Pacific Symposium on Biocomputing 13:15-24(2008)

LOCAL RELIABILITY MEASURES FROM SETS OF CO-OPTIMAL MULTIPLE SEQUENCE ALIGNMENTS

GIDDY LANDAN
DAN GRAUR

*Department of Biology & Biochemistry, University of Houston,
Houston, TX 77204*

The question of multiple sequence alignment quality has received much attention from developers of alignment methods. Less forthcoming, however, are practical measures for quantifying alignment reliability in real life settings. Here, we present a method to identify and quantify uncertainties in multiple sequence alignments. The proposed method is based upon the observation that under any objective function or evolutionary model, some portions of reconstructed alignments are uniquely optimal, while other parts constitute an arbitrary choice from a set of co-optimal alternatives. The co-optimal portions of reconstructed alignments are, thus, at most half as reliable as the uniquely optimal portions. For pairwise alignments, this irreducible uncertainty can be quantified by the comparison of the high-road and low-road alignments, which form the co-optimality envelope for the two sequences. We extend this approach for the case of progressive multiple sequence alignment by forming a large set of equally likely co-optimal alignments that bracket the co-optimality space. This set can, then, be used to derive a series of local reliability measures for any candidate alignment. The resulting reliability measures can be used as predictors and classifiers of alignment errors. We report a simulation study that demonstrates the superior power of the proposed local reliability measures.

1. Introduction

Multiple sequence alignment (MSA) is the first step in comparative molecular biology. It is the foundation of a multitude of subsequent biological analyses, such as motif discovery, calculation of genetic distances, identification of homologous strings, phylogenetic reconstruction, identification of functional domains, three-dimensional structure prediction by homology modeling, functional genome annotation, and primer design [1]. The fundamental role of multiple sequence alignment is best demonstrated by noting that a paper describing a popular multiple-alignment reconstruction method, ClustalW [2], has been cited close to 25,000 times since its publication (i.e., an average of five times a day). Being a fundamental ingredient in a wide variety of analyses, the reliability and accuracy of multiple sequence alignment is an issue of utmost importance; analyses based on erroneously reconstructed alignments are bound to be severely handicapped [e.g., 3-9]. The question of multiple sequence alignment quality has received much attention from developers of alignment methods [10-15]. Unfortunately, practical measures for addressing alignment-quality issues in real life settings are sorely missing.

Multiple sequence alignment is frequently treated as a "black box"; the possibility that it may yield artifactual results is usually ignored. Moreover, in a manner reminiscent of basic laboratory disposables, the vast majority of multiple sequence alignments are produced robotically and discarded unthinkingly on the road to some other goal, such as a phylogenetic tree or a 3D structure. We speculate that more than 99% of all multiple sequence alignments that ultimately yield publishable results are never even looked at by a human being. Yet, when an occasional alignment is actually inspected, it is usually found wanting. Multiple sequence alignments are so notoriously inadequate, that the literature is littered with phrases such as "the alignment was subsequently corrected by hand" [e.g., 16-22]. Unfortunately, "hand correction" is neither objective nor reproducible, and as such we should strive to replace it by a scientifically legitimate method.

Errors in reconstructed alignments are typically attributed to the inadequacy of the evolutionary model and its parameters. Understandably, then, the recent proliferation of new reconstruction methods is mainly concerned with developing new optimality criteria and optimization heuristics. Unfortunately, the second source of reconstruction errors, i.e., the fact that the objective function usually possesses multiple optima even when the evolutionary model is adequate, is rarely addressed. Moreover, the full co-optimal solution set is often far too large to enumerate explicitly [23], and current MSA programs arbitrarily report only one of these co-optimal solutions. Reporting only one alternative from among the multitude of equally optimal or co-optimal alignments obscures the fact that the entire set of co-optimal alignments possesses valuable information; some portions of the alignments are uniquely optimal and are reproduced in every solution, while other portions differ among the solutions. Since the choice between such co-optimal alternatives is necessarily arbitrary, these portions of the alignments represent inherent irreducible uncertainty.

