Gene 441 (2009) 141–147

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Characterization of pairwise and multiple sequence alignment errors

Giddy Landan *, Dan Graur

Department of Biology & Biochemistry, University of Houston, Houston, TX, USA

ARTICLE INFO

Article history:
Received 27 February 2008
Received in revised form 21 May 2008
Accepted 22 May 2008
Available online 3 June 2008

Received by A. Bernardi

Keywords:
Multiple sequence alignment
Pairwise sequence alignment
Alignment errors

ABSTRACT

We characterize pairwise and multiple sequence alignment (MSA) errors by comparing true alignments from simulations of sequence evolution with reconstructed alignments. The vast majority of reconstructed alignments contain many errors. Error rates rapidly increase with sequence divergence, thus, for even intermediate degrees of sequence divergence, more than half of the columns of a reconstructed alignment may be expected to be erroneous. In closely related sequences, most errors consist of the erroneous positioning of a single indel event and their effect is local. As sequences diverge, errors become more complex as a result of the simultaneous mis-reconstruction of many indel events, and the lengths of the affected MSA segments increase dramatically. We found a systematic bias towards underestimation of the number of gaps, which leads to the reconstructed MSA being on average shorter than the true one. Alignment errors are unavoidable even when the evolutionary parameters are known in advance. Correct reconstruction can only be guaranteed when the likelihood of true alignment is uniquely optimal. However, true alignment features are very frequently sub-optimal or co-optimal, with the result that optimal albeit erroneous features are incorporated into the reconstructed MSA. Progressive MSA utilizes a guide-tree in the reconstruction of MSAs. The quality of the guide-tree was found to affect MSA error levels only marginally.

1. Introduction

Sequence alignment is the most basic analysis used in the comparative study of molecular sequences (nucleic acids and proteins). It entails the identification of the location of insertions and deletions (indels) that might have occurred since the divergence of the sequences from a common molecular ancestor. In essence, sequence alignment is an inference algorithm designed to identify positional homologies, i.e., residues that had descended from one ancestral residue. Sequence alignment is the starting point of almost all analyses that involve the comparison of molecular data (Mullan, 2002), e.g., derivation of sequence similarity measures, identification of homologous sites, phylogenetic reconstruction, identification of functional domains, and three-dimensional structure prediction. The fundamental role of multiple sequence alignment is best demonstrated by noting that papers describing multiple-alignment reconstruction methods, in particular ClustalW (Thompson et al., 1994), are among the most cited papers in the literature. Being a fundamental ingredient in a wide variety of analyses, an issue of utmost importance is MSA 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; Ogden and Rosenberg, 2006; Kumar and Filipski, 2007).

The alignment of molecular sequences was first described by Needleman and Wunsch (1970). Since then the theory and art of sequence alignment reconstruction has witnessed a proliferation of alignment algorithms aiming at improving computational feasibility and performance, on the one hand, and the biological relevance and quality of the deduced alignments, on the other (for reviews, see McClure et al., 1994; Hirosawa et al., 1995; Waterman, 1995; Gusfield, 1997; Thompson et al., 1999; Nicholas et al., 2002; Notredame, 2002; Edgar and Batzoglou 2006; Notredame, 2007). By a huge margin, the most widely used alignment method is ClustalW (Thompson et al., 1994). ClustalW produces 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 use ClustalW as the standard in MSA reconstruction.

Many researchers routinely rely on reconstructed MSAs implicitly. This is so even though deduced sequence alignments are known to be unreliable and inaccurate (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). In some studies, different alignment algorithms were compared in terms of alignment quality, mostly focusing on their ability to reconstruct large-scale features of reference alignments (McClure et al., 1994; Thompson et al., 1999; Lassmann and Sonnhammer, 2002). In contrast, little attention has

Abbreviations: MSA, Multiple sequence alignment; PWA, pairwise alignment.

* Corresponding author. Department of Biology & Biochemistry, University of Houston, 369 Science & Research Building 2, 4800 Calhoun Road, Houston, TX 77204-5001, USA. Tel.: +1 713 7437236; fax: +1 713 7432636.
E-mail addresses: giddy.landan@gmail.com (G. Landan), dgraur@uh.edu (D. Graur).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.05.016

been given to the fine-detail quality of multiple sequence alignment (but, see Thorne and Kishino, 1992; Thorne et al., 1992; Wheeler, 1995; Holmes and Durbin; 1998, Hickson et al., 2000; Golubchik et al., 2007.)

