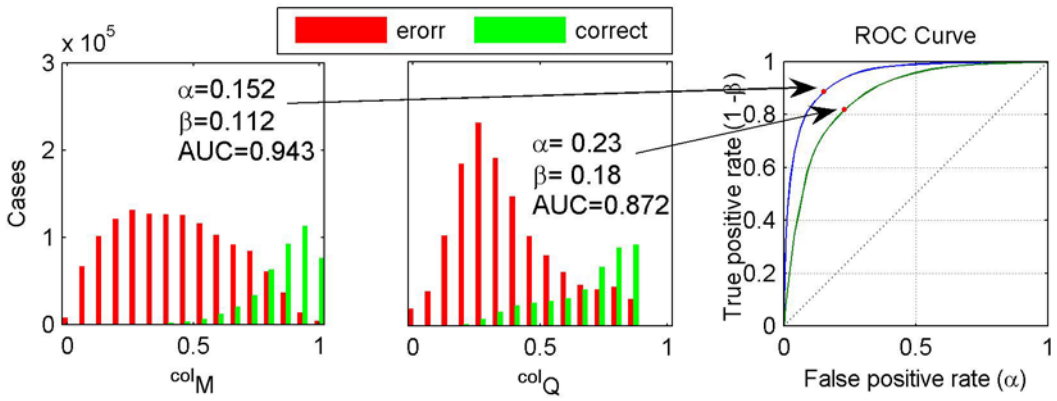


Figure 24: ***The residue-pairs reliability measure as a classifier of reconstructed MSA errors/correct features.***
Histograms (left) presents the different distributions of the two populations: ***H0****:error(red) vs.* ***H1****:correct(green). ROC curves (right) report the level of classification errors and the power of the classifier.*

The most useful level of MSA scoring is the column level. Current methods employ Shannon's entropy as a measure of MSA quality, that is, column quality is judged by its residue variability. The entropy-based column quality measure reported by ClustalX, $^{col}\boldsymbol{Q}$ (Thompson *et al.*, 1997), is inferior to our $^{col}\boldsymbol{M}$ local reliability measure. A ROC analysis (Figure 25.a), reveals that $^{col}\boldsymbol{M}$ separates the two populations better than $^{col}\boldsymbol{Q}$, with AUCs of ~0.94 and ~0.87, respectively.

a. ROC analysis of the two measures



b. Correlation of the two measures to the true error rate

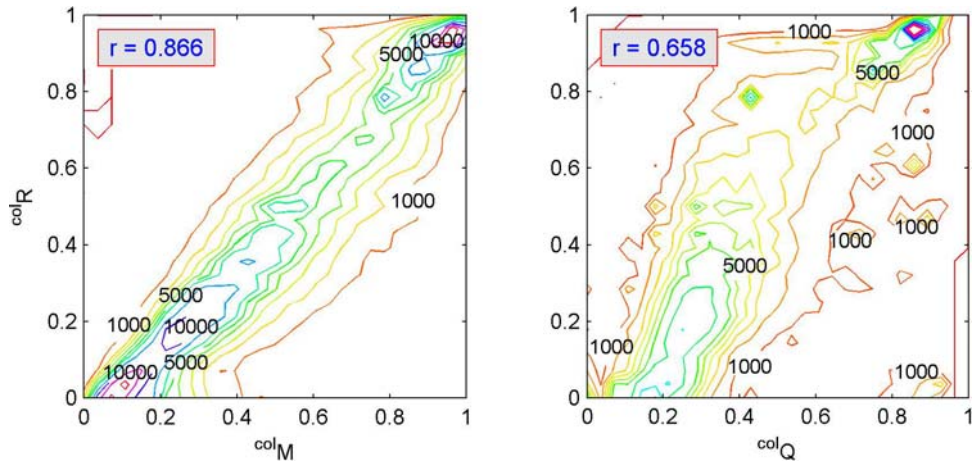


Figure 25: ***Comparison of two column quality measures,*** ^{col}M ***and*** ^{col}Q. *(a) as classifiers of MSA reconstruction errors, and (b) as estimates of error rates.*

When interpreting the LRMs $^{*}\boldsymbol{M}$ as estimates of the reconstruction rates $^{*}\boldsymbol{R}$, we find extremely high correlations between the two types of measures, one derived from the comparison to the true MSA, $^{*}\boldsymbol{R}$, and the other from the MSA set, $^{*}\boldsymbol{M}$. The correlation coefficients are $\boldsymbol{r} = 0.94$ for the residue-base measure and $\boldsymbol{r} = 0.87$ for the column measure. Once again, the entropy-based column quality measure is inferior to our $^{col}\boldsymbol{M}$, the correlation between $^{col}\boldsymbol{Q}$ and $^{col}\boldsymbol{R}$, though significant, being only $\boldsymbol{r} = 0.66$ (Figure 25.b).

Application of LRMs in phylogenetic reconstruction

One use of LRMs is to account for MSA errors while reconstructing a phylogeny based on an MSA. This can be achieved by weighting or filtering MSA columns and residues by their LRM when estimating pairwise distances. Hopefully, distance matrices that are less affected by erroneous segments of the MSA will be more accurate, and will therefore produce better phylogenies. This hope is largely groundless.

We find that filtering of MSAs by the removal of errors results in a deterioration of reconstruction rates (data not shown), which may be explained by the reduction of the sample size for distance estimation.

When weighting, rather than filtering, MSA columns or residue-pairs by their LRMs, we find (Figure 26) that although the resulting phylogenies may differ from the unweighted ones, there is no significant improvement in phylogenetic reconstruction rates, although the mean values are marginally better when weighting (red lines).

This result may be traced back to the observation that MSA errors tend to reduce the support for poorly resolved internal branches. Thus, poorly supported internal branches are not sufficiently represented in the MSA to begin with, and weighting down of erroneous internal branches cannot sufficiently enhance their signal.

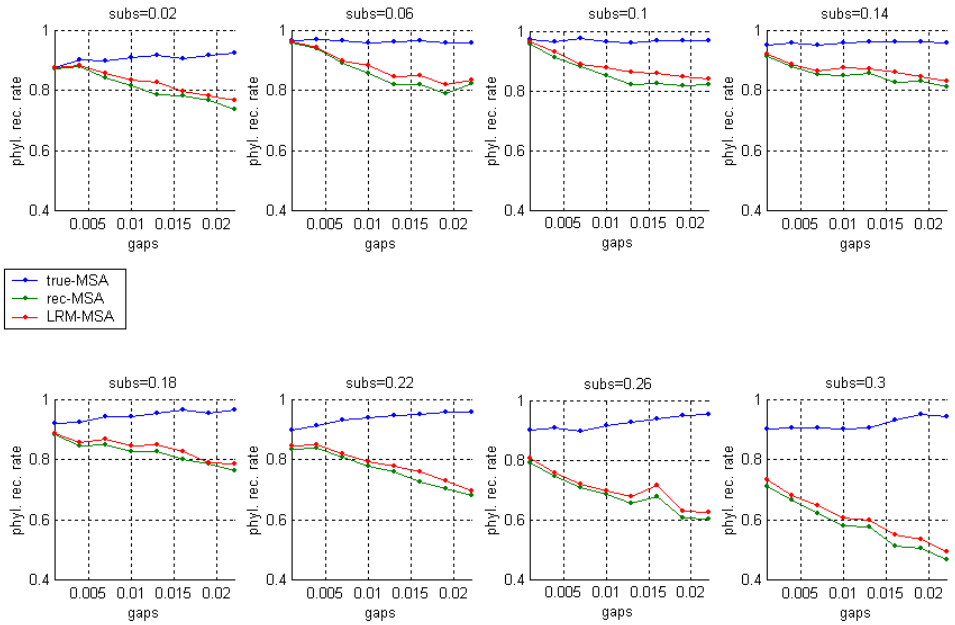


Figure 26: ***Phylogenetic reconstruction rates for LRM weighting.*** *(a) Distances derived from true MSAs (blue), (b) distances derived from reconstructed MSAs (green), and (c) distances derived from reconstructed MSAs weighted by our residue-pair reliability measure (red).*

We reach the unfortunate conclusion that while LRMs can detect errors in a reconstructed MSA, their utility in improving phylogenetic reconstruction is small.

