The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs

Sam Griffiths-Jones* and Alex Bateman

The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

Received on November 15, 2001; revised on February 18, 2002; accepted on March 14, 2002

ABSTRACT

Motivation: The best quality multiple sequence alignments are generally considered to derive from structural superposition. However, no previous work has studied the relative performance of profile hidden Markov models (HMMs) derived from such alignments. Therefore several alignment methods have been used to generate multiple sequence alignments from 348 structurally aligned families in the HOMSTRAD database. The performance of profile HMMs derived from the structural and sequence-based alignments has been assessed for homologue detection.

Results: The best alignment methods studied here correctly align nearly 80% of residues with respect to structure alignments. Alignment quality and model sensitivity are found to be dependent on average number, length, and identity of sequences in the alignment. The striking conclusion is that, although structural data may improve the quality of multiple sequence alignments, this does not add to the ability of the derived profile HMMs to find sequence homologues.

Supplementary information: A list of HOMSTRAD families used in this study and the corresponding Pfam families is available at http://www.sanger.ac.uk/Users/sgj/alignments/map.html.

Contact: sgj@sanger.ac.uk

INTRODUCTION

Profile hidden Markov models (HMMs) are powerful and popular computational tools for building a mathematical model of a multiple sequence alignment. This model can then be used to detect related sequences. It has been shown that these methods significantly outperform methods such as BLAST (Altschul et al., 1990) and FASTA (Pearson and Lipman, 1988), which rely purely on pairwise comparisons (Park et al., 1998).

A number of methods have been proposed to address the problem of multiple sequence alignment, and these fall broadly into two categories—progressive and iterative approaches. The most popular method in the past has been a progressive approach, where the multiple sequence alignment is built up by aligning the closest sequences first and then adding in the more distant ones. Algorithms which use this method include CLUSTALW (Thompson et al., 1994), MULTALIGN (Barton and Sternberg, 1987) and MULTAL (Taylor, 1998). These methods differ mainly in the determination of the order of alignment which generally involves the building of a guide tree. The T-COFFEE algorithm also uses a progressive alignment approach, but first pre-computes all pairwise alignments between the sequences (Notredame et al., 2000). Each stage of the progressive alignment then involves comparison of the overall alignment with each of the aligned pairs. The alternative is an iterative approach to refine and improve an initial multiple alignment. The program HMMT (Eddy, 1995) uses a simulated annealing approach to maximize the probability that an HMM represents a sequence alignment in an iterative fashion. Other iterative algorithms include PRRP (Gotoh, 1996) and SAGA (Notredame and Higgins, 1996). The above methods are mainly based on a global alignment approach—alternative local approaches are implemented by a number of algorithms including PIMA (Smith and Smith, 1992) and DIALIGN (Morgenstern et al., 1998).

Several previous studies have compared the ability of a number of multiple sequence alignment programs to align sequences of varying similarity and length (Thompson et al., 1999b). The BAliBASE database of protein sequences has been constructed to benchmark alignment algorithms with regard to such sequence parameters (Thompson et al., 1999a). Studies with BAliBASE suggest amongst other general findings that the best performance is often obtained by iterative strategies, and global methods generally outperform local alignment methods (Thompson et al., 1999b).

Independent experimentally-based multiple sequence alignments can be derived from the superposition of three-dimensional structures. The HOMSTRAD database (Mizuguchi et al., 1998) is a collection of such structurally
aligned families. For many purposes, such as identifying key functional residues, these alignments can be regarded as optimal. However, the current study makes use of these available data to compare the performance of the sequence alignment methods discussed above with respect specifically to homology detection using profile HMMs. To this end, we compare the sensitivity and specificity of profile HMMs constructed using structural alignments with those made with computational alignment methods that rely only on sequence information.

Structure alignments are not the only way to incorporate structural information into homologue detection procedures. An alternative approach is highlighted by the FUGUE algorithm for sequence–structure homology recognition (Shi et al., 2001). This method incorporates structural environment-specific substitution tables into a sequence alignment algorithm. This is reported to give significantly improved alignment accuracy over purely sequence-based methods. The same report also describes a large improvement in homology recognition on the incorporation of such structural information.