Here, 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 in the reconstructions.

2. Methods

2.1. Evolutionary simulations

Sequence evolution was simulated using ROSE (Stoye et al., 1998). We simulated DNA sequences of length 500 on average. The guiding phylogeny was a 16 OTU balanced binary tree. All branches were of the same length, both in terms of substitution and indel rates. Substitution rates spanned values that produce an average pairwise distance ranging from 0.02 to 0.30 substitutions per site. Insertion and deletion rates and length distribution were equal, producing an average pairwise gap content ranging from 0.001 to 0.022 gaps per site. Overall, we used 8 substitution levels and 8 indel levels, and for each of the 64 combinations we simulated 100 datasets for a total of 6400 datasets. The results relating to pairwise alignments were obtained by considering only two of the most distant sequences within each 16 OTU dataset. The range of simulation parameters in ROSE was chosen to reflect present knowledge of real-life substitution, deletion, and insertion patterns.

2.2. Alignment reconstruction methods

Pairwise alignments were reconstructed with ALIGN (Pearson and Lipman, 1988). ClustalW (Thompson et al., 1994) was used for MSA reconstruction.

2.3. Comparison of MSAs

The true alignments from simulation were compared to reconstructed alignments using the method described in Thompson et al. (1999). Our "column reconstruction rate" is Thompson et al.'s (1999) CS measure, and the "column error rate" is its complement. Our "residue-pair reconstruction rate" is Thompson et al.'s (1999) SPS measure, and the "residue-pair error rate" is its complement. Note that for the case of pairwise alignment, the columns and residue-pair measures are identical.

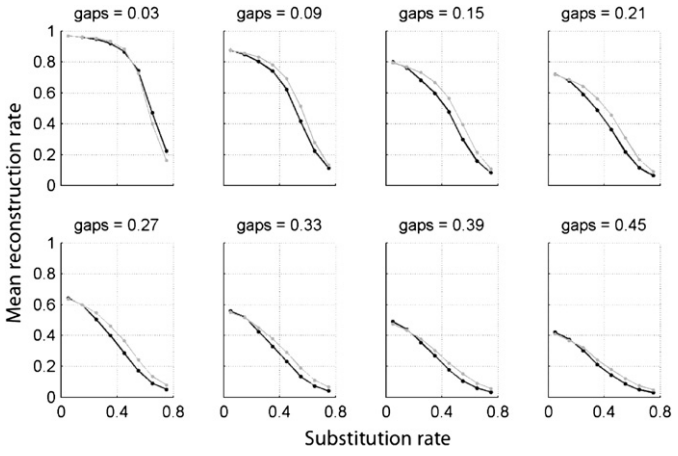

Fig. 1. Mean pairwise alignment reconstruction rates with default (black) and true (gray) penalties. The abscissa is the substitution rate and the different panels indicate different indel rates. Each point is an average over 100 simulation runs at one combination of 8 substitution levels and 8 indel levels. To avoid clutter, standard errors were left out and are reported in the text where appropriate.

The column reconstruction score for individual columns was used to decompose the two alignments into alternating segments of correct and erroneous reconstruction. By definition, the correct segments are identical between the true and reconstructed alignments.

3. Results

3.1. Pairwise alignment errors

We start our characterization of alignment errors by considering the simplest case of pairwise alignment (PWA). In addition to being a special case of MSA, pairwise alignments are also the building blocks of MSAs. Pairwise alignments were reconstructed using the ALIGN program at its default parameter values (Pearson and Lipman, 1988; match = 5; mismatch = −4; gap-open = −16; gap-extent = −4).

The overall reconstruction rate depends on the actual divergence of the sequences, with reconstruction rates rapidly deteriorating with divergence (Fig. 1, black lines). The default parameters are thought to be adequate for a wide range of practical problems, and are indeed reasonable when no prior knowledge of evolutionary parameters is available. It is expected, however, that using penalty scores that correspond to the true evolutionary parameters will produce better quality alignments, and that the default penalty values may introduce a bias that will result in reconstruction errors.

To quantify the level of errors resulting from inadequate penalties, we repeated the analysis using the exact penalty scores corresponding to the true alignment (Fig. 1, gray lines). The PWA reconstruction rates achieved when the true evolutionary parameters are known in advance are only marginally higher (3–10%) than those achieved with default values. It follows that although appropriate penalties are desirable, using the default values is by no means the principal source of error. Since in providing the true parameters we used all the available prior knowledge, the resulting reconstruction rates represent the maximum reconstruction level that can be attained by PWA. We must emphasize, however, that in real life it is impossible to provide true parameters, because the true alignment in not known in advance. Even under such favorable albeit unrealistic conditions, the alignment error rate is quite high.