Phylogenetic analysis of alignment sets

The alignment set ${}^{gt}\boldsymbol{AS}$, contains much more information than is captured by our LRMs. Another approach to utilize this information is to infer phylogenies directly from ${}^{gt}\boldsymbol{AS}$ as a whole.

For each alignment in an alignment set, we reconstruct a phylogeny using some standard phylogenetic reconstruction method. In this study we used BioNJ (Gascuel, 1997) as the tree reconstruction method. The resulting phylogeny set, ${}^{gt}\boldsymbol{TS}$, is than used to infer a consensus phylogeny for the alignment set, ${}^{as}\boldsymbol{T}$ (figure 27).

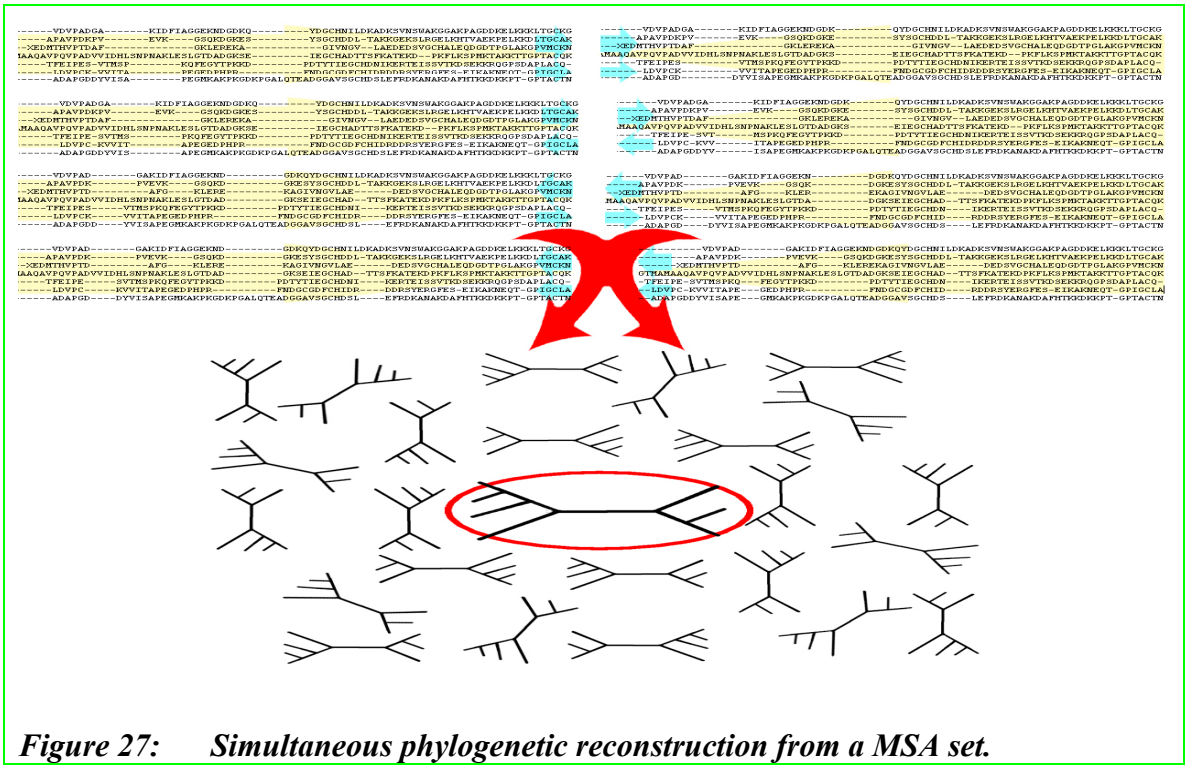


Figure 27: ***Simultaneous phylogenetic reconstruction from a MSA set.***

To avoid situations where the standard majority-rule consensus method (e.g., Felsenstein, 2004) yields partially resolved trees, we adopted a variant consensus method where the inferred phylogeny is the best supported tree from among the sets trees.

The phylogenetic reconstruction rate of the alignment-set consensus tree $^{as}\boldsymbol{T}$ (Figure 28, green) is significantly higher than that of phylogenies derived from a single ClustalW MSAs, $^{cw}\boldsymbol{T}$ (red). The overall mean improvement is ~6% (Wilcoxson signed-rank test p-value < 10^{-16}). Interestingly, for closely related sequences, the $^{as}\boldsymbol{T}$ phylogenies may be more accurate than phylogenies derived from the true MSA (blue).

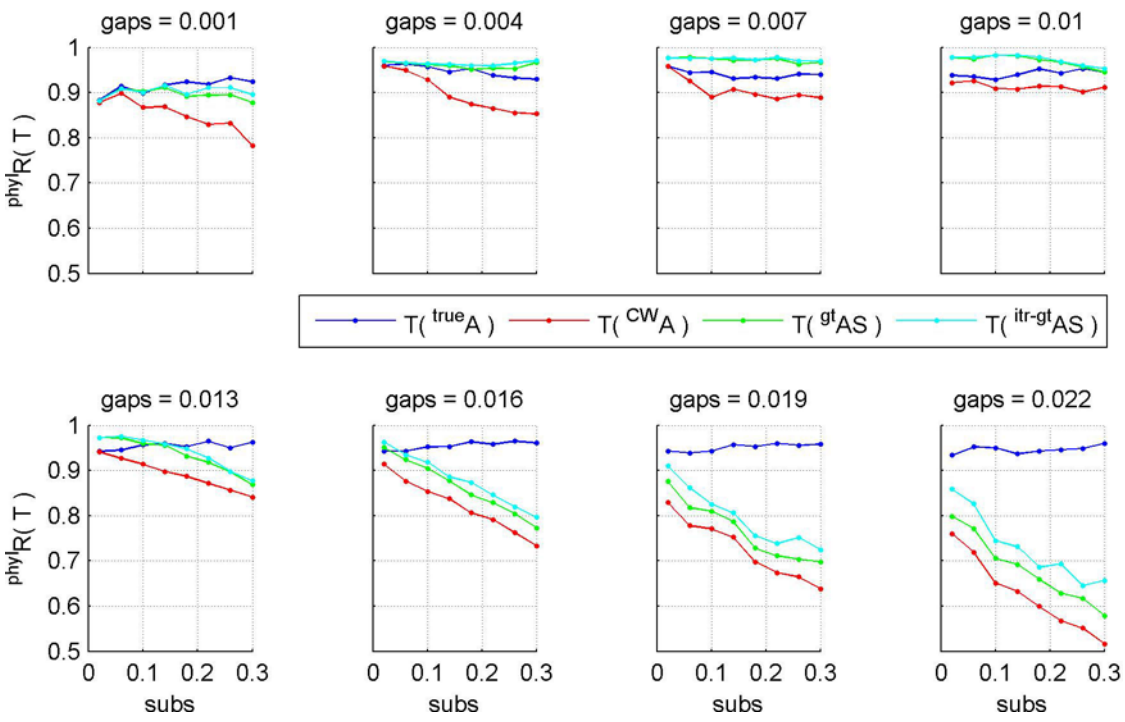


Figure 28: ***Phylogenetic reconstruction rates.***
(a) BioNJ based on true MSAs (blue), (b) BioNJ based on standard ClustalW MSA (red), (c) Consensus phylogeny based on MSA set (green), and (d) Consensus phylogeny from iterative construction of MSA sets (cyan.)

We note that the construction of an alignment set, a phylogeny set, and the resulting consensus tree ${}^{as}\boldsymbol{T}$, is dependent upon the initial choice of guide-tree. We therefore repeat the process, using as a guide-tree the ${}^{as}\boldsymbol{T}$ of the previous iteration, which is our best estimate of the phylogeny so far. The analysis is iterated until the guide-tree and the inferred phylogeny converge, or until a pre-specified number of iterations is reached. We term the final tree the "iterative alignment set phylogeny", ${}^{itr\text{-}as}\boldsymbol{T}$ (cyan). In practice, nearly all cases converge within 6 iterations of the alignment set analysis. For closely related sequences, iteration does not improve upon the basic ${}^{as}\boldsymbol{T}$. For more distantly related sequences, the improvement of ${}^{itr\text{-}as}\boldsymbol{T}$ over ${}^{as}\boldsymbol{T}$, is yet again as large as the improvement of ${}^{as}\boldsymbol{T}$ over ${}^{cw}\boldsymbol{T}$.