When dealing with pairwise alignments, we can capture this information by considering two extreme cases, termed the high-road and the low-road [24-25], which bracket the set of all co-optimal alignments. Alignment programs usually report either the high-road or the low-road as the final alignment. In such cases the other extreme alignment can be easily obtained by reversing the sequence residue order in the input [26]. Reversing the sequences amounts to inverting the direction of the two axes of the alignment dot matrix, thereby converting the high road to the low road and the low road to the high road. Columns that are identical in the two alignments define parts of the alignment where a single optimum of the objective function exists, whereas columns that differ between the two alignments define those portions of the alignments where there exist two or more co-optimal solutions.

A simple extension of this principle to the case of multiple sequence alignment is the “Heads or Tails” (HoT) methodology [26], where the original sequence set (the Heads set) is first reversed to create a second set (the Tails set). The two sequence sets are, subsequently, aligned independently, and the two resulting alignments are compared to produce a measure of their internal consistency. While the HoT method can be applied to any MSA reconstruction method, it produces only two alignments, and its statistical power is, therefore, limited.

Here we present a more powerful extension of the HoT methodology for the case of progressive multiple sequence alignment. Progressive alignment proceeds in a series of pairwise alignments of profiles, or sub-alignments, whose order is determined by an approximate guide tree. At each of these alignment steps, the resulting sub-alignment is an arbitrary choice from among many co-optimal alternative alignments. Our extension derives a large set of alternative MSAs that explores the co-optimality envelope of the several pairwise profile alignments that can be defined for a given guide-tree.

The set of alternative alignments is then analyzed to score specific elements of the alignments by their frequency of reproduction within the set. The reproduction scores can be applied to any candidate MSA to derive a series of local reliability measures that can identify and quantify uncertainties and errors in the reconstructed MSA.

2. Methods

2.1. *Construction of the co-optimality MSA set*

We implemented the derivation of the alignment set for ClustalW [2], which uses progressive alignment. Given the ClustalW approximate guide-tree for $\boldsymbol{N}$ sequences, we define the guide-tree alignment set, $^{gt}\boldsymbol{AS}$, as follows (Fig. 1):

For each of the $\boldsymbol{(N\text{-}3)}$ internal branches of the guide tree, partition the sequences into two subgroups (Fig. 1a). Construct two sub-alignments for each of the two sequence groups (Fig. 1b):

- Heads: The ClustalW alignment of the sequence subgroup.
- Tails: The ClustalW alignment of the reversed sequences, reversed to the original residue order.

Next, use the 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 for each internal branch (Fig. 1c).

The process is repeated for all internal branches of the guide-tree (Fig. 1d). All in all, then, $^{gt}\boldsymbol{AS}$ contains ***8·(N-3)*** alignments. These alignments differ from each other in two respects: (a) the partitioning of sequences and profiles to create the final MSA, and (b) the Heads or Tails selection of co-optimal sub-alignments and profile alignments. Any alignment in the set can be qualified as a *bona-fide* progressive alignment. Thus, the alignments in the guide-tree alignment set can be considered as equally likely alternatives that uniformly sample the co-optimality envelope.

2.2. *Local reliability measures for MSA*

Given a candidate reconstructed MSA, ***A***, we first construct the corresponding guide-tree alignment set, $^{gt}\boldsymbol{AS}$, and score the elements of ***A*** by their reproduction in $^{gt}\boldsymbol{AS}$ (Fig. 1e). For each pair of residues that are aligned as homologs in ***A***, we define our basic reliability measure, the *residue-pair reliability measure*, $^{pair}M^{c}_{i,j}$ (where ***c*** is the column index and ***i,j*** are the sequence indices), as the proportion of alignments in $^{gt}\boldsymbol{AS}$ that reproduce the pairing of the residue pair. The measure takes values within the interval [***0..1***], where ***1*** denotes total support. Averaging of the residue-pair support gives rise to a series of reliability measures:

- The *residue reliability* is the mean of the residue-pair reliability over all pairings involving the residue:

$$^{res}M^{c}_{i} = \overline{^{pair}M^{c}_{i,*}}$$

- The *column reliability* is the mean of the residue-pair reliability over all pairs in a column:

$$^{col}M^{c} = \overline{^{pair}M^{c}_{*,*}}$$

- The *alignment reliability* is the mean of the residue-pair reliability over all residues-pairs in the alignment:

$$^{ali}M = \overline{^{pair}M^{*}_{*,*}}$$

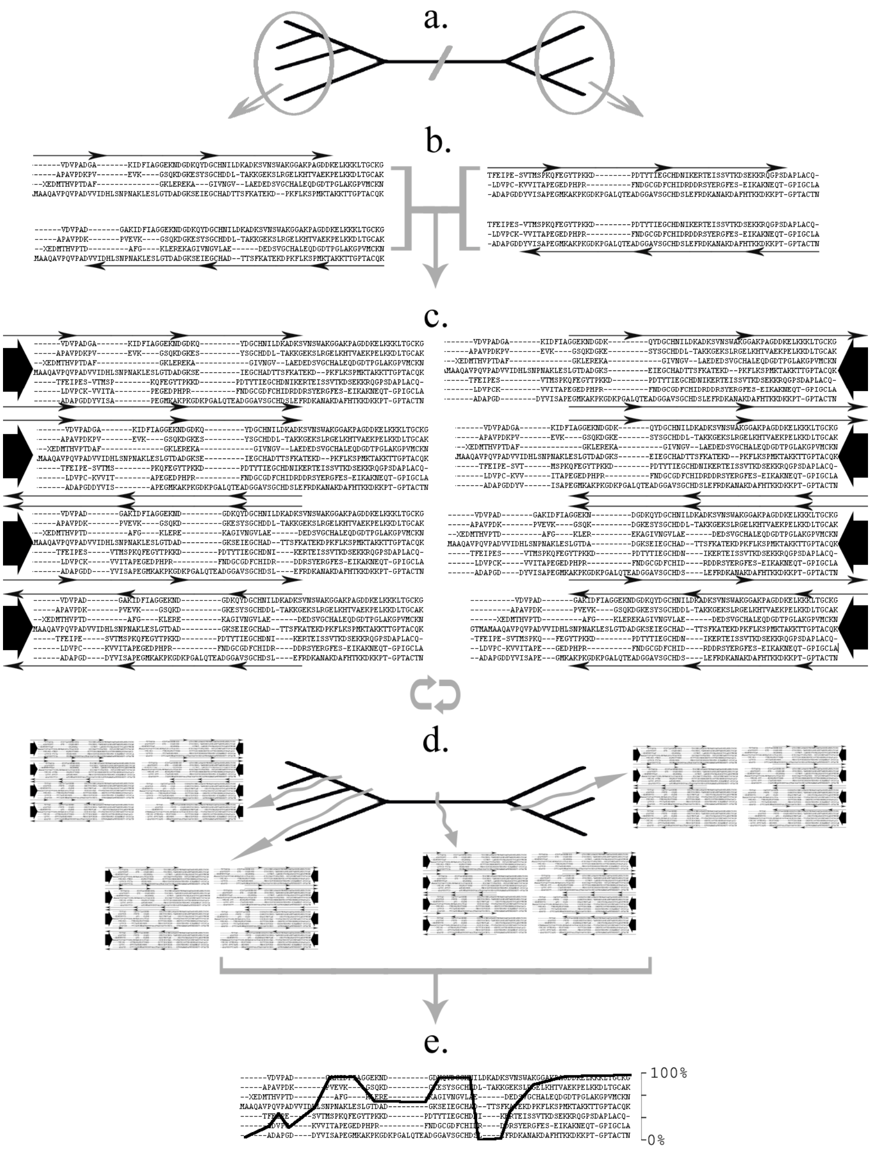

Figure 1: Construction of the guide-tree alignment set and the local reliability measures: (a) Use an internal branch of the guide tree to partition the sequences; (b) Align each subset in both heads and tails orientations, to produce 4 sub-alignments; (c) Align the four combinations of sub-alignments, in both heads and tails directions, for a total of 8 alignments; (d) Repeat a-c for each of the ***N-3*** internal branches, to produce ***8·(N-3)*** alternative alignments (***32*** for ***N=7***); (e) score elements of a candidate alignment by their frequency of reproduction (vertical axis) in the alignment set. (For more details, see text).