Let us, now, characterize these unavoidable errors. Given a reconstructed PWA and the corresponding true alignment, the two 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 error segments in the reconstructed PWA correspond to mis-reconstructed segments of the true alignment.

Considering the actual objective function scores, reconstruction errors can be classified into co-optimal or sub-optimal alignments. First, under any scoring function, many different alignments may attain the same maximal score. All these alignments are equivalent (co-optimal), and without outside knowledge the alignment produced by PWA programs is merely an arbitrary choice from the set of co-optimal alignments. Second, the true alignment, being some concrete realization of a stochastic process, may be sub-optimal. This leads to the situation where an erroneous alignment segment may be assigned a higher score than the true alignment segment even by an exact scoring function. In other words, a true alignment may be sub-optimal in many of its elements. In contrast, the reconstructed PWA is always, by definition, optimal according to the objective function, as are all its segments. 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 (Fig. 2). We note that even under the most favorable circumstances of close sequence relatedness, sub-optimality

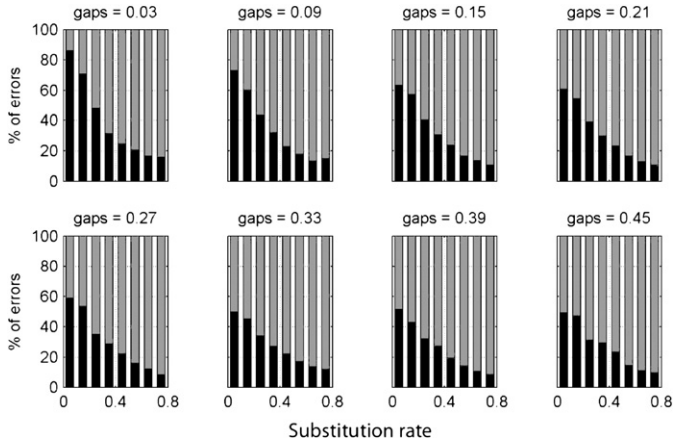

Fig. 2. Relative frequency of error pairwise alignment segments where the corresponding true segment is co-optimal (black) or sub-optimal (gray). (See Fig. 1 for details of layout).

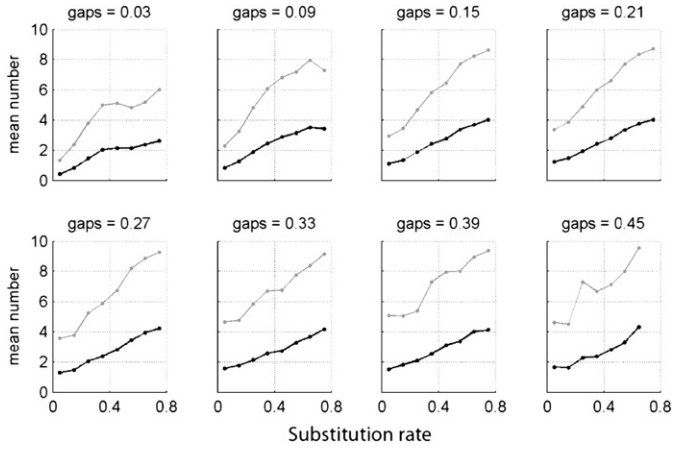

Fig. 4. Mean number of indels (black) and gap characters (gray) per error PWA segment as a function of sequence divergence. (See Fig. 1 for details of layout).

accounts for at least 50% of all errors. That is, the alignment is over-fitted spuriously to maximize the objective function score.

Next, we note that the mean length of error segments (Fig. 3, gray lines) increases dramatically with substitution rate, while the mean length of correctly reconstructed segments remains fairly stable (Fig. 3, black lines). While the length of error segments increase with divergence, we note that erroneously reconstructed segments contain fewer indels (and gap characters) and are shorter than the corresponding true segments. This is a systematic 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.

The mean numbers of wrongly inferred indels and gap-character states increases with substitution rate (Fig. 4). For closely related sequences, the error segments are short and frequently result from a single indel being erroneously positioned. As the two sequences farther diverge, the errors multiply. At the same time, neighboring indels in the true alignment begin interfering with one another to produce error segments where several 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.

