Multiple sequence alignment is an essential tool in many areas of biological research, and the accuracy of an alignment can strongly affect the accuracy of a downstream application such as phylogenetic analysis, identification of functional motifs, or polymerase chain reaction primer design. The heads or tails (HoT) method (Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.) assesses the consistency of an alignment by comparing the alignment of a set of sequences with the alignment of the same set of sequences written in reverse order. This study shows that HoT scores and the alignment accuracies are positively correlated, so alignments with higher HoT scores are preferable. However, HoT scores are overestimates of alignment accuracy in general, with the extent of overestimation depending on the method used for multiple sequence alignment.

Introduction

Multiple sequence alignment is an essential tool in many areas of biological research including tertiary protein structure prediction, identification of functional motifs and domains, polymerase chain reaction primer design, reconstruction of ancient ancestral protein sequences, and phylogenetic inference (Mullan 2002; Ogden and Rosenberg 2006). Multiple sequence alignment is a way to write a set of homologous sequences of unequal length so that homologous nucleotides or amino acids are written one above the other in a column, that is accomplished by inferring the presence of gaps, corresponding to historical insertions and deletions. Multiple sequence alignments are estimates of the positions of those gaps, not truth, and like all estimates may be more or less accurate. The accuracy of multiple sequence alignments is of considerable concern to downstream analyses, such as phylogenetic analyses, because the accuracies of the downstream analyses depend strongly on the accuracy of the alignment (Hall 2005; Rosenberg 2005; Kumar and Filipski 2006; Ogden and Rosenberg 2006). A variety of studies have compared the accuracies of different multiple sequences alignment programs based either on “gold standard” structural alignments such as the BAliBASE collection (Thompson et al. 1999a, 2005; Lassmann and Sonnhammer 2002, 2005; Do et al. 2005; Katoh et al. 2005; Carroll et al. 2007) or on simulated data sets (Katoh et al. 2002; Nuin et al. 2006). Those studies, however, do not allow the individual investigator to judge whether a particular alignment is sufficiently accurate to justify trusting the results of a downstream analysis.

Some studies offer a degree of guidance with respect to multiple alignment accuracy. For instance, when aligning protein sequence, when the average amino acid identity is >25% the alignment accuracy is on average >80%, and when amino acid identity is <20% alignment accuracy is <50% (Thompson et al. 1999a). For DNA alignments, when aligned sequence identity is >60% accuracy is >80%, but accuracy drops below about 50% when identity is below 50% (Ogden and Rosenberg 2006). Those guidelines, while better than nothing, are not terribly helpful because they are based on sequence identities that are averages of closely related sequences with distantly related sequences. They also fail to take into account gap density.

In general, the “gapper” the alignment the lower is the accuracy (Nuin et al. 2006). Most alignment programs allow users to set gap penalties and other parameters, and in some situations, accuracy can be improved by modifying the default settings of those parameters. However, I would speculate that >99% of alignments are performed using the default parameter settings simply because there is no easy way to assess whether modifying the parameters makes the alignment more or less accurate.

To help remedy the situation, Landan and Graur (2007) proposed the heads or tails (HoT) method to check the reliability of multiple alignments. Landan and Graur use the term “reliability,” but the HoT method actually measures how consistent a method is in the face of alternate data sets and not how accurate the method is. In the classic dartboard analogy, the HoT approach is a way of measuring whether a method managed to place all the darts close together or spreads them around the board. It does not measure how close they are to the bull’s eye. Accordingly, I shall simply refer to the HoT score as a measure of HoT consistency.

According to the HoT-score method, the unaligned sequences are first reversed to produce a “tails” data set, then alignments are generated from the original (heads) and the tails data sets. The heads and tails alignments are compared by 2 criteria, columns and residue pairs. The columns comparison shows the fraction of columns that are identical between the 2 alignments and is the more conservative measure. The residue pairs measure shows the proportion of residue pairs that are identical in the 2 alignments. Although Landan and Graur provide measures of consistency, we have no idea how those measures correspond to measures of accuracy.