2.3. *Implementation*

Construction of the co-optimality MSA set and derivation of the local reliability measures were implemented in MATLAB scripts, available from the authors upon request.

3. Results

The local reliability measures can be used to identify and quantify errors in the reconstructed MSAs. We demonstrate their performances in a simulation study where MSAs reconstructed by ClustalW are compared to the true alignment from ROSE simulations [27]. We used 6400 datasets where the sequence evolution was simulated along a 16 taxa balanced depth-3 phylogeny, with an average branch length ranging from 0.02 to 0.30 substitutions per site, and an indel to substitution ratio of 0.015. The average sequence length was 500 nucleotides. Comparison of the true MSA to the ClustalW MSA yields rates of correct reconstruction at several resolution levels: residue-pairs, residue, column, and the entire alignment.

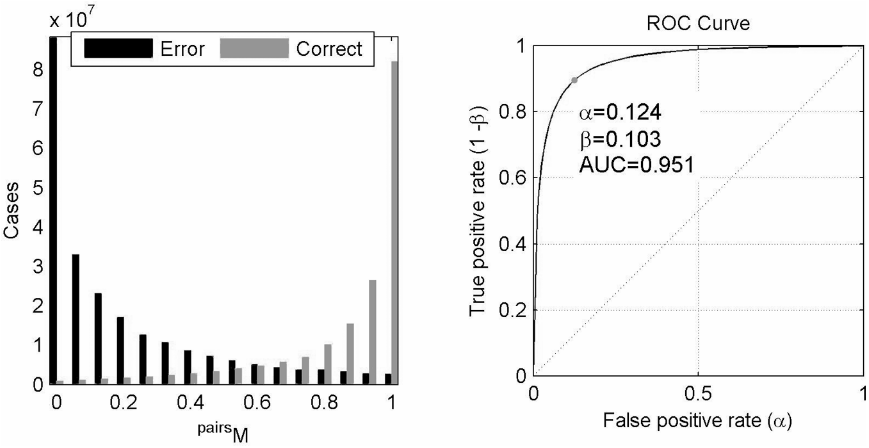

Figure 2: The residue-pairs reliability measure, $^{pairs}\boldsymbol{M}$, as a classifier of erroneous or correct residue-pairs in reconstructed MSAs. Histograms (left) presents the distributions of the two populations: **H0**:error (black) vs. **H1**:correct (gray). ROC curve (right) report the level of classification errors and the power of the classifier.

One use of the reliability measures is as binary classifiers of local MSA features as correct or erroneous. Figure 2 presents a receiver-operating characteristic (ROC) analysis [28] of $^{pairs}\boldsymbol{M}$ as a classifier of residue-pairs errors. Since the residue-pairs reconstruction rate, $^{pairs}\boldsymbol{R}$, is binary, the two populations, error (***H0***, black) or correct (***H1***, gray) reconstructions, are strictly defined. Our measure $^{pairs}\boldsymbol{M}$ is capable of separating the two populations, with a very high power (area under curve, AUC=0.95).

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. In figure 3 we compare the column reliability measure, colM to the entropy-based column quality measure reported by ClustalX, colQ [29], as classifiers of the true column errors. An ROC analysis reveals that colM separates the two populations, of erroneous and correct columns, better than colQ, with AUCs of ~0.94 and ~0.87, respectively.