METHODS
The 1 August 2001 freeze of HOMSTRAD (Mizuguchi et al., 1998) was obtained from http://www-cryst.bioc.cam.ac.uk/homstrad/. The sequences of each HOMSTRAD family were searched against Pfam 6.5 (http://www.sanger.ac.uk/Software/Pfam/; Bateman et al., 2002) using the HMMER2 software (http://hmmer.wustl.edu/) on a hardware accelerated Decypher box to build up a mapping of HOMSTRAD and Pfam families. The 585 multi-member HOMSTRAD families were then filtered to give a list of 348 families which each correspond with only one Pfam family.

The HOMSTRAD families were realigned using two multiple sequence alignment methods with default parameters—CLUSTALW (Thompson et al., 1994) and T-COFFEE (Notredame et al., 2000). These methods are generally recognized as being amongst the best multiple sequence alignment programs (Thompson et al., 1999b). For comparison, three sets of intermediate quality alignments were generated by random introduction of gap characters into the HOMSTRAD alignment. The three sets were produced with different probability (p) of gap introduction at each position in the alignment. These sets are referred to as Gap1 (p = 0.01), Gap5 (p = 0.05) and Gap80 (p = 0.8) in the following analysis. These intermediate quality alignments are not intended to be biologically meaningful, rather they serve to illustrate the exaggerated effects of poor alignments on profile HMM performance.

COMPALIGN v1.5 (S.R.Eddy, unpublished) was used to assess similarity of each of the sequence-based methods to the structure-based alignment from the HOMSTRAD database. This program calculates the fraction of pairs of aligned residues which are identical between two alignments. We then built profile HMMs from each of these differently aligned families and used them to search the Pfamseq protein sequence database (Bateman et al., 2002), built from SWISS-PROT 39 and TrEMBL 14; (Bairoch and Apweiler, 2000). A blanket e-value threshold of 0.01 was chosen and family membership compared with the corresponding Pfam family. We have assumed that the Pfam database provides a list of correct matches for each HOMSTRAD family. There are a few instances where the model built from the HOMSTRAD family finds members not present in the corresponding Pfam family. This is likely to be due to incomplete coverage of these families in Pfam. However, the removal of these examples from the analysis does not significantly alter the values and findings presented here.

Distributions of sequence number, identity and length were calculated and used to choose the following categories of alignments based on maintaining significant numbers within each group: sequence number—2, 3, 4, 5, 6, 7–10, 11–41; average sequence identity within an alignment—<20%, 20–30%, 30–40%, 40–50%, 50–60%, 60–70%, 70–80%, >80%; average sequence length—<100, 100–200, 200–300, 300–400, 400–500, >500 amino acids.

Model sensitivity and specificity were calculated with respect to Pfam family membership, and each alignment-derived model normalized with respect to the performance of the HOMSTRAD-derived model. Statistical analyses of the data (pairwise t-test) were carried out to determine the significance of sensitivity differences between each sequence-based model and the structure-based model.

RESULTS AND DISCUSSION
Alignment similarities
The proportions of pairs of aligned residues which are identical between alignments are shown in Table 1. If we assume the structural alignment data to be the optimal alignment in each case, the data show that T-COFFEE and CLUSTALW both perform well with nearly 80% of residues aligned correctly. T-COFFEE has a slight advantage in alignment quality. The lower quality alignment sets Gap1, Gap5 and Gap80 have on average only 40%, 19% and 8% of residues correctly aligned. The Gap80 set of alignments can be considered as essentially unaligned.

These averaged results may hide specific variations in alignment quality, that is, each of the alignment methods may be affected in different ways by the number of sequences to align, the similarity amongst the sequences or by their length. Previous studies of alignment methods have in fact shown this to be the case (Thompson et al., 1999a). To separate out these effects we have
Use of structure information to increase alignment accuracy

Table 1. Average similarity to structural alignment as measured by the proportion of pairs of aligned residues (standard deviation) which are identical between the two alignments

<table>
<thead>
<tr>
<th>Method</th>
<th>Average Identity (s.d.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-COFFEE</td>
<td>0.804 (0.200)</td>
</tr>
<tr>
<td>CLUSTALW</td>
<td>0.782 (0.225)</td>
</tr>
<tr>
<td>Gap1</td>
<td>0.395 (0.257)</td>
</tr>
<tr>
<td>Gap5</td>
<td>0.189 (0.136)</td>
</tr>
<tr>
<td>Gap80</td>
<td>0.080 (0.060)</td>
</tr>
</tbody>
</table>

categorized our set of alignments according to the number of sequences in the alignment, the average identity of sequences within an alignment, and average length of the sequences, as detailed in the methods. The distributions of these parameters within our dataset are shown in Figure 1. The number of sequences per alignment is highly skewed towards low numbers—the possible effects of this are discussed below. The average sequence length is skewed towards shorter sequences, but this distribution is similar to that observed in the sequence databases. The average sequence identity is slightly skewed to <50% identity.