When the number of indels involved in a single error segment is relatively small, it is possible to describe the detailed structure of the errors. We define "simple" errors to be those involving at most two indels in both the true PWA segment and the corresponding erroneously reconstructed segment. Such simple alignment errors may be classified into several types. "Shift" (Fig. 5a) is the erroneous positioning of a single indel whose 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 distance between the true and the erroneous positions. The affected region increases with substitution rates. "Split" (Fig. 5b) is a single true indel that in the reconstruction was split into two, either on the same sequence or one indel per sequence. The true indel length may not be preserved in any of the two erroneous indels, but the difference in gap content between the two sequences should remain the same. "Merge" (Fig. 5c) is an error in which two indels, whether on the same sequence or one on each sequence, are reconstructed as a single indel. "Ex-nihilo" (Fig. 5d) is the erroneous inference of two indels of equal lengths, one in each sequence, where no indel is present in the true alignment. This type of error, in which non-existing indels are inferred to exist, can be regarded as the extreme case of a split error. "Purge" (Fig. 5e) involves the disappearance of two indels of equal length, one in each sequence. The resulting error segment is gapless. Purge may be regarded as an extreme form of merge. All other errors are complex (Fig. 5f), i.e., compounds of two or more alignment errors.

With increasing sequence divergence, the simple errors account for fewer and fewer cases out of the overall errors (Fig. 6). Among the errors affecting two indels, the errors that result in fewer indels, merge and purge (blues), are much more frequent than the errors resulting in more indels, split and ex-nihilo (oranges). This is another demonstration of the bias towards the minimization of inferred indel events.

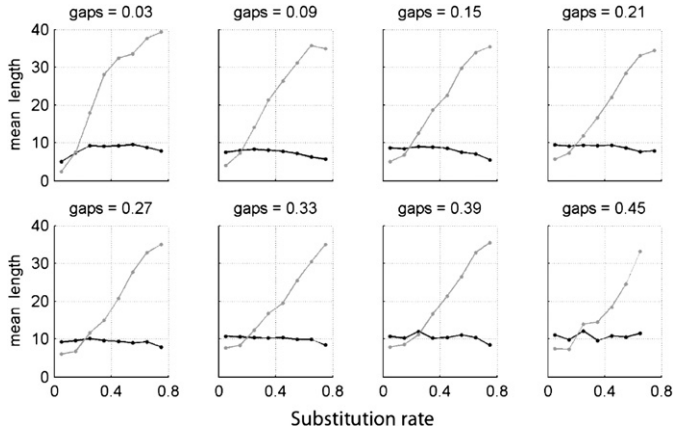

Fig. 3. Mean number of residues in correctly (black) and erroneously (gray) reconstructed PWA segments as a function of sequence divergence. (See Fig. 1 for details of layout).

3.2. Multiple sequence alignment errors

To study the errors in MSA reconstruction, we compared true MSAs from simulations to reconstructed MSAs produced by ClustalW (Thompson et al., 1994) with default parameters. Note that ClustalW employs internal estimation of evolutionary parameters to derive penalty values, so the default values are less critical than those used in PWA algorithms, such as ALIGN (Pearson and Lipman, 1988).

First we present the overall error rates in MSA reconstruction. Fig. 7 summarizes the mean error rates as a function of sequence divergence. The residue-pairs error rates, (Fig. 7 black lines), range from $\sim 5 \pm 2\%$ for very closely related sequences to $90 \pm 7\%$ for very distantly related sequences, with a monotonic dependency on the evolutionary rates. Apart from very closely related sequences, the column error rate, (Fig. 7 gray lines), is higher than 50%, and rapidly reaches 100%, that is, all the columns in the MSA are mis-reconstructed.

(a) Shift

```
gcA-taTcaActCTcagaatCGt    TAC----------TacTGTA    TTAc----TTa    TcG-gGaTGGa   true
cgAcgcTtgAtcCTggtttgCGg    TACtcgtcgcctcTgtTGTA    TTAatcccTTg    TgGccGgTGGg

gcAtatcaacTCtcaGaaT-CGt    TACTacT----------GTA    TTA----CTTa    TcGg-GaTGGa   error
cgAcgcttgaTCctgGttTgCGg    TACTcgTcgcctctgttGTA    TTAatccCTTg    TgGccGgTGGg
```