Chapter 5: Application of methods to case studies

In this part of the study we apply the methods of the previous chapter to the analysis of alignment and phylogenetic reconstruction problems of real biological sequences.

Data

The case studies we analyze are taken from the BaliBase database (Bahr *et al.*, 2001). BaliBase is a database for benchmarking MSA programs, which have been developed by the authors of the ClustalW algorithm, and is widely used for the comparison of MSA algorithms (e.g., Karplus and Hu, 2001, Lassmann and Sonnhammer, 2002, Wallace *et al.*, 2005). The MSAs in BaliBase are curated, and for each there is a definition of core segments within the alignment which should/must be reconstructed by programs. These are basically the highly conserved domains of the proteins, while the non-core segments are the more variable, especially in gaps, and are considered as so ambiguous that any alignment over those segments is admissible. BaliBase contains several datasets, each presenting the MSA programs with different sorts of reconstruction difficulties.(see http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE2/)

We use the core segments of BaliBase MSAs to reconstruct reference phylogenies. These phylogenies are based on high quality alignments, and we regarded them as the best reference phylogenies for the sequences. The variable segments of BaliBase MSAs, on the other hand, represent cases where the sequence alignment is highly ambiguous, and therefore amenable to MSA error management methods.

The case study consists of 320 MSAs derived from BaliBase (see Chapter 2). Figure 29 presents the distribution of the MSAs sizes and sequence divergence.

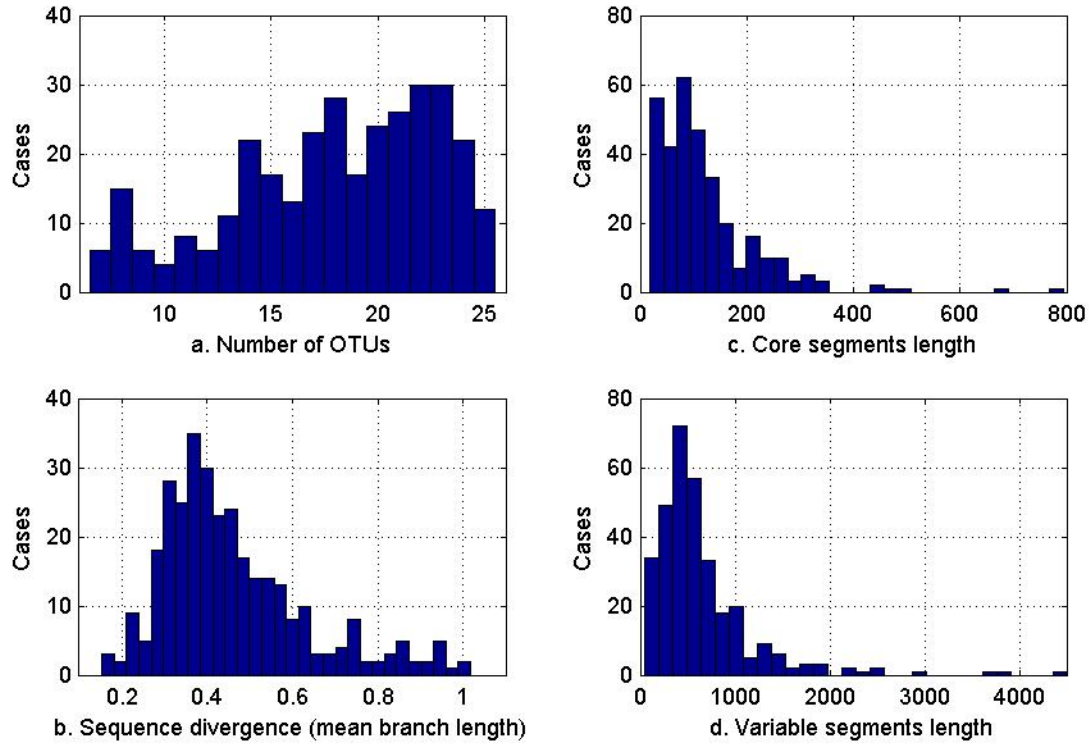


Figure 29: ***Distributions of the 320 BaliBase test cases.***
(a) Number of OTUs; (b) Sequence divergence; (c) Core segments length; (d) Variable Segments length;

The variable regions of BaliBase MSAs were first analyzed using standard ClustalW alignment, followed by a BioNJ phylogenetic reconstruction. Comparing those ClustalW-BioNJ phylogenies, $^{cw}\boldsymbol{T}$, to the reference phylogenies derived from the core segments, $^{ref}\boldsymbol{T}$, we find that the mean phylogenetic reconstruction rate is 38% (Figure 30).

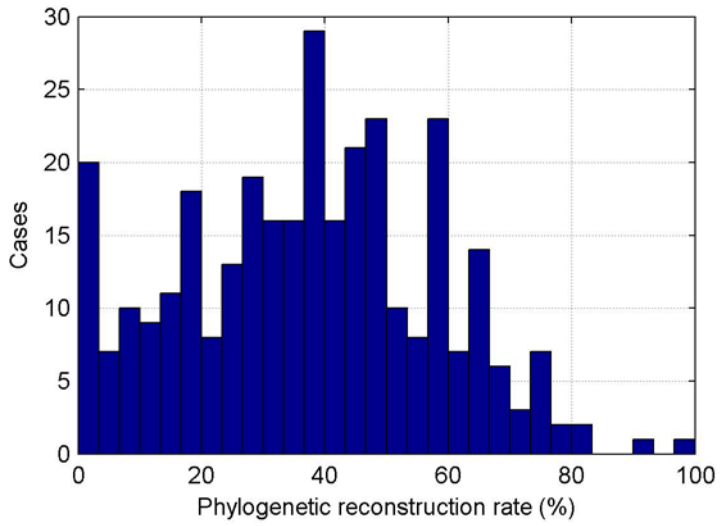


Figure 30: ***Phylogenetic reconstruction rate of BaliBase test cases.*** *Phylogenies reconstructed by BioNJ from ClustalW MSAs.*

The ClustalW MSAs of the variable segments were further scored with our proposed reliability measures. Weighting of pairwise sequence distances by the reliability measures ${}^{pairs}\boldsymbol{M}$ and ${}^{col}\boldsymbol{M}$ did not produce significant improvement of the phylogenetic reconstruction rates (data not shown).

Phylogenetic reconstruction using alignment sets

For each BaliBase MSA, we reconstruct three phylogenies:

a. ${}^{ref}\boldsymbol{T}$: Reference tree, is a BioNJ based on the core blocks of the BaliBase alignments. Parts of ${}^{ref}\boldsymbol{T}$ may be poorly resolved, and these are identified by a bootstrap analysis (Felsenstein, 1985).

b. ${}^{cw}\boldsymbol{T}$: ClustalW tree, BioNJ tree based on a standard ClustalW alignments of the variable regions.

c. ${}^{itr\text{-}as}\boldsymbol{T}$: Iterative alignment-set tree, our proposed method, derived from the variable regions alone.

The ClustalW and alignment-set trees are than compared to the reference tree. We used the symmetric tree distance (e.g., Felsenstein, 2004), normalized by the number of branches, to produce the reconstruction rate ${}^{phy}\boldsymbol{R}$ (see chapter 2).

We find that the mean improvement of ${}^{itr\text{-}as}\boldsymbol{R} = {}^{phy}\boldsymbol{R}({}^{itr\text{-}as}\boldsymbol{T})$ over ${}^{cw}\boldsymbol{R} = {}^{phy}\boldsymbol{R}({}^{cw}\boldsymbol{T})$ is 4.4%, which is significant at the 10^{-11} level (Figure 31). Relative to ${}^{cw}\boldsymbol{R}$, the improvement is about 12%.

