Heads or Tails: A Simple Reliability Check for Multiple Sequence Alignments

Giddy Landan and Dan Graur

Department of Biology & Biochemistry, University of Houston

The question of multiple sequence alignment quality has received much attention from developers of alignment methods. Less forthcoming, however, are practical measures for addressing alignment quality issues in real life settings. Here, we present a simple methodology to help identify and quantify the uncertainties in multiple sequence alignments and their effects on subsequent analyses. The proposed methodology is based upon the a priori expectation that sequence alignment results should be independent of the orientation of the input sequences. Thus, for totally unambiguous cases, reversing residue order prior to alignment should yield an exact reversed alignment of that obtained by using the unreversed sequences. Such "ideal" alignments, however, are the exception in real life settings, and the two alignments, which we term the heads and tails alignments, are usually different to a greater or lesser degree. The degree of agreement or discrepancy between these two alignments may be used to assess the reliability of the sequence alignment. Furthermore, any alignment dependent sequence analysis protocol can be carried out separately for each of the two alignments, and the two sets of results may be compared with each other, providing us with valuable information regarding the robustness of the whole analytical process. The heads-or-tails (HoT) methodology can be easily implemented for any choice of alignment method and for any subsequent analytical protocol. We demonstrate the utility of HoT for phylogenetic reconstruction for the case of 130 sequences belonging to the chemoreceptor superfamily in *Drosophila melanogaster*, and by analysis of the BaliBASE alignment database. Surprisingly, Neighbor-Joining methods of phylogenetic reconstruction turned out to be less affected by alignment errors than maximum likelihood and Bayesian methods.

Introduction

Multiple sequence alignment is the most basic tool in the comparative study of molecular sequences. It is also the foundation of subsequent biological analyses, such as the derivation of sequence similarity measures, identification of homologous sites, phylogenetic reconstruction, identification of functional domains, 3-dimensional structure prediction, and primer design (Mullan 2002). The fundamental role of multiple sequence alignment is best demonstrated by noting that a paper describing a popular multiple-alignment reconstruction method, ClustalW (Thompson et al. 1994), is one of the most cited papers in biology. 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 (Morrison and Ellis 1997; O'Brien and Higgins 1998; Hickson et al, 2000; Ogden and Rosenberg 2006; Kumar and Filipski 2007). The question of multiple sequence alignment quality has received much attention from developers of alignment methods (Thompson et al. 1999, Elofsson 2002, Lassmann and Sonnhammer 2002, Thompson et al. 2005, Nuin et al. 2006). Less forthcoming, however, are practical measures for addressing alignment quality issues in real life settings.

Multiple sequence alignment is frequently taken for granted, and little thought is devoted to the possibility that this little "black box" may yield artifactual results. Moreover, in a manner reminiscent of basic laboratory disposables, the vast majority of multiple sequence alignments are produced automatically and discarded unthinkingly on the road to some other goal, such as a phylogenetic tree or a 3-dimensional structure. We conjecture that more than 99% of all multiple sequence alignments that are used to produce publishable results are never even seen by a human being. Yet, when a rare alignment is actually inspected by a researcher, 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 visual or manual inspection" (e.g., O'Callaghan et al. 1999; Kawasaki et al. 2000; Kullnig-Gradinger et al. 2002). Unfortunately, visual inspection is neither an objective nor a reproducible method, and as such we should strive to replace it by a scientifically acceptable tool.

Here, we present a simple methodology for the rapid identification and quantification of uncertainties in multiple sequence alignments and subsequent analyses.

Methods

Given a set of sequences (the heads set), we first create a second set containing the same sequences in reversed residue order (the tails set). The heads-or-tails (HoT) methodology entails the independent multiple alignment of the heads and tails sets and the comparison of the results obtained by using these two alignments in subsequent analytical protocols. For example, if the ultimate purpose is the inference of phylogenetic relationships, one may use the two alignments to construct two phylogenetic trees that can then be compared with each other. Typically, the first step of the analysis will be the reconstruction of multiple sequence alignments from the two sequences sets. For some subsequent analyses that are sensitive to direction (e.g., protein structure prediction), it may be important at this stage to reverse the tails alignment to the original residue order.

The degree of agreement between the two alignments may be assessed using two measures: 1) the fraction of identical alignment columns (termed column score by Thompson et al. [1999]) and 2) the proportion of residue pairs that are paired identically in the two alignments (equivalent to sum-of-pairs score of Thompson et al. [1999]). The column measure is, of course, the more conservative of the two. The effects of multiple sequence alignment uncertainties on subsequent downstream analyses may be evaluated

Key words: multiple sequence alignment, phylogenetic reconstruction, chemoreceptors.