(b) Split

```
ACggtacTtCagaTag    TaC----------TactGTAatTtg    ATAgaacggtacttcAgAtagTaaTc   true
ACat--gTaCctcTcc    TgCaaggcgcctcTgtaGTAccTca    ATAttgact-----tAaAaccTcgTt

ACg-GTACtTCagaTag   TaCtA-------CTGTAaT---Ttg    ATAgaacggtACTTcAgAtagTaaTc   error
ACatGTACcTC---Tcc   TgCaAggcgcctCTGTAgTaccTca    ATA----ttgACTTaAaAcc-TcgTt
```

(c) Merge

```
GGatttcaTTtGcaTtC-GaaCtagcagccacaaAGtA    GcAgt-tAgA    AAC-Gg-tccGGaaATaCGgGcaATAc   true
GGggcgttTTaGagTgCtGggCcct---------AGcA    GtA--ccAcA    AACaGatagaGGttATtCGaGttATAt

GGatttcatttGCaTTcgAactaGCaGccaCaaAGtA     GcAgttAgA     AACgGtccG--GaaATaCGgGcaATAc   error
GGg--------GCgTTttAgagtGCtGggcCctAGcA     GtAcc-AcA     AACaGataGagGttATtCGaGttATAt
```

(d) Ex-nihilo

```
GGcaGgGcgcaaaCTTgC   true
GGggGcGgattgcCTTcC

GGcaGGGCGcAaa--CTTgC   error
GG--GGGCGgAttgcCTTcC
```

(e) Purge

```
TGggg-tACAT    AGaAaCatgTacctctccTAA-TC   true
TGta-ccACAT    AG-AcCgccTtgagactaTAAaTC

TGgggtACAT     AGAaacatgTacctCTcctAATC    error
TGtaccACAT     AGAccgcctTgagaCTataAATC
```

(f) Complex

```
GaaATaCGgGcaATAcc----------cactgtaatt----TgCG-acGCgGacGcaGttagaTTtGcaTtCcGaaCtaT   true
GttATtCGaGttATAt-aatgcgtcga--------cgtagtTaCGgtgGCcGgtGggG-cgttTTaGagTgCtGggCccT

GaaATaCGgGcaATAccactGT--AatTtGcgACGcgGaC--GcaGttaGaTTTgcAtTcC-GaaCtaT   error
GttATtCGaGttATAtaatgcGTcgAcgTaGttACGgtGgCcgGtgGggcGtTTTagAgTgCtGggCccT
```

Fig. 5. Examples of the five simple pairwise alignment errors and a complex error. Gray parts of the alignments are correctly reconstructed and delimit the error segments.

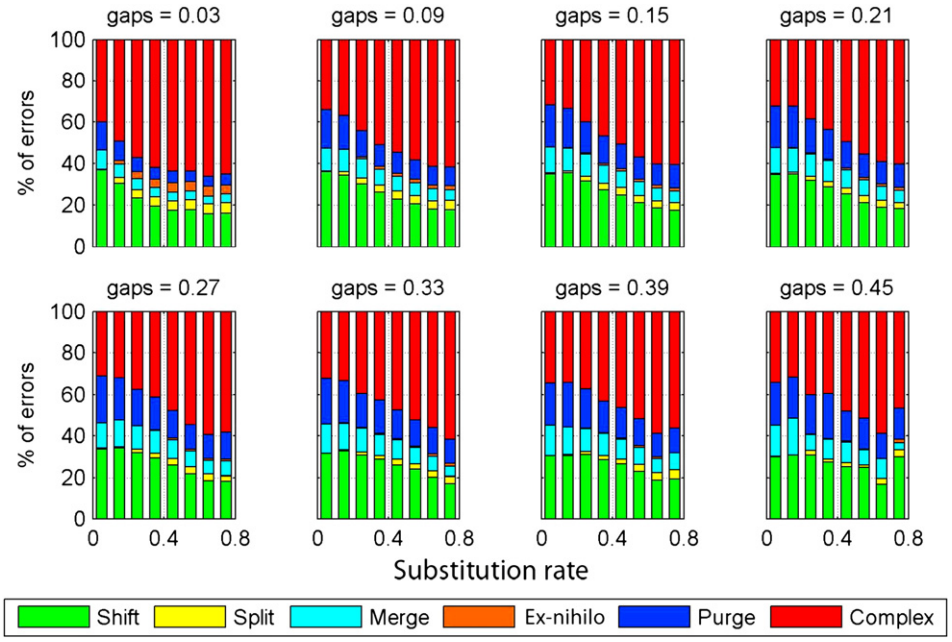

Fig. 6. Relative frequencies of the six pairwise alignment error types. (See Fig. 1 for details of layout).