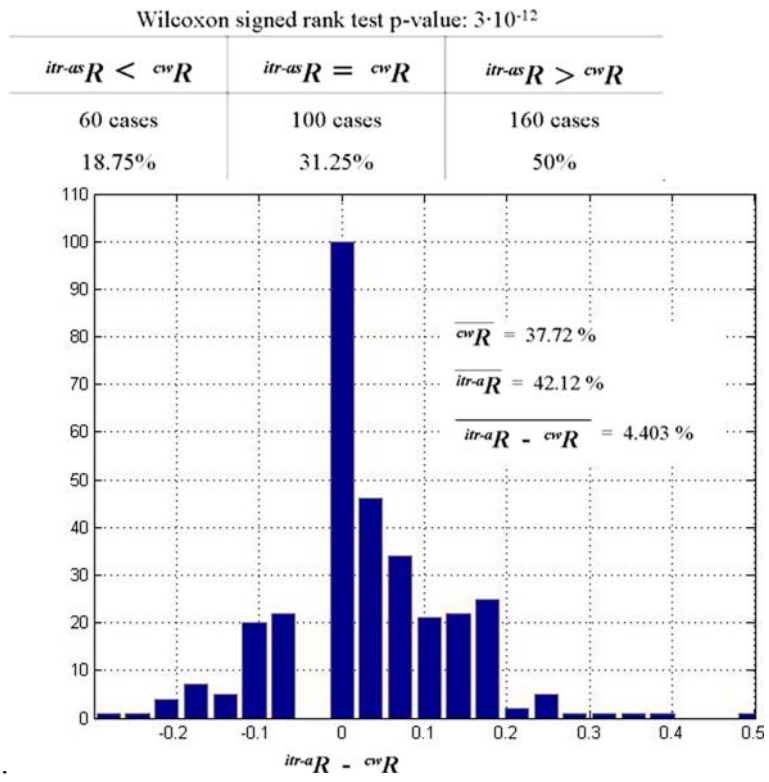


Figure 31: Improvement in phylogenetic reconstruction rate.

We take this opportunity to provide a brief comparison of ClustalW to two other MSA reconstruction methods: PileUp (Dolz, 1994, Womble, 2000) and MUSCLE (Edgar, 2004). For BioNJ phylogenies based on MUSCLE MSAs of the test cases, the mean reconstruction rate is 38%, which is comparable to the phylogenetic reconstruction rate of ${}^{cw}\boldsymbol{T}$. Phylogenies derived from PileUp MSAs are less accurate, with mean phylogenetic reconstruction rate of 32%.

In the above, we take ${}^{ref}\boldsymbol{T}$, which is based on the core segments, to be the best estimate we can have for the underlying phylogeny. However, ${}^{ref}\boldsymbol{T}$ may contain errors, which may lead to a lowering of ${}^{itr\text{-}as}\boldsymbol{T}$ and ${}^{cw}\boldsymbol{T}$ scores. We therefore repeated the analysis using only the highly supported internal branches of ${}^{ref}\boldsymbol{T}$, that is, branches with bootstrap score of more than 50%. We find that the improvement gained by our iterative alignment set method is indeed larger, with average over the cases of 6.2% (Figure 32).

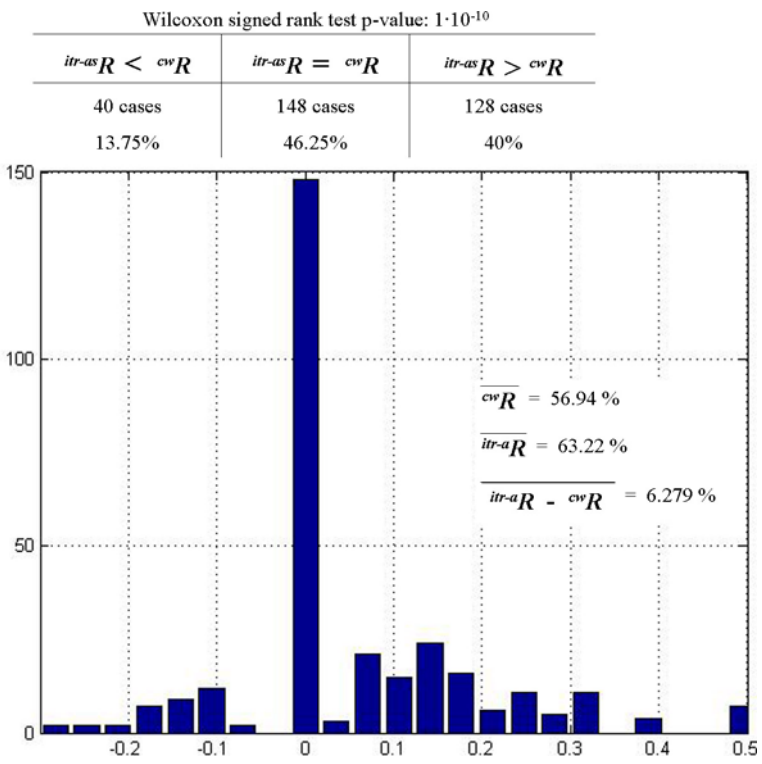


Figure 32: ***Improvement in phylogenetic reconstruction rate for branches of the reference phylogeny supported by bootstrap proportion > 50%.***

Our proposed method uses the consensus tree of the iterative phylogeny set ${}^{gt}\boldsymbol{TS}$ as the final reconstructed phylogeny, ${}^{itr\text{-}as}\boldsymbol{T}$. More elaborate methods for choosing the best tree from ${}^{gt}\boldsymbol{TS}$ will certainly improve the performance of our method. The tree choice problem is deferred to another study, but its probable performances can be assessed by

assuming we know which is the best phylogeny in ${}^{gt}\boldsymbol{TS}$ (Figure 33). In this case, the mean improvement is 16.3%, which is 43% relative to ${}^{cw}\boldsymbol{T}$, with improvement in 90% of the cases.

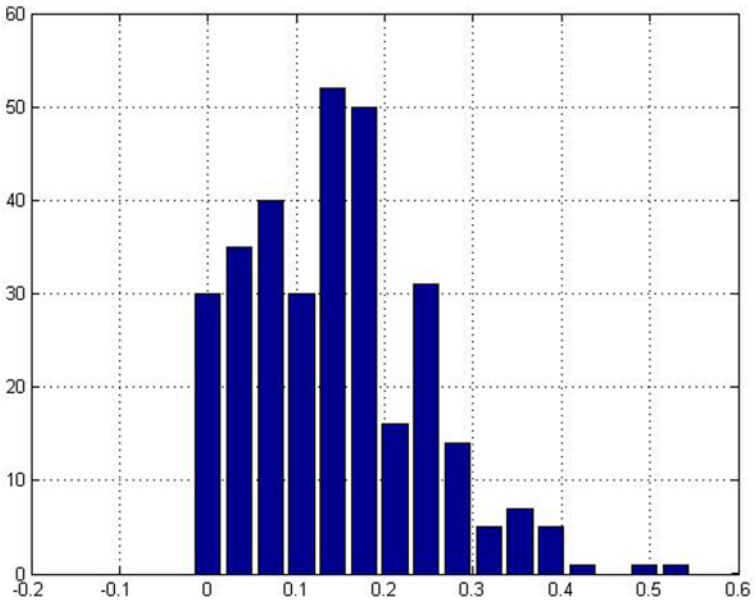


Figure 33: ***Phylogenetic reconstruction improvement for the best tree in the phylogeny set.***

Chapter 6: DISCUSSION

Although there are very many MSA reconstruction programs available, we have opted to characterize the errors of only one such method, ClustalW (Thompson *et al.*, 1994a.) ClustalW is by far the most widely used MSA reconstruction program. Studies comparing the performances of competing MSA reconstruction methods always take ClustalW as their "gold standard," and usually report only marginal differences between the methods compared. This lack of difference is expected since most methods incorporate two common ingredients: progressive alignment along a guide-tree and an affine gap-cost objective function. In our case studies (Chapter 5), we have included a comparison to two other MSA reconstruction programs, PileUp (Dolz, 1994, Womble, 2000) and MUSCLE (Edgar, 2004), and found that the accuracy of phylogenetic reconstruction based on MUSCLE-MSAs is comparable to that of ClustalW, and that PileUp-MSAs produces less accurate phylogenies.