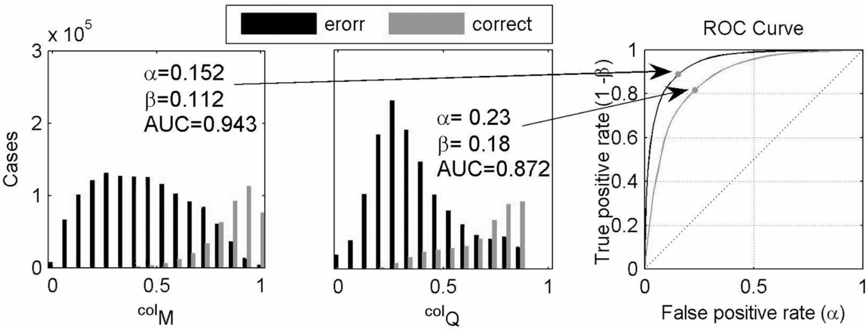

Figure 3: Comparison of two column reliability measures, $^{col}\boldsymbol{M}$ and $^{col}\boldsymbol{Q}$ as classifiers of erroneous or correct columns in reconstructed MSAs: Histograms (left) presents the different distributions of the two populations: **H0**:error (black) vs. **H1**:correct (gray). ROC curves (right) report the level of classification errors and the power of the classifier.

When interpreting the local reliability measures, $^{*}\boldsymbol{M}$**,** 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, $^{*}\boldsymbol{R}$; 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 $^{col}\boldsymbol{R}$, although significant, is only $\boldsymbol{r}$ = 0.66 (Fig. 4).

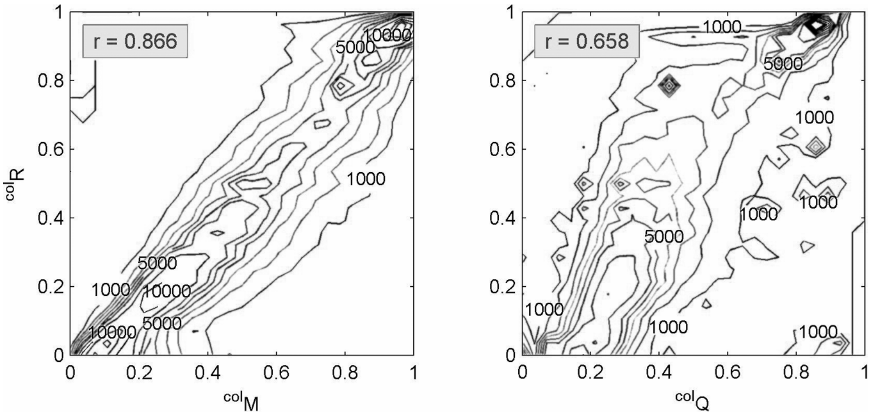

Figure 4: Comparison of two column quality measures, ***colM*** and ***colQ***, as estimates of the true reconstruction rates.

4. Discussion

The local reliability of reconstructed MSAs is usually viewed as related to the local divergence of the sequences. Thus, current local reliability measures are based on the column entropy or variation [e.g., 29]. While it is true that highly preserved segments of an MSA are more easily reconstructed by MSA algorithms, column entropies do not take into account the algorithmic sources of reconstruction errors. In contrast, our approach specifically addresses one common source of alignment errors, namely, the irreducible uncertainty stemming from the arbitrary choice from a set of co-optimal solutions. Hence its superiority to previous local quality measures.

The equivalence of co-optimal solutions is only one source of reconstruction errors. Two other sources of errors merit mention here: (a) the approximate nature of the guide-tree and the estimated evolutionary parameters, and (b) stochastic errors, where the true alignment is sub-optimal even when the objective function is exact [30]. It is interesting to note that although our reliability measures do not address these sources of errors directly, they do manage to correctly identify about 90% of the errors, while maintaining a low false positive rate.

The guide-tree alignment set does not exhaust the co-optimality space. In fact, it is not computationally feasible to enumerate the entire set of co-optimal alignments [23]. Even tracking every high-road low-road combination in a progressive alignment will yield a set whose size grows exponentially with the number of sequences. Our guide-tree alignment set of size ***8·(N-3)*** was designed as a practical compromise between computational feasibility and statistical power. Since the construction of the guide-tree already requires ***$O(N^2)$*** pairwise alignment steps, the additional ***$O(N^2)$*** steps required by our method amount to tripling the processing time.