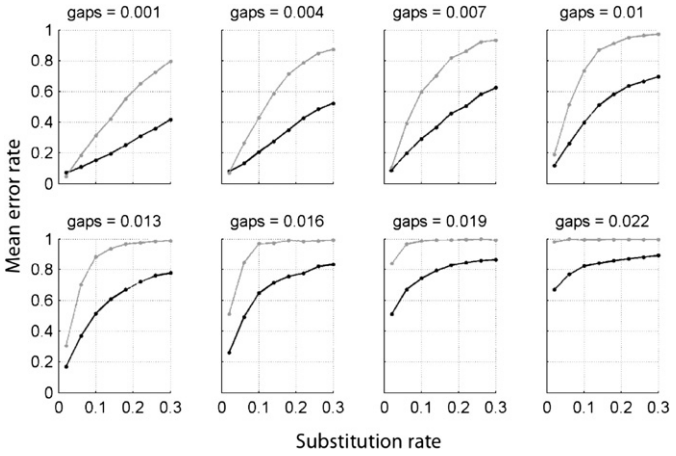

Fig. 7. Mean MSA error rates as a function of sequence divergence. Two errors rates are reported: residue-pairs error rate (black) and column error rate (gray).

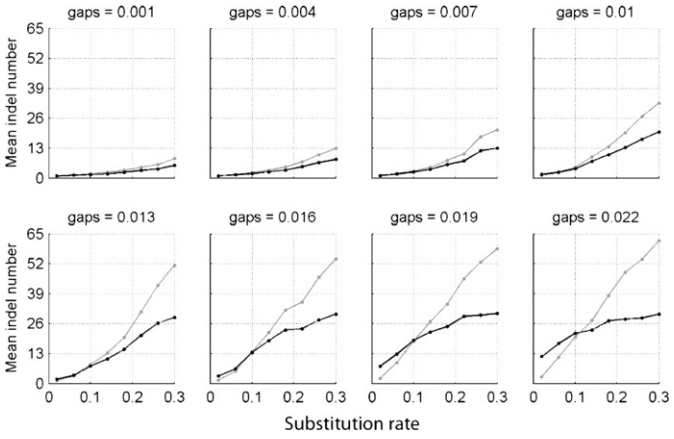

Fig. 8. Error MSA segment sizes as a function of sequence divergence. Mean length of error segments (black) and mean length of corresponding true segments (gray). (See Fig. 1 for details of layout).

The first step in progressive MSA reconstruction methods is the estimation of 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, as well as the penalties for the several pairwise alignment steps. Thus, we need to consider the possibility that the errors in the guide-tree produce errors in the reconstructed alignments (Lake, 1991). To assess the contribution of guide-tree inaccuracies to the MSA error rates, we consider MSAs that are guided by the true underlying phylogeny. We find that such "assisted" MSAs are only marginally better than the standard MSAs produced by employing the 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 a major source of error in MSA reconstruction.

Comparing the reconstructed alignments to the true alignments from the simulation we first note that reconstruction errors occur much more frequently in columns with gaps than in columns with no gaps ("anchor" columns). For example, for one combination of simulation parameters of intermediate divergence only 40% of the columns are correctly reconstructed and the vast majority (80%) of those are anchor columns. The error rate in anchor columns is 47%, whereas in gapped columns it reaches 79%. The difference of error rates between anchor and gapped columns reflects the nature of the problem, that is, 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 effect on both anchor and gapped columns.

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. In high quality reconstructions, error segments are short and wide apart, and encompass only a few indels. As the overall error rate increases, so does the length of error segments (Fig. 8). An erroneously reconstructed segment can contain any number of anchor and gapped columns that are different in the native and reconstructed alignment. As far as erroneously reconstructed MSA segments are concerned, the true mean length (Fig. 8, gray) is longer than the reconstructed length (Fig. 8, black), and the discrepancy increases with sequence divergence. Since the number of residues in both segments is identical, the length decrease of reconstructed segments is wholly due to lower content of gap characters in these segments.

To probe the fine details of error segments, we categorized errors by the number of indels involved in the true and erroneous segments.