Accuracy is most frequently measured by comparing an estimated alignment with a protein structure–based alignment such as a BAliBASE alignment. Although they are often treated as gold standards, structure-based alignments are not “true” alignments, they are simply better estimates of the true historical relationships among homologous sites. Measuring accuracy requires knowing the true alignment, which is only possible with simulated data. EvolveAGene 3 (Hall 2008), a coding sequence simulation program that incorporates indels, provides true alignments that can be used to measure the accuracy of estimated DNA and protein sequence alignments. That accuracy can then be compared with the HoT score of the alignment to assess how well the HoT score reflects accuracy.
In this study I used simulated data sets to obtain known, true, alignments of both protein and DNA sequences. The unaligned sequences were then aligned to produce a heads alignment. The heads alignment was compared with the true alignment to measure the “accuracy” of the alignment. The sequences were then reversed and aligned to produce a tails alignment. The heads and tails alignments were compared by the residue pairs score to measure the HoT score. I then examined the relationships between accuracy and HoT scores over a realistic range of alignment conditions. I also examined whether the HoT-score method can be used to accurately assess the effects of changes in alignment program parameter settings or to decide which alignment program performs better on a particular data set.

Methods
Simulations

Sequence evolution was simulated by EvolveAGene 3 (Hall 2008) based on 20 data sets from the BAliBASE collection of structure-based alignments (Thompson et al. 1999b, 2005; Bahr et al. 2001), with 5 independent simulations based on each of the BAliBASE data sets. Each simulation was based on the topology and mean branch length of a maximum likelihood tree of the BAliBASE data sets. Each simulation was initiated from the first sequence in the BAliBASE data set. The simulated data sets are runs 1–5 in the simulation sets in table 3 of Hall (2008). The number of sequences in the simulated data sets ranged from 10 to 142, the mean number of indels per base substitution ranged from 0.0026 to 0.0046, the tree lengths ranged from 8.9 to 72.9 substitutions per site, and the pairwise percent amino acid identities ranged from 11.2% to 83.8%. EvolveAGene 3 output includes the true DNA and protein sequence alignments and the unaligned DNA and protein sequences. The unaligned sequences were termed the “heads” sequences.

Alignments

Protein sequences were aligned by ClustalW 1.83 (Thompson et al. 1994) and by ProbCons (Do et al. 2005), whereas DNA sequences were aligned by ClustalW 1.83 and by ProbConsRNA beta http://probcons.stanford.edu/download.html. In the text, ProbConsRNA beta is simply referred to as ProbCons. ProbCons and ProbConsRNA beta alignments used the default settings, as did ClustalW DNA alignments. Three different sets of parameter settings were used to align protein sequences by ClustalW: default settings (pairwise alignment gap opening = 10.0, pairwise alignment gap extension = 0.1, multiple alignment gap opening = 10.0, multiple alignment gap extension = 0.2, and minimum gap separation distance = 4); BioX http://www.macupdate.com/info.php/id/20630 implementation of ClustalW default settings (same as ClustalW settings except that minimum gap separation distance = 8); and improved settings (pairwise alignment gap opening = 10.0, pairwise alignment gap extension = 0.1, multiple alignment gap opening = 3.0, multiple alignment gap extension = 1.8, and minimum gap separation distance = 1). The improved settings are similar to those that I have previously recommended (Hall 2007).

HoT Score

The Perl script HoT_reverse (Landan and Graur 2007) was used to reverse the heads sequences to create the tails sequences. The Perl script HoT_compare (Landan and Graur 2007) was used to compare the heads alignment with the tails alignment, and the HoT score is given as the total residue pairs score reported by HoT_compare. The residue pairs score is the proportion of residue pairs that are paired identically in the heads and tails alignments and is the equivalent of the sum-of-pairs score of Thompson et al. (1999a; Landan and Graur 2007).

Alignment Accuracy

Alignment accuracy was measured by comparison of the true alignment with the heads alignment and is the proportion of residue pairs that are paired identically in the 2 alignments, equivalent to the residue pairs score of HoT_compare. Accuracy was determined by the Perl script CompareAlign, which is available upon request to barryhall@zeninternet.com. HoT scores and alignment accuracies of the 100 data sets are in supplementary table S1 (Supplementary Material online).