E-mail: giddy.landan@gmail.com.

Mol. Biol. Evol. 24(6):1380–1383. 2007
doi:10.1093/molbev/msm060
Advance Access publication March 25, 2007

Table 1
Agreement between Heads and Tails Runs for Three Alignment Methods: Percentage of Identical Alignment Columns, Percentage of Identically Aligned Residue Pairs, and Percentage of Shared Phylogenetic Partitions in BioNJ (JTT) Trees Inferred from the Two Alignments

	ClustalW	MUSCLE	ProbCons
Columns	18.0%	8.7%	6.7%
Residue pairs	52.1%	53.7%	60.8%
Internal branches	64.6%	65.4%	59.1%

Table 2
Percentage of Shared Phylogenetic Partitions between Phylogenies Based on ClustalW Heads and Tails Runs for Three Phylogenetic Reconstruction Methods

	BioNJ (JTT)	ProML (JTT)	MrBayes (mixed)
BioNJ (JTT)	64.6% (HT)	59.8% (HH)	55.9% (HH)
ProML (JTT)	56.7% (TT)	52.8% (HT)	57.5% (HH)
MrBayes (mixed)	55.1% (TT)	50.4% (TT)	54.3% (HT)

NOTE.—Comparison of heads and tails trees for each method are reported on the diagonal (HT), values above the diagonal are among-method comparisons for the heads runs (HH), below the diagonal are among-method comparisons for the tails runs (TT).

by comparing the results obtained with the heads set with those obtained with the tails set.

Two PERL scripts for creating the tails set and for comparing the heads and tails alignments are available at http://nsm.uh.edu/~dgraur/scripts/HoT/. These scripts were kindly contributed by Dr Tal Dagan (Heinrich-Heine Universität Düsseldorf).

Results

We illustrate the type of results that can be obtained from a HoT analysis by reanalyzing a protein data set consisting of 130 homologous chemoreceptors (olfactory and gustatory receptors) from *Drosophila melanogaster* (Robertson et al. 2003). The average sequence length of these proteins is 404 amino acids. The sequences were aligned twice: once in the ordinary amino-to-carboxy direction (heads) and the other in the opposite carboxy-to-amino direction (tails). The degree of agreement between the two alignments was assessed using the columns and residue pairs measures (Thompson et al. 1999).

Table 1 summarizes the agreement between the heads and tails alignments for three alignment reconstruction programs: ClustalW (Thompson et al. 1994), MUSCLE (Edgar 2004), and ProbCons (Do et al. 2005). For the *D. melanogaster* chemoreceptor superfamily, all the three alignment methods fail to reproduce more than 80% of alignment columns. Even though the columns score is intuitive, it is admittedly too sensitive because it ignores columns with partial agreement. A more adequate measure is the residue pair agreement, where we find that only between 50% and 60% of the residue pairs are shared between the two alignments.

Given the large discrepancy between the heads and tails alignments, it is expected that any subsequent analysis that uses those alignments may also yield incongruent results between the two runs. We used the heads–tails alignment pairs to reconstruct phylogenies by the BioNJ method (Gascuel 1997) with the Jones-Taylor-Thornton distance measure (Jones et al. 1992) as implemented in the PHYLIP ProtDist program (Felsenstein 2005). The similarity of the resulting heads–tails phylogeny pairs can be quantified by the fraction of internal branches that are shared by the two trees (Felsenstein 2004). As expected the phylogeny pairs differ substantially from one another, failing to reproduce about 35% of the inferred phylogenetic partitions (table 1).

BioNJ being one of the simplest phylogenetic reconstruction methods, we repeated the ClustalW HoT analysis using two highly elaborate (and computationally intensive) phylogenetic reconstruction methods: maximum likelihood, as implemented in the PHYLIP ProML program (Felsenstein 2005), and MrBayes (Ronquist and Huelsenbeck 2003). The internal agreement between the heads and tails phylogenies inferred by these methods is even smaller than that achieved by BioNJ, with about 45% of the inferred phylogenetic partitions failing to reproduce (table 2). Moreover, there are fewer shared phylogenetic partitions between the heads and tails trees inferred by MrBayes and ProML than there are between trees inferred by the different methods but based on the same alignment orientation (table 2).