Fig. 9 presents the relative abundance of shift errors (green), errors involving two indels only (blue), and complex errors involving three or more indels (red). Errors consisting of misplacement of very few indels are prevalent when the number of substitutions is small, when indels are rare, and when intervening anchor stretches are long. The presence of conserved anchor stretches isolates and limits the range of error segments. For example, in a subset of closely related sequences with an overall error rate of 10%, we find that 68% of the error segments involve just one shifted indel, whereas only 1% of the segments involve more than three indels. As evolutionary distances increase, the density of gapped columns increases, and errors at neighboring positions are merged to produce longer error segments, comprising many simultaneously misplaced indels. In such cases, the overall result is of a compounded nature and is hard to interpret. For example, in a subset of sequences of intermediate divergence with an overall error rate of 40%, only 2.5% of the error segments involve one misplaced indel, whereas 90% of the segments involve more than three indels.

Considering the relative abundance of the error types as a function of sequences divergence (Fig. 9), we note that the transition from simple errors to complex ones is much sharper than was observed earlier for PWA errors (Fig. 6). This can be understood by noting that MSAs are reconstructed by a series of pairwise profile alignments, so that even if in each pairwise step the errors are strictly shift errors, compounding them will produce complex errors in the resulting MSA.

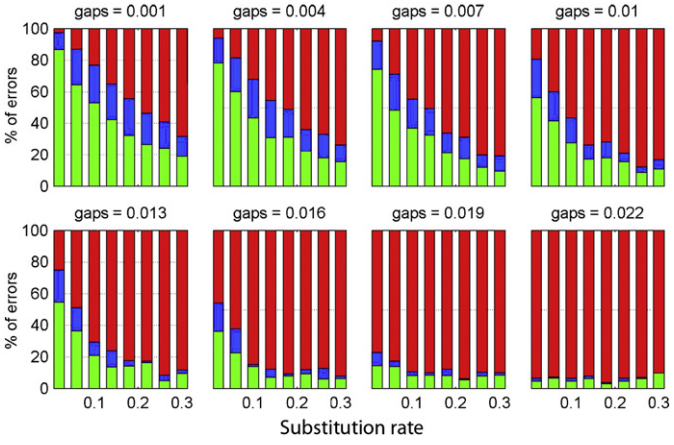

Fig. 9. Relative frequencies of three MSA error types. Shift errors are in green, errors involving two indels only are in blue, and complex errors involving three or more indels are in red.

4. Discussion

We have opted to characterize the errors of only one MSA reconstruction method, ClustalW (Thompson et al., 1994a), because it is the most widely used by a huge margin, and because it fairly represents the broad class of progressive alignment methods. Although several newer MSA methods (e.g., MAFFT, Katoh et al., 2005; PROBCONS, Do et al., 2005; M-COFFEE, Wallace et al., 2006) have been shown to perform significantly better than ClustalW, the improvement in accuracy is less than 10% (Wong et al., 2008). Our evolutionary simulation process was kept simple, with substitutions and indels as the only types of sequence change, and with no among-site rate variation. 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 evolutionary phenomena. It is, therefore, expected that the reconstruction rates we report here represent an upper limit of the performance of MSA reconstruction, and that MSAs of real biological sequences will typically have even higher error rates.

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.

The immediate source of MSA reconstruction errors is in the erroneous deduction and positioning of gaps. 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. In other words, correct reconstruction can only be guaranteed in the exceptionally unlikely case when the likelihood of the true alignment is uniquely optimal.

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. A possible explanation for this finding is that the expectation is valid only for those segments in 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 evolutionary rates.

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 final conclusion is that meaningful alignments can only be obtained if the homologous sequences are long, very closely related, and have accumulated only very few and far-between deletions and insertions. Unfortunately, in the real world, sequences are frequently short, distantly related, and have accumulated a great number of spatially clustered deletions and insertions. Thus, at even moderate evolutionary distances, reconstructed alignments are correct for only about half of their length. What happens in subsequent analyses that implicitly assume that the alignments they use are correct can only be described as calamitous. This situation clearly requires methods for the identification and management of MSA errors, such as HoT and COS (Landan and Graur, 2007; 2008).

Acknowledgment

This paper is dedicated to Prof. Masami Hasegawa, a pioneer in the field of statistical molecular phylogeny. This work was supported by NSF grant DBI-0543342.