Acknowledgments

This work was supported by NSF grant DBI-0543342.

References

1. L.J. Mullan, *BriefBioinform* **3**:303-305 (2002).
2. J.D. Thompson, D.G. Higgins, and T.J. Gibson, *Nucleic Acids Res* **22**: 4673-4680 (1994).
3. D.A. Morrison and J.T. Ellis, *Mol Biol Evol* **14**:428-441 (1997).
4. L. Florea, G. Hartzell, Z. Zhang, G.M. Rubin, and W. Miller *Genome Res.* **8**:967-974 (1998).
5. E.A. O'Brien and D.G. Higgins, *Bioinformatics* **14**:830-838 (1998).
6. R.E. Hickson, C. Simon, and S.W. Perrey, *Mol Biol Evol* **17**:530-539 (2000).
7. L. Jaroszewski L. Rychlewski, and A. Godzik, *Protein Science* **9**:1487–1496 (2000).
8. T.H. Ogden and M.S. Rosenberg, *Syst. Biol.* **55**:314–328 (2006).
9. S. Kumar and A. Filipski, *Genome Res.* **17**:127-135 (2007).
10. J.D. Thompson, F. Plewniak, and O. Poch, *Nucleic Acids Res* **27**:2682-2690 (1999).
11. A. Elofsson, *Proteins* **46**:330-339 (2002).
12. T. Lassmann and E.L. Sonnhammer, *FEBS Lett* **529**:126-130 (2002).
13. J.D. Thompson, P. Koehl, R. Ripp, and O. Poch, *Proteins* **61**:127-36 (2005).
14. Y. Chen and G.M. Crippen, *Structural bioinformatics* **22**:2087–2093 (2006).
15. P.A. Nuin, Z. Wang, and E.R. Tillier, *BMC Bioinformatics* **7**:471 (2006).
16. D. O'Callaghan, C. Cazevieille, A. Allardet-Servent, M.L. Boschiroli, G. Bourg, V. Foulongne, P. Frutos, Y. Kulakov, and M. Ramuz, *Mol. Microbiol.* **33**:1210–1220 (1999).
17. K. Kawasaki, S. Minoshima, and N. Shimizu, *J. Exp. Zool.* **288**:120-134 (2000).
18. C.M. Kullnig-Gradinger, G. Szakacs, and C.P. Kubicek, *Mycol. Res.* **106**:757-767 (2002).

19. J.L.M. Rodrigues, M.E. Silva-Stenico, J.E. Gomes, J.R.S. Lopes, and S.M. Tsai, *Applied and Environmental Microbiology* **69**:4249–4255 (2003).
20. S.B. Mohan, M. Schmid, M. Jetten AND J. Cole, *FEMS Microbiology Ecology* **49**:433–443 (2004).
21. E. Bapteste, R.L Charlebois, D. MacLeod and C. Brochier, *Genome Biology* **6**:R85 (2005).
22. M. Levisson, J. van der Oost and S.W.M. Kengen, *FEBS Journal* **274**:2832–2842 (2007).
23. D. Naor and D.L. Brutlag, *J. Comp. Biol.* **1**: 349-366 (1994).
24. D.J. States, and M.S. Boguski, *In M. Gribskov and J. Devereux, eds., Sequence Analysis Primer* pp:124–130, Oxford University Press, New York (1995).
25. T.G. Dewey, *J. Comp. Biol.* **8**: 177-190 (2001).
26. G. Landan and D. Graur, *Mol. Biol. Evol.* **24**:1380–1383 (2007).
27. J. Stoye, D. Evers, and F. Meyer, *Bioinformatics* **14**: 157-163 (1998).
28. M.H. Zweig and G. Campbell, *Clin. Chem.* **39**:561-577 (1993).
29. J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins, *Nucleic Acids Res.* **25**:4876-4882 (1997).
30. G. Landan, *In Zoology*, pp. 93. Tel Aviv University, Tel Aviv (2005).