It may be suspected that the internal branches that conflict between heads and tails phylogenies are the relatively poorly resolved parts of the phylogeny. This is indeed the case for the BioNJ trees, where only 1 out of the 90 conflicting phylogenetic partitions is supported at the 70% bootstrap proportion level. However, of the 116 conflicting partitions in the MrBayes trees, 21 have posterior probabilities greater than 0.7, whereas of the 120 conflicting branches in the ProML trees, 74 are significant at the 0.05 level.

The discrepancy between heads and tails alignments is expected to increase with increasing divergence of the sequences under considerations. To demonstrate this dependency, we conducted a HoT analysis of the BaliBASE alignment database (Thompson et al. 2005). The analysis consisted of ClustalW alignments followed by BioNJ (JTT) phylogenetic reconstructions of 196 sequence sets, where the number of sequences ranged from 5 to 142 and the alignment length from 87 to 8,052. Figure 1 presents the HoT agreement fraction as a function of the total tree length, which we use as a measure of sequence divergence. As expected, the agreement deteriorates with increasing sequence divergence for both the alignments (fig. 1*a* and *b*) and the inferred phylogeny (fig. 1*c*). Moreover, even at relatively low degrees of sequence divergence, the range of agreement values often spans quite low values, and high discrepancy outliers are abundant.

Discussion

DNA and proteins are polymers that exhibit an operational directionality: amino to carboxy in proteins and 5′ to 3′ in nucleic acids. This directionality is so entrenched in our perception of these molecules, that we unthinkingly process every sequence from left to right, regardless of

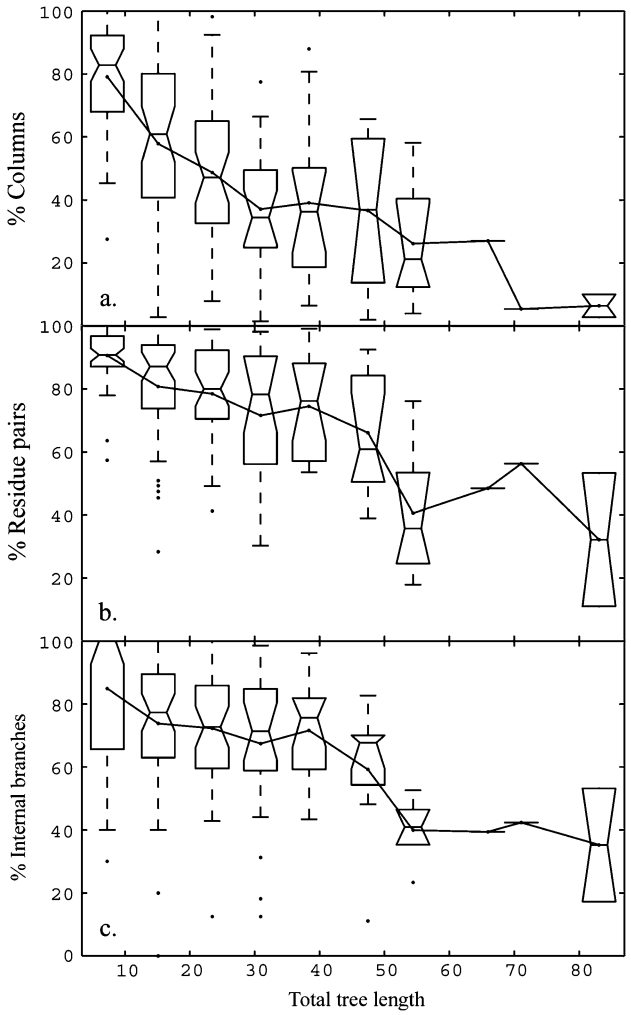

FIG. 1.—Agreement between heads and tails alignments and phylogenies for 196 sequence sets from BaliBASE, as a function of total length of the inferred trees. (*a*) Percentage of identical alignment columns, (*b*) Percentage of identically aligned residue pairs, and (*c*) Percentage of shared phylogenetic partitions in BioNJ (JTT) trees inferred from the two alignments. Box plots summarize medians, quartiles and range; lines pass through the mean values.

the purpose of the analysis. The essence of our "reliability check" is to transcend this directionality and align the sequences ambidextrously.