In the earlier parts of this study we have used simulations of sequence evolution to provide us with true MSAs and true phylogenies against which to compare reconstructed MSAs and phylogenies. A standard criticism of simulation studies is that they may not be relevant to real-life circumstances. We have taken great care to render our simulations as biologically realistic as possible, by restricting their range so that descriptive statistics of the simulated MSAs match the descriptive statistics of biological MSAs that have been deposited in alignment databases.

Our simulation process was kept simple, with substitutions and indels as the only types of sequence change, and with equal rates along independent sequence residues. These

settings replicate the assumptions inherent in MSA reconstruction methods. In this sense, the MSA reconstruction process was tested in a best-case evolutionary scenario. This allows us to focus on the most basic errors that are characteristic of the reconstruction process, without obfuscating the analysis with errors resulting from more complex sequence evolution phenomena. It is, therefore, expected that the reconstruction rates we have reported represent an upper limit of the performance of MSA reconstruction, and that MSAs of real biological sequences will typically have even higher error rates.

MSA reconstruction errors and their effects

The primary conclusion from the comparison of reconstructed alignments to native alignments from simulations is that reconstructed alignments are highly uncertain in their details. Only very closely related sequences can produce accurate alignments, while many sequence sets of biological interest are expected to produce reconstructed alignments with error in more than half of their columns.

Errors in reconstructed MSAs are expected to affect adversely subsequent analyses that use MSAs as their input. For the case of phylogenetic reconstruction in our simulation setting, we showed that phylogenies derived from reconstructed MSAs are much less accurate than those derived from true MSAs (Figure 6). In fact, even a relatively simple phylogenetic reconstruction method such as BioNJ (Gascuel, 1997) is robust when based on the true MSA. Thus, the low phylogenetic accuracy in real-life settings can be almost wholly attributed to the poor quality of reconstructed MSAs. In actual sequence

analysis problems, the true MSA is never known, and we may only hope to be able to identify and correct the errors.

The immediate source of MSA reconstruction errors is in the erroneous deduction and positioning of gaps. In other words, more errors occur in gapped columns than in anchor columns. For closely related sequences, in which the error rate is low, most reconstruction errors can be classified as simple shift errors. These errors preserve the alignment length, and their effect is usually local. As sequences diverge and indels accumulate, errors resulting from the simultaneous rearrangement of many indel events become more and more prominent. Such complex errors affect larger and larger portions of the reconstructed MSA, so that even for intermediate levels of sequence divergence, most of the length of the MSA may be erroneously reconstructed.

In such cases, it is generally the rule that the erroneous MSA is shorter in length and contains fewer gaps than the true MSA. In addition, there is a bias in the ability to correctly reconstruct insertions and deletions. Deletions in a few OTUs or insertions in many OTUs are better dealt with by the MSA reconstruction program than insertions in a few OTUs and deletions in many OTUs. In both cases, this reflects an algorithmic bias towards the minimization of the number and size of gaps.

These biases are the result of applying optimization techniques to highly variable stochastic processes. In sequence evolution, the likelihood of actually realized random events is often far below the maximum likelihood of the true stochastic parameters, leading to over-fitting of the MSA structure to the evolutionary parameters. This is demonstrated by the observation that in most cases where the reconstructed alignment differs from the true one, the objective function score of the true historical alignment is

lower than the optimum, that is, the true MSA is sub-optimal. Moreover, even when the true alignment attains the optimum score, correct reconstruction is not guaranteed. Alternative co-optimal alignments are very frequent, and the choice among them is arbitrary.

Progressive MSA reconstruction utilizes an approximate phylogeny, or guide-tree, to determine the addition order of sequences to the partially reconstructed MSA, and to provide the objective functions for the scoring of the successive pairwise alignment steps. It is natural to expect that the quality of the guide-tree will critically affect the quality of the resulting MSA. Contrary to this expectation, we find that providing the true phylogeny as the guide-tree improves the resulting MSA only marginally (Figure 19). A possible explanation of this finding is that the expectation is valid only for those segments of an MSA where the true MSA is uniquely optimal under the correct evolutionary parameters. In cases in which there are other co-optimal possible MSAs in addition to the true MSA, or when the true MSA is sub-optimal, reconstructions errors are bound to occur even under perfect knowledge of the phylogeny and the evolutionary rates.

Phylogenies and MSAs

Sequence phylogenies and multiple sequence alignments are two descriptions of a single underlying evolutionary history, and should always be treated as dual aspects of the same phenomenon. As such, they also present a typical case of circular reasoning: approximate phylogenies govern the progressive reconstruction of MSAs, while the resulting MSAs are used to reconstruct phylogenies. It is not surprising, then, that

phylogenetic reconstruction rates are affected more by the quality of the initial guide tree than by actual quality of the MSAs (Figures 19 and 20). This is due to the fact that although the approximate nature of the guide-tree does not drastically affect the frequency of errors in reconstructed MSAs, it does introduce a substantial bias in the phylogenetic signal that becomes apparent in the erroneous columns of the reconstructed MSA.

The phylogenetic signal of reconstructed MSA columns was found to be biased in towards the topology of the guide tree. This frequently tends to lend spurious support to erroneous inner branches of the guide tree, while disrupting phylogenetic signal in support of poorly resolved true inner branches. Of course, such spurious heightening of the phylogenetic signal is benign only when the guide tree is actually the true phylogeny. When the guide tree is only approximate, i.e., some true internal branches are missing from it and are replaced by erroneous internal branches, overfitting to the erroneous internal branches is accompanied by a reduction in the strength of the phylogenetic signal supporting the absent true internal branches. The overall result, therefore, is an MSA with spuriously heightened support of both true and erroneous internal branches of the guide-tree. Clearly, this is a case of circular reasoning, where the quality of our prior expectation determines the accuracy of our final conclusions. In this respect, phylogenetic reconstruction is extreme among MSA-dependent analyses, since the information provided to the reconstruction process is of the same class as the information deduced from the reconstructed MSA, thus, creating a vicious cycle.

Such considerations led some authors (e.g., Thorne and Kishino, 1992, Vinga and Almeida, 2003) to abandon altogether the use of MSAs in phylogenetic reconstruction.

(We note that although such an approach may be acceptable in phylogenetic reconstruction, it may not be applicable for other types MSA-dependent analyses.) An alternative approach to circular reasoning is to use it in a Bayesian fashion, with posterior refinement of approximate priors. In our proposed method for phylogenetic reconstruction based on MSA sets, such an iterative approach proved to be of practical value in improving the accuracy of reconstructed phylogenies. Needless to say, better guide-trees are always welcome.

The quality of the guide-tree is mainly determined by the accuracy of the pairwise distance-matrix derived from pairwise alignments. The estimated distances, in turn, gain accuracy with increasing sample size (i.e., sequence lengths). Thus, MSAs of long sequences start off with better guide trees and their error rate is lower than MSAs of short sequences. This is in contrast to the situation in pairwise alignment, where error levels are almost unaffected by sequence lengths.

Our overall conclusion is that only very closely related, long sequences, with few indels to be reconstructed, and long between-gap anchors, are amenable to meaningful alignment reconstruction. However, in the real world, homologous sequences are frequently short and characterized by a high gap content. The result is that even for moderate distances, reconstructed alignments are expected to be correct for only about half of their total length. This situation clearly requires methods for the identification and management of MSA errors.

The proposed methodology

Dealing with alignment errors is predicated upon our ability to identify them and reduce their effects in subsequent analyses. To these ends, it is profitable to examine sets of alternative, equally likely, alignments. The alignment set should be sufficiently variable to support robust statistics, while at the same time small enough so as to keep the amount of processing needed to a practical level. Clearly, not any arbitrary choice of alignments will qualify as equally likely biologically.