Results
Using HoT Scores to Set Alignment Parameters

Figure 1A is a probability plot of the HoT scores. The further to the right is the curve, the more consistent is the method. HoT-score consistencies are ClustalW BioX < ClustalW default < ClustalW improved < ProbCons. If the HoT-score method is useful for choosing alignment parameters or alignment methods, then the order of method with respect to accuracy should be the same as that with respect to HoT-score consistency. Figure 1B shows that that is the case; however, HoT scores are generally higher than is accuracy.

Another approach to judging the value of HoT scores to choosing alignment parameters or methods is to consider, for each data set, the HoT score or accuracy of one method relative to another method; for example, the accuracy of ClustalW preferred settings relative to ClustalW default settings. If the preferred settings relative to the default settings are close to 1.0, it probably does not matter which settings are used. I have arbitrarily chosen 0.95 and 1.05 as the range that suggests insignificant differences in HoT scores or accuracy. Table 1 shows that 100% of ClustalW BioX alignments have significantly lower HoT scores than ClustalW default alignments, and none have higher HoT scores, whereas 96% are less accurate and none more accurate. In general, determination of HoT scores would result in choosing the more accurate method.

The HoT-score approach would err if it led to choosing the significantly more HoT-consistent method when that method was, in fact, significantly less accurate. Table 1 shows that in deciding between the ClustalW BioX and ClustalW default settings, the HoT-score approach would lead to the wrong choice only 12% of the time; in choosing between the ClustalW default and ClustalW improved method, it would err only 3% of the time; and in choosing
between ProbCons and ClustalW improved, it would err only 3% of the time. In deciding between ProbCons and ClustalW to align DNA sequences, HoT scores would err only 2% of the time.

Using HoT Scores to Assess Alignment Accuracy

Comparison of figure 1A with figure 1B suggests that overall HoT scores are a bit higher than accuracy. To determine how well HoT scores correspond to accuracy, HoT scores were plotted versus accuracy for ClustalW improved and ProbCons protein alignments and for ClustalW and ProbCons DNA alignments (fig. 2). In figure 2, points that are below and to the right of the diagonal line have higher HoT scores than accuracy, whereas those above and to the left of the diagonal have lower HoT scores than accuracy.

If HoT scores precisely reflected accuracy the points would all fall along the diagonal lines in figure 2. Instead, the majority of the points fall to the right of and below the diagonal, that is, in general HoT scores overestimate accuracy. For the protein alignments (table 2), the ProbCons method is 0.06 more accurate than the ClustalW improved method, and ProbCons HoT scores overestimate accuracy by only 0.03, compared with ClustalW HoT scores over estimating accuracy by 0.06. Clearly, for protein alignments, ProbCons is the method of choice. For the DNA alignments, ProbCons is also the more accurate method, but ProbCons HoT scores overestimate accuracy by a whopping 0.12, whereas the less accurate ClustalW HoT scores only overestimate accuracy by 0.03.

Landan and Graur used total tree length as a measure of sequence divergence and noted that HoT scores decrease as tree length increases (Landan and Graur 2007). Figure 3 shows that accuracy also decreases as tree length increases, but the correlations are not strong. Indeed, for the ProbCons method of protein sequence alignment, accuracies of trees whose lengths ranged from 60 to 75 substitutions per site were higher than accuracies of trees whose lengths ranged from 40 to 50 substitutions per site.

When alignment accuracy is >60%, phylogenetic trees are, on average, about as accurate as when estimated from the true alignments (Ogden and Rosenberg 2006). Although I am unaware of similar studies of other downstream applications of alignments, it is reasonable that we should be most concerned with those difficult to align data sets whose accuracies are <60%. By the <60% accurate criterion 18 of the ClustalW improved, 11 of the ProbCons protein alignments, 23 of the ClustalW DNA, and 16 of the ProbCons DNA alignments were difficult to align. HoT scores failed to identify 10 of the 18 ClustalW difficult protein alignments as “difficult,” correctly identified all the ProbCons difficult protein alignments, but misidentified 2 ProbCons acceptable alignments as being difficult. HoT scores failed to identify only 1 of the 21 difficult