Under any objective function, there may be many co-optimal solutions to any sequence alignment problem. Among these co-optimal alignments, two extreme cases, termed the high road and the low road (States and Boguski 1992; Dewey 2001), bracket the set of all cooptimal alignments. Pairwise alignment programs usually report either the high road or the low road as the final alignment. In such cases, the heads and tails runs amount to retrieving both the high- and low-road alignments. Columns that are identical in the two alignments define parts of the alignment where only 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 and frequently more cooptimal solutions. Arbitrarily reporting only one of these multiple alternatives obscures the fact that co-optimal portions of the alignment are at most half as reliable as the strictly optimal portions.

The information gained from a HoT analysis is not restricted to the alignment part of the analysis only. For example, our heads and tails analysis also sheds some new light on the merits of different approaches to phylogenetic reconstruction. In recent years, simple distance methods, such as Neighbor-Joining (Saitou and Nei 1987) and BioNJ (Gascuel 1997), have lost their popularity, and more "sophisticated" and computationally intensive methods, such as maximum likelihood and Bayesian inference, are preferred. In our example, we find the simple distance methods to be more robust to the uncertainties in the alignment input. Perhaps, the more elaborate methods, by their fastidious examination of the minutiae of alignment, merely engage in deriving extremely accurate estimates of the wrong parameters. Perhaps, also, the extreme reduction in the amount of information entailed by the compaction of the character state alignment into a handful of pairwise distances serves as a smoothing out mechanism that filters out some of the noise inherent in the data.

The proposed HoT methodology is widely applicable and can be easily implemented for any sequence analysis procedure involving sequence alignment, regardless of choice of alignment program or other computational building blocks. In its most general terms, it consists of two steps bracketing the entire analysis protocol: before the analysis, produce a second sequence data set consisting of the reversed sequences, and after the analysis of both sets, compare and examine the results. HoT should be used as a reality check for internal consistency.

Acknowledgments

We thank Tal Dagan for kindly providing the PERL scripts for the HoT analysis. This work was supported by National Science Foundation grant DBI-0543342.

Literature Cited

Dewey TG. 2001. A sequence alignment algorithm with an arbitrary gap penalty function. J Comp Biol. 8:177–190.

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. 15:330–340.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Elofsson A. 2002. A study on protein sequence alignment quality. Proteins. 46:330–339.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Felsenstein J. 2005. PHYLIP (phylogeny inference Package). Version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

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.

Hickson RE, Simon C, Perrey SW. 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.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comp Appl Biosci. 8:275–282.
Kawasaki K, Minoshima S, Shimizu N. 2000. Propagation and maintenance of the 119 human immunoglobulin V genes and pseudogenes during evolution. J Exp Zool. 288: 120–134.
Kullnig-Gradinger CM, Szakacs G, Kubicek CP. 2002. Phylogeny and evolution of the genus *Trichoderma*: a multigene approach. Mycol Res. 106:757–767.
Kumar S, Filipski A. 2007. Multiple sequence alignment: in pursuit of homologous DNA positions. Genome Res. 17:127–135.
Lassmann T, Sonnhammer EL. 2002. Quality assessment of multiple alignment programs. FEBS Lett. 529:126–130.
Morrison DA, Ellis JT. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. Mol Biol Evol. 14:428–441.
Mullan LJ. 2002. Multiple sequence alignment—the gateway to further analysis. Brief Bioinformatics. 3:303–305.
Nuin PA, Wang Z, Tillier ER. 2006. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics. 7:471.
O'Brien EA, Higgins DG. 1998. Empirical estimation of the reliability of ribosomal RNA alignments. Bioinformatics. 14:830–838.
O'Callaghan D, Cazevieille C, Allardet-Servent A, Boschiroli ML, Bourg G, Foulongne V, Frutos P, Kulakov Y, Ramuz M. 1999. A homologue of the *Agrobacterium tumefaciens* VirB and *Bordetella pertussis* Ptl type IV secretion systems is essential for intracellular survival of *Brucella suis*. Mol Microbiol. 33:1210–1220.
Ogden TH, Rosenberg MS. 2006. Multiple sequence alignment accuracy and phylogenetic inference. Syst Biol. 55:314–328.
Robertson HM, Warr CG, Carlson JR. 2003. Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. Proc Natl Acad Sci USA. 100:14537–14542.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.
States DJ, Boguski MS. 1992. Similarity and homology. In: Gribskov M, Devereux J, editors. Sequence analysis primer. New York: Oxford University Press. pp. 124–130.
Thompson JD, Higgins DG, Gibson TJ. 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 JD, Koehl P, Ripp R, Poch O. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins. 61:127–136.
Thompson JD, Plewniak F, Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 27:2682–2690.

William Martin, Associate Editor

Accepted March 15, 2007