References

Do, C.B., Mahabhashyam, M.S., Brudno, M., Batzoglou, S., 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. 15 (2), 330–340.
Edgar, R.C., Batzoglou, S., 2006. Multiple sequence alignment. Curr. Opin. Struck. Biol. 16 (3), 368–373.
Ellis, J., Morrison, D., 1995. Effects of sequence alignment on the phylogeny of Sarcocystis deduced from 18S rDNA sequences. Parasitol. Res. 81, 696–699.
Feng, D.F., Doolittle, R.F., 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25, 351–360.
Frommlet, F., Futschik, A., Bogdan, M., 2004. On the significance of sequence alignments when using multiple scoring matrices. Bioinformatics 20, 881–887.
Goldstein, L., Waterman, M.S., 1992. Poisson, compound Poisson and process approximations for testing statistical significance in sequence comparisons. Bull. Math. Biol. 54, 785–812.
Golubchik, T., Wise, M.J., Easteal, S., Jermiin, L.S., 2007. Mind the gaps: evidence of bias in estimates of multiple sequence alignments. Mol. Biol. Evol. 24 (11), 2433–2442.
Gotoh, O., 1990. Consistency of optimal sequence alignments. Bull. Math. Biol. 52, 509–525.
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.
Hickson, R.E., Simon, C., Perrey, S.W., 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.
Hirosawa, M., Totoki, Y., Hoshida, M., Ishikawa, M., 1995. Comprehensive study on iterative algorithms of multiple sequence alignment. Comput. Appl. Biosci. 11, 13–18.
Holmes, I., Durbin, R., 1998. Dynamic programming alignment accuracy. J. Comput. Biol. 5, 493–504.
Katoh, K., Kuma, K., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33 (2), 511–518.
Kumar, S., Filipski, A., 2007. Multiple sequence alignment: in pursuit of homologous DNA positions. Genome Res. 17, 127–135.
Lake, J.A., 1991. The order of sequence alignment can bias the selection of tree topology. Mol. Biol. Evol. 8, 378–385.
Landan, G., Graur, D., 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol. Biol. Evol. 24, 1380–1383.
Landan, G., Graur, D., 2008. Local reliability measures from sets of co-optimal multiple sequence alignments. Pac. Symp. Biocomput. 13, 15–24.
Lassmann, T., Sonnhammer, E.L., 2002. Quality assessment of multiple alignment programs. FEBS Lett. 529, 126–130.

McClure, M.A., Vasi, T.K., Fitch, W.M., 1994. Comparative analysis of multiple protein-sequence alignment methods. Mol. Biol. Evol. 11, 571–592.
Morrison, D.A., Ellis, J.T., 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.
Needleman, S.B., Wunsch, C.D., 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., Ropelewski, A.J., Deerfield, D.W., 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., 2007. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput. Biol. 3 (8), e123.
O'Brien, E.A., Higgins, D.G., 1998. Empirical estimation of the reliability of ribosomal RNA alignments. Bioinformatics 14, 830–838.
Ogden, T.H., Rosenberg, M.S., 2006. Multiple sequence alignment accuracy and phylogenetic inference. Syst. Biol 55, 314–328.
Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U. S. A. 85, 2444–2448.
Stoye, J., Evers, D., Meyer, F., 1998. Rose: generating sequence families. Bioinformatics 14, 157–163.
Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. 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., Plewniak, F., Poch, O., 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 27, 2682–2690.
Thorne, J.L., Kishino, H., 1992. Freeing phylogenies from artifacts of alignment. Mol. Biol. Evol. 9, 1148–1162.
Thorne, J.L., Kishino, H., Felsenstein, J., 1992. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. 34, 3–16.
Wallace, I.M., O'Sullivan, O., Higgins, D.G., Notredame, C., 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. 34 (6), 1692–1699.
Waterman, M.S., 1994. Estimating statistical significance of sequence alignments. Philos. Trans. R. Soc. Lond., B Biol. Sci. 344, 383–390.
Waterman, M.S., 1995. Introduction to Computational Biology: Maps, Sequences, and Genomes. Chapman & Hall, London.
Waterman, M.S., Vingron, M., 1994. Rapid and accurate estimates of statistical significance for sequence data base searches. Proc. Natl. Acad. Sci. U. S. A. 91, 4625–4628.
Wheeler, W., 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst. Biol. 44, 321–331.
Wong, K.M., Suchard, M.A., Huelsenbeck, J.P., 2008. Alignment uncertainty and genomic analysis. Science 319, 473–476.
Yu, L., Smith, T.F., 1999. Positional statistical significance in sequence alignment. J. Comput. Biol. 6, 253–259.