Table 1

Relative HoT Reliabilities and Accuracies of Alignment Methods

<table>
<thead>
<tr>
<th>Comparison</th>
<th>HoT Scores</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Significantly Worse</td>
<td>Significantly Better</td>
</tr>
<tr>
<td>Protein alignments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ClustalW BioX relative to ClustalW default</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>ClustalW improved relative to ClustalW default</td>
<td>5</td>
<td>58</td>
</tr>
<tr>
<td>ProbCons relative to ClustalW improved</td>
<td>9</td>
<td>34</td>
</tr>
<tr>
<td>DNA alignments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ProbCons relative to ClustalW</td>
<td>0</td>
<td>58</td>
</tr>
</tbody>
</table>

<sup>a</sup> Errors = the fraction of comparisons in which HoT reliability would lead to choosing the less accurate method.
ClustalW DNA alignments but failed to identify 9 of the 16 difficult ProbCons DNA alignment.

To use the dartboard analogy, HoT scores do not show that as the darts cluster more tightly they tend to cluster around the bull’s eye.

Discussion

Because multiple sequence alignment is such a basic tool in comparisons of molecular sequence, it is very important to be able to easily evaluate the accuracy of alignments. As a tool for judging the accuracy of alignments, the value of the HoT-score method varies according to the alignment method that is used (table 2). For the 2 protein alignment methods that are considered here, HoT scores are pretty good for the ProbCons method and fair for the ClustalW improved method. For the DNA alignment methods, there is conflict between accuracy of the method and how well HoT scores reflect that accuracy. ProbCons is more accurate, but HoT scores greatly overestimate that accuracy. The method dependence of the degree to which HoT scores reflect accuracy suggests that HoT scores should not be used as an estimator of absolute accuracy. However, the value of Hot scores as an indicator of “relative” accuracy is an entirely different matter.

There are a variety of alignment methods and programs available, and for many of those programs, there are a variety of parameters that can be set. Although there are numerous studies comparing the accuracies of those programs on test data sets, it is often not obvious which method is best for any particular real data set and it is extraordinarily difficult to judge how adjusting alignment parameters affects alignment accuracy. Those judgments do not require estimating absolute accuracy, but they do require reliable estimates of relative accuracy. HoT scores

![Figure 2](http://mbe.oxfordjournals.org/)  
FIG. 2.—HoT scores versus accuracy. (A) and (B), protein alignments. (C) and (D), DNA alignments. Dashed lines are least squares linear regression lines.

Table 2

<table>
<thead>
<tr>
<th>Mean Accuracy and HoT Scores</th>
<th>Protein Alignments</th>
<th>DNA Alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ClustalW Improved</td>
<td>ProbCons</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.81 ± 0.02</td>
<td>0.87 ± 0.02</td>
</tr>
<tr>
<td>HoT scores</td>
<td>0.87 ± 0.02</td>
<td>0.90 ± 0.02</td>
</tr>
</tbody>
</table>

*a* Mean ± standard error.
are an effective estimator of relative accuracy for both protein and DNA alignments (table 1) and would rarely lead to choosing the wrong method or parameter settings. It is important to stress that it is not the purpose of this study to suggest that particular parameter settings or methods are universally superior. Instead, it is to encourage readers to apply the HoT-score method to choose among methods and parameter settings for their own, unique data sets.

The HoT consistency measure is both clever and easy to apply and it serves a valuable purpose as a measure of relative accuracy for choosing among alignment methods or parameter settings, but it can be misleading as an estimator of absolute alignment accuracy. Although it is very much needed, we still lack a good way to estimate accuracies of real multiple sequence alignments.

Supplementary Material

Supplementary table S1 is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

I thank Reviewer #1 for his insight and considerable effort in raising issues with respect to the validity of using the HoT method to choose among parameter setting for alignment programs. I thank Reviewer #2 for the dartboard analogy.

Literature Cited


Sudhir Kumar, Associate Editor

Accepted April 27, 2008