We presented one such alignment set, the guide-tree alignment set (^{gt}AS), which contains $8\cdot(N^{otu}-3)$ MSAs. The alignments in ^{gt}AS share the same guide tree, but differ in the addition order in which the progressive process proceeds, and the arbitrary choice from among co-optimal alternatives. Since even the construction of the guide tree requires $O(N_{otu}^{2})$ alignment steps, the additional $O(N_{otu})$ steps of our method are negligible in terms of processing time, with at most a doubling of CPU time for the worst case of 4 OTUs. Although the utility of this alignment set is demonstrable, we find it to be very conservative. It may be worthwhile, than, to develop equally likely alignment sets that span larger portions of the MSA space.

One use of alignment sets is to score some specific candidate MSA. We presented a series of local reliability measures that score elements of a candidate MSA by the frequency in which they are reproduces in the set's alignments. The local reliability measures we developed proved to be very good predictors of MSA errors. Unfortunately, we found that identification of MSA errors is not sufficient to improve phylogenetic accuracy when analyzing a single MSA. Yet, our family of quality measures may be of use in other types of alignment-dependent analyses.

Filtering of MSA errors by local reliability measures is similar to the current practice of ignoring gapped columns of the MSA when reconstructing phylogenies. This practice seems to be justified by the fact that errors occurs more frequently in gapped MSA columns. However, the errors also disrupt the local structure of neighboring anchor columns, resulting in erroneously reconstructed anchor columns.

Both types of filtering, either by gapped columns or by our local reliability measures, suffer from two drawbacks. First, filtering reduces the sample size, in many cases drastically, thus increasing the variance of estimated distances. In addition, indels occur more frequently in more variable domains of the sequences. Usually these are also the most informative domains from a phylogenetic standpoint, and contribute the most to the divergence signal when estimating mean sequence distances. Thus, removal of those regions results in underestimation of pairwise distances, and a systematic bias in the resulting distance matrix and the phylogeny derived from it.

Even without filtering of variable columns, distance matrices derived from reconstructed alignments are systematically biased towards underestimation of divergence rates. This results from the overfitting of the reconstructed MSA to maximize the objective function. An issue for further study is whether distances can be corrected for this systematic bias. Such a correction should transform observed differences to distance estimates, taking into account not only the phenomenon of multiple substitutions, but also the local statistics of MSA biases of specific reconstruction methods.

Local reliability measures average out, and extremely reduce, the amount of information available in the full alignment set. Moreover, their context is still a single reconstructed

MSA. Unfortunately, even with perfect knowledge of the location of the errors, it is not possible to transform a poor quality MSA into high quality one.

We propose, than, that the prudent approach is never to use a single reconstructed MSA as the basis for further analyses. Rather, MSA-dependent methods should be enhanced and adapted to accommodate the simultaneous analysis of MSA sets.

In the context of phylogenetic reconstruction, we applied a simple consensus method to derive phylogenies from alignment sets, and found that the resulting phylogenies are significantly more accurate than those based on a single MSA. We note that our proposed MSA set, ${}^{gt}\boldsymbol{AS}$, is dependent upon an approximate guide-tree. Application of our method in an iterative fashion, using the deduced phylogeny as a guide-tree for the next iteration, enhances the phylogenetic reconstruction rate even further. Interestingly, when the sequences are relatively closely related, the phylogenetic reconstruction rate may be even higher than that attained when using the true error-free MSA (Figure 28).

We demonstrated the utility of the proposed phylogenetic reconstruction method in the analysis of real biological sequences from the BaliBase database (Bahr *et al.*, 2001). We found that our method is significantly more accurate then the standard single-MSA analysis, with a mean improvement of about 5% in phylogenetic reconstruction rates.

Our consensus method is very simple and does not always retrieve the best phylogeny from the alignment set. Therefore, we find it probable that refinement of the selection method from among the phylogenies in the phylogeny set may further enhance the phylogenetic accuracy. Such refinement may draw on the phylogenetic signal of MSA columns on the one hand and on Bayesian analysis on the other.

MSAs are ubiquitous tools in molecular biology, and in a manner similar to buffers, they are taken for granted. Moreover, most MSAs in actual use are produced and discarded automatically on the road to some other goal. I conjecture that more than 99% of MSAs that are used to produce publishable results, are never even seen by a human being. (This is certainly the case for this study.) Yet, when a rare MSA is actually inspected by a researcher, it is usually found wanting. MSAs are so notoriously inadequate, that the literature is littered with phrases such as "The MSA was subsequently corrected by visual inspection." In fact, Thompson *et al.* (1994a) in their seminal paper clearly state: "CLUSTAL W is… a very useful starting point for manual refinement…"

I would like to augment my conclusions with the following advice (with apologies to Antoine Saint Exupéry):

> The danger of the MSA is so little understood, and such
> considerable risks would be run by anyone who might get lost in
> a phylogeny, that for once I am breaking through my reserve.
> "Children," I say plainly, "watch out for the MSAs!"

Literature cited

Allison, L., C.S. Wallace, and C.N. Yee. 1992. Finite-state models in the alignment of macromolecules. J Mol Evol **35:** 77-89.

Althaus, E., A. Caprara, H.P. Lenhof, and K. Reinert. 2002. Multiple sequence alignment with arbitrary gap costs: Computing an optimal solution using polyhedral combinatorics. Bioinformatics **18:** S4-S16.

Altschul, S.F. 1993. A protein alignment scoring system sensitive at all evolutionary distances. J Mol Evol **36:** 290-300.

Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol **219:** 555-565.

Altschul, S.F. and B.W. Erickson. 1986. Optimal sequence alignment using affine gap costs. Bull Math Biol **48:** 603-616.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389-3402.

Arslan, A.N., O. Egecioglu, and P.A. Pevzner. 2001. A new approach to sequence comparison: normalized sequence alignment. Bioinformatics **17:** 327-337.

Bahr, A., J.D. Thompson, J.C. Thierry, and O. Poch. 2001. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Res **29:** 323-326.

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

Bucka-Lassen, K., O. Caprani, and J. Hein. 1999. Combining many multiple alignments in one improved alignment. Bioinformatics **15:** 122-130.

Carrilo, H. and D. Lipman. 1988. The multiple sequence alignment problem in biology. SIAM J Appl Math **48:** 1073-1082.

Depiereux, E. and E. Feytmans. 1992. MATCH-BOX: a fundamentally new algorithm for the simultaneous alignment of several protein sequences. Comput Appl Biosci **8:** 501-509.

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

Dolz, R. 1994. GCG: production of multiple sequence alignment. Methods Mol Biol **24:** 83-99.

Dress, A., G. Fullen, and S. Perrey. 1995. A divide and conquer approach to multiple alignment. Proc Int Conf Intell Syst Mol Biol **3:** 107-113.

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

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

Ellis, J. and D. Morrison. 1995. Effects of sequence alignment on the phylogeny of Sarcocystis deduced from 18S rDNA sequences. Parasitol Res **81:** 696-699.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Felsenstein, J. 1985. Confidence-Limits on Phylogenies - An Approach Using the Bootstrap. Evolution **39:** 783-791.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Feng, D.F. and R.F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol **25:** 351-360.

Frommlet, F., A. Futschik, and M. Bogdan. 2004. On the significance of sequence alignments when using multiple scoring matrices. Bioinformatics **20:** 881-887.

Gascuel, O. 1997. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14:** 685-695.

Goldstein, L. and M.S. Waterman. 1992. Poisson, compound Poisson and process approximations for testing statistical significance in sequence comparisons. Bull Math Biol **54:** 785-812.

Gonnet, G.H., M.A. Cohen, and S.A. Benner. 1992. Exhaustive matching of the entire protein sequence database. Science **256:** 1443-1445.

Gotoh, O. 1993b. Optimal alignment between groups of sequences and its application to multiple sequence alignment. Comput Appl Biosci **9:** 361-370.

Gotoh, O. 1990. Consistency of optimal sequence alignments. Bull Math Biol **52:** 509-525.

Gotoh, O. 1986a. Alignment of three biological sequences with an efficient traceback procedure. J Theor Biol **121:** 327-337.

Gupta, S.K., J.D. Kececioglu, and A.A. Schaffer. 1995. Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. J Comput Biol **2:** 459-472.

Gusfield, D. 1997. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, New York, NY.

Henikoff, S. 1991. Playing with blocks: some pitfalls of forcing multiple alignments. New Biol **3:** 1148-1154.

Henikoff, S. and J.G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A **89:** 10915-10919.

Hickson, R.E., C. Simon, and S.W. Perrey. 2000. The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. Mol Biol Evol **17:** 530-539.

Higgins, D.G., A.J. Bleasby, and R. Fuchs. 1992. CLUSTAL V: improved software for multiple sequence alignment. Comput Appl Biosci **8:** 189-191.

Higgins, D.G. and P.M. Sharp. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene **73:** 237-244.

Hirosawa, M., M. Hoshida, M. Ishikawa, and T. Toya. 1993a. MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming. Comput Appl Biosci **9:** 161-167.

Hirosawa, M., Y. Totoki, M. Hoshida, and M. Ishikawa. 1995b. Comprehensive study on iterative algorithms of multiple sequence alignment. Comput Appl Biosci **11:** 13-18.

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

Holmes, I. and R. Durbin. 1998. Dynamic programming alignment accuracy. J Comput Biol **5:** 493-504.

Jukes, T.H. and C.R. Cantor 1969. Evolution of Protein Molecules. Academic Press, New York.

Karplus, K. and B. Hu. 2001. Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. Bioinformatics **17:** 713-720.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res **30:** 3059-3066.

Kent, W.J. 2002. BLAT - The BLAST-like alignment tool. Genome Res **12:** 656-664.

Kobayashi, H. and H. Imai. 1998. Improvement of the A(*) Algorithm for Multiple Sequence Alignment. Genome Inform Ser Workshop Genome Inform **9:** 120-130.

Lake, J.A. 1991. The order of sequence alignment can bias the selection of tree topology. Mol Biol Evol **8:** 378-385.

Lassmann, T. and E.L. Sonnhammer. 2002. Quality assessment of multiple alignment programs. FEBS Lett **529:** 126-130.

Lawrence, C.E., S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, and J.C. Wootton. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science **262:** 208-214.

Lee, C., C. Grasso, and M.F. Sharlow. 2002. Multiple sequence alignment using partial order graphs. Bioinformatics **18:** 452-464.

Lombard, V., E.B. Camon, H.E. Parkinson, P. Hingamp, G. Stoesser, and N. Redaschi. 2002. EMBL-Align: a new public nucleotide and amino acid multiple sequence alignment database. Bioinformatics **18:** 763-764.

Lukashin, A.V., J. Engelbrecht, and S. Brunak. 1992. Multiple alignment using simulated annealing: branch point definition in human mRNA splicing. Nucleic Acids Res **20:** 2511-2516.

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

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

Miller, W. 1993. Building multiple alignments from pairwise alignments. Comput Appl Biosci **9:** 169-176.

Morgenstern, B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15:** 211-218.

Morgenstern, B., A. Dress, and T. Werner. 1996a. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc Natl Acad Sci U S A **93:** 12098-12103.

Morgenstern, B., K. Frech, A. Dress, and T. Werner. 1998b. DIALIGN: finding local similarities by multiple sequence alignment. Bioinformatics **14:** 290-294.

Morrison, D.A. and J.T. Ellis. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. Mol Biol Evol **14:** 428-441.

Mullan, L.J. 2002. Multiple sequence alignment--the gateway to further analysis. Brief Bioinform **3:** 303-305.

Myers, E.W. and W. Miller. 1988. Optimal alignments in linear space. Comput Appl Biosci **4:** 11-17.

Naor, D. and D.L. Brutlag. 1994. On near-optimal alignments of biological sequences. J Comput Biol **1:** 349-366.

Needleman, S.B. and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol **48:** 443-453.

Nicholas, H.B., A.J. Ropelewski, and D.W. Deerfield. 2002. Strategies for multiple sequence alignment. Biotechniques **32:** 572-4, 576, 578.

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

Notredame, C. and D.G. Higgins. 1996. SAGA: sequence alignment by genetic algorithm. Nucleic Acids Res **24:** 1515-1524.

Notredame, C., D.G. Higgins, and J. Heringa. 2000a. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol **302:** 205-217.

Notredame, C., L. Holm, and D.G. Higgins. 1998b. COFFEE: an objective function for multiple sequence alignments. Bioinformatics **14:** 407-422.

O'Brien, E.A. and D.G. Higgins. 1998. Empirical estimation of the reliability of ribosomal RNA alignments. Bioinformatics **14:** 830-838.

Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A **85:** 2444-2448.

Pearson, W.R. and W. Miller. 1992. Dynamic programming algorithms for biological sequence comparison. Methods Enzymol **210:** 575-601.

Sadreyev, R. and N. Grishin. 2003. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol **326:** 317-336.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4:** 406-425.

Sammeth, M., J. Rothganger, W. Esser, J. Albert, J. Stoye, and D. Harmsen. 2003. QAlign: quality-based multiple alignments with dynamic phylogenetic analysis. Bioinformatics **19:** 1592-1593.

Schwikowski, B. and M. Vingron. 1997. The deferred path heuristic for the generalized tree alignment problem. J Comput Biol **4:** 415-431.

Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. J Mol Biol **147:** 195-197.

Sokal, R.R. and F.J. Rohlf 1995. Biometry: the principles and practice of statistics in biological research. 3rd edition. W. H. Freeman and Co., New York.

Srinivasarao, G.Y., L.S. Yeh, C.R. Marzec, B.C. Orcutt, W.C. Barker, and F. Pfeiffer. 1999. Database of protein sequence alignments: PIR-ALN. Nucleic Acids Res **27:** 284-285.

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

Taylor, W.R. 1987. Multiple sequence alignment by a pairwise algorithm. Comput Appl Biosci **3:** 81-87.

Thompson, J.D. 1995. Introducing variable gap penalties to sequence alignment in linear space. Comput Appl Biosci **11:** 181-186.

Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res **25:** 4876-4882.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994a. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673-4680.

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

Thompson, J.D., F. Plewniak, R. Ripp, J.C. Thierry, and O. Poch. 2001c. Towards a reliable objective function for multiple sequence alignments. J Mol Biol **314:** 937-951.

Thorne, J.L. and H. Kishino. 1992. Freeing phylogenies from artifacts of alignment. Mol Biol Evol **9:** 1148-1162.

Thorne, J.L., H. Kishino, and J. Felsenstein. 1991a. An evolutionary model for maximum likelihood alignment of DNA sequences. J Mol Evol **33:** 114-124.

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

Vinga, S. and J. Almeida. 2003. Alignment-free sequence comparison-a review. Bioinformatics **19:** 513-523.

Wallace, I.M., O. O'Sullivan, and D.G. Higgins. 2005. Evaluation of iterative alignment algorithms for multiple alignment. Bioinformatics. **21:** 1408-1414.

Wang, Y. and K.B. Li. 2004. An adaptive and iterative algorithm for refining multiple sequence alignment. Comput Biol Chem **28:** 141-148.

Waterman, M.S. 1994. Estimating statistical significance of sequence alignments. Philos Trans R Soc Lond B Biol Sci **344:** 383-390.

Waterman, M.S. 1986. Multiple sequence alignment by consensus. Nucleic Acids Res **14:** 9095-9102.

Waterman, M.S. 1995. Introduction to Computational Biology: Maps, Sequences, and Genomes. Chapman & Hall,

Waterman, M.S. and M. Vingron. 1994. Rapid and accurate estimates of statistical significance for sequence data base searches. Proc Natl Acad Sci U S A **91:** 4625-4628.

Webb, B.J., J.S. Liu, and C.E. Lawrence. 2002. BALSA: Bayesian algorithm for local sequence alignment. Nucleic Acids Res **30:** 1268-1277.

Wheeler, W. 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst Biol **44:** 321-331.

Womble, D.D. 2000. GCG: The Wisconsin Package of sequence analysis programs. Methods Mol Biol **132:** 3-22.

Yu, L. and T.F. Smith. 1999. Positional statistical significance in sequence alignment. J Comput Biol **6:** 253-259.

Zar, J.H. 1999. Biostatistical Analysis. 4th edition. Prentice-Hall, Inc., Upper Saddle River, NJ.

Zhang, M.Q. and T.G. Marr. 1995. Alignment of molecular sequences seen as random path analysis. J Theor Biol **174:** 119-129.

Zhu, J., J. Liu, and C. Lawrence. 1997. Bayesian adaptive alignment and inference. Proc Int Conf Intell Syst Mol Biol **5:** 358-368.

Zweig, M.H. and G. Campbell. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine [published erratum appears in Clin Chem 1993 Aug;39(8):1589]. Clin Chem **39:** 561-577.

Appendix - A Brief History of MSA

I do not pretend to provide an exhaustive, all-encompassing, definitive, textbook-like review of the whole literature pertaining to sequence alignment. Such a compilation would exceed the space limit for a Ph.D. thesis by several hundred percents. I believe that my treatment of the literature covers all the basic works pertaining to my thesis, and is a sufficient starting point for any potential student of the field.

The early years: 1970-1988

In the early days of sequencing, published sequences were few, and they were predominantly amino-acids sequences. In 1970, Needleman and Wunsch first described a method for the pairwise alignment of two protein sequences (Table A.1). With the accumulation of sequences, the need arose for the identification of possible homologous sequences, and for the simultaneous alignment of more than two sequences.

Table A.1: Timeline of major developments in sequence alignment: the early years

Year	Authors	Title
1970	Needleman and Wunsch	A general method applicable to the search for similarities in the amino acid sequence of two proteins
1981	Smith and Waterman	Identification of common molecular subsequences
1986	Waterman	Multiple sequence alignment by consensus
1986	Altschul and Erickson	Optimal sequence alignment using affine gap costs
1986	Gotoh	Alignment of three biological sequences with an efficient traceback procedure
1986	Bishop and Thompson	Maximum likelihood alignment of DNA sequences
1987	Taylor	Multiple sequence alignment by a pairwise algorithm
1987	Feng and Doolittle	Progressive sequence alignment as a prerequisite to correct phylogenetic trees
1988	Higgins and Sharp	CLUSTAL: a package for performing multiple sequence alignment on a microcomputer

Searching sequence databases was first addressed in 1981 by Smith and Waterman, with the development of local pairwise alignment. Local alignment, which facilitates sequence searches, had to be distinguished from the alignment of sequences in their entirety for detailed comparative purposes, a task that was rechristened as "global" alignment. In this study I have addressed only global alignment issues.

Global alignments were next improved by Altschul and Erickson in 1986, who introduced affine gap costs which greatly enhanced their biological relevance. The simultaneous alignment of more than two sequences followed shortly after with the work of Gotoh (1986) and Taylor (1987), and was culminated by the introduction of progressive multiple sequence alignment by Feng and Doolittle (1987), and by the first version of standard MSA reconstruction software, CLUSTAL (Higgins and Sharp, 1988).

Consolidation: 1988-1994

The following years (Table A.2) were dominated by improvements in the performances of MSA alignment methods. First, purely algorithmic aspects were improved: run-times and space requirements were reduced, resulting in the ability to analyze larger data sets. Scoring systems were also improved to provide MSAs that were more realistic biologically.

Table A.2: Timeline of major developments in sequence alignment: consolidation

Year	Authors	Title
1988	Carrilo and Lipman	The multiple sequence alignment problem in biology
1988	Myers and Miller	Optimal alignments in linear space
1991	Thorne *et al.*	An evolutionary model for maximum likelihood alignment of DNA sequences
1992	Allison *et al.*	Finite-state models in the alignment of macromolecules
1992	Depiereux and Feytmans	MATCH-BOX: a fundamentally new algorithm for the simultaneous alignment of several protein sequences
1992	Higgins *et al.*	CLUSTAL V: improved software for multiple sequence alignment
1992	Lukashin *et al.*	Multiple alignment using simulated annealing: branch point definition in human mRNA splicing
1992	Thorne *et al.*	Inching toward reality: an improved likelihood model of sequence evolution
1993	Altschul	A protein alignment scoring system sensitive at all evolutionary distances
1993	Gotoh	Optimal alignment between groups of sequences and its application to multiple sequence alignment
1993	Hirosawa *et al.*	MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming
1993	Miller	Building multiple alignments from pairwise alignments
1993	Lawrence *et al.*	Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment
1994	Dolz	GCG: production of multiple sequence alignment
1994	Thompson *et al.*	CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice

At the same time, the view was extended towards alternative evolutionary models and algorithmic approaches. Most of the biologically relevant improvements were implemented in the ClustalW program (Thompson *et al.*, 1994a), which became the standard tool for MSA reconstruction.

The proliferation era: 1994-present

Since the publication of ClustalW in 1994, four major trends can be discerned (Table A.3):

a. New MSA reconstruction methods are constantly being developed. Some are motivated by algorithmic and statistical considerations, others introduce new evolutionary models, and yet others address specific biological problems. Yet, ClustalW is still considered the standard and most reliable method. Only recently were possible heirs to ClustalW developed: the MUSCLE program (Edgar, 2004), and the ProbCons program (Do *et al.*, 2005).

b. MSAs became essential in structural analysis, and methods that reconstruct MSAs from the structural non-historical viewpoint were developed.

c. MSA reconstruction became a standard ingredient of high-throughput analysis systems, addressing various biological problems, and resulting in numerous biological databases. Once again, in most cases MSAs are produced by ClustalW.

d. MSA quality issues were started to be studied from various perspectives.

Table A.3: Timeline of major developments in sequence alignment: proliferation

Year	Authors	Title
1995	Dress *et al.*	A divide and conquer approach to multiple alignment
1995	Gupta *et al.*	Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment
1995	Eddy	Multiple alignment using hidden Markov models
1995	Hirosawa *et al.*	Comprehensive study on iterative algorithms of multiple sequence alignment
1995	Thompson	Introducing variable gap penalties to sequence alignment in linear space
1995	Zhang and Marr	Alignment of molecular sequences seen as random path analysis
1996	Morgenstern *et al.*	Multiple DNA and protein sequence alignment based on segment-to-segment comparison
1996	Notredame and Higgins	SAGA: sequence alignment by genetic algorithm
1997	Altschul *et al.*	Gapped BLAST and PSI-BLAST: a new generation of protein database search programs
1997	Schwikowski and Vingron	The deferred path heuristic for the generalized tree alignment problem
1997	Zhu *et al.*	Bayesian adaptive alignment and inference
1998	Kobayashi and Imai	Improvement of the A(*) Algorithm for Multiple Sequence Alignment
1998	Morgenstern *et al.*	DIALIGN: finding local similarities by multiple sequence alignment
1998	Notredame *et al.*	COFFEE: an objective function for multiple sequence alignments
1999	Bucka-Lassen *et al.*	Combining many multiple alignments in one improved alignment
1999	Morgenstern	DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment
1999	Thompson *et al.*	A comprehensive comparison of multiple sequence alignment programs
2000	Notredame *et al.*	T-Coffee: A novel method for fast and accurate multiple sequence alignment
2001	Arslan *et al.*	A new approach to sequence comparison: normalized sequence alignment

Year	Authors	Title
2001	Holmes and Bruno	Evolutionary HMMs: a Bayesian approach to multiple alignment
2001	Thompson *et al.*	Towards a reliable objective function for multiple sequence alignments
2002	Althaus *et al.*	Multiple sequence alignment with arbitrary gap costs: Computing an optimal solution using polyhedral combinatorics
2002	Katoh *et al.*	MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform
2002	Kent	BLAT - The BLAST-like alignment tool
2002	Lee *et al.*	Multiple sequence alignment using partial order graphs
2002	Miklos	An improved algorithm for statistical alignment of sequences related by a star tree
2002	Webb *et al.*	BALSA: Bayesian algorithm for local sequence alignment
2003	Sadreyev and Grishin	COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance
2003	Sammeth *et al.*	QAlign: quality-based multiple alignments with dynamic phylogenetic analysis
2004	Edgar	MUSCLE: multiple sequence alignment with high accuracy and high throughput
2004	Wang and Li	An adaptive and iterative algorithm for refining multiple sequence alignment
2005	Do *et al.*	ProbCons: Probabilistic consistency-based multiple sequence alignment
2005	Wallace *et al.*	Evaluation of iterative alignment algorithms for multiple alignment