Figure 2 shows pairwise identity between the sequence-based methods and the structural alignment for each subset of alignments. The data clearly show that CLUSTALW and T-COFFEE alignments become more similar to the structural alignments with increasing sequence identity, from only 40% to nearly 100% alignment identity across the range shown. In contrast, increasing numbers of sequences and sequence length seem to have relatively little effect on the quality of the T-COFFEE and CLUSTALW alignments. The variation in quality of the artificially generated alignments purely reflects the probability algorithm used to insert gaps, with a clear decrease in the similarity to structural alignments with increasing sequence length. These data are included to allow later analysis of how these changes affect profile HMM performance.

Model sensitivity and specificity

Much of the data presented up to this point overlaps with and is consistent with previous studies of alignment methods (Thompson et al., 1999b). However, our primary aim is to explore the ability of models built from these multiple sequence alignments to find homologues, something which has not been discussed previously in the literature.

Our approach is to use the alignments derived above to study, not similarity to structural alignments, but sensitivity and specificity of profile HMMs built from these alignments. All sensitivity and specificity values are calculated with respect to Pfam family membership, and are normalized against the performance of models built from the structural alignments. It is important to note that Pfam is used only to provide a measure of trusted family membership—no HMMs are built from Pfam data so we are not biasing the performance of any models. Average sensitivity and specificity values over all models are shown in Table 2.
The striking feature of these data is the overall performance of these models. The average sensitivity of the structural alignment-based profile HMMs with respect to Pfam is 0.687, and reflects the relatively small size of the HOMSTRAD families. Those built from T-COFFEE and CLUSTALW aligned sequences appear to match the performance of the structural model (mean relative sensitivities 0.995 and 0.991 respectively). The models built from the poorly aligned sequences perform surprisingly well with the Gap1 set finding 84% of matches detected by the structural alignment-derived models, and even the essentially unaligned Gap80 set of alignments able to generate models with mean sensitivity of 0.576. Paired t-test analysis shows that models built from Gap sets differ significantly in performance from those built from structural alignments ($p < 0.001$), with CLUSTALW bordering on significance ($p = 0.018$), and the performance difference between HOMSTRAD and T-COFFEE aligned profile HMMs not significant ($p = 0.22$).

The relative specificities of the models built by all methods are highly invariant. This reflects the small number of sequences in the seed alignments—the vast majority have five or fewer sequences (see Figure 1b). This is in marked contrast to the seed alignments from which the Pfam HMM library is built, which frequently contain hundreds or even thousands of sequences. The corresponding Pfam families generally subsume the HOMSTRAD families and models built from only sequences with structures from HOMSTRAD therefore find very few additional sequences—on average 0.43 sequences per family (which we count as false positives here). Interestingly, model specificity actually increases slightly as the alignment quality decreases, concomitant with the significant decrease in sensitivity. This is simply due to the fact that the lower quality models always find a smaller subset of those proteins detected with models derived from the best alignments, and hence tend to find even fewer false positives. If we disregard all families where additional sequences are detected, the relative sensitivities are: T-COFFEE, 0.998; CLUSTALW, 0.996; Gap1, 0.852; Gap5, 0.767; Gap80, 0.601. These numbers are within the error quoted in Table 2.

It is important to note that the above analysis is reliant upon using a blanket e-value threshold of 0.01 for matching the profile HMM. In some specific cases it

Table 2. Relative mean (standard deviation) sensitivity of profile HMMs built from multiple sequence alignments normalized against values for HOMSTRAD-derived profile HMMs utilizing an e-value threshold of 0.01.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-COFFEE</td>
<td>0.995 (0.141)</td>
<td>0.999 (0.025)</td>
</tr>
<tr>
<td>CLUSTALW</td>
<td>0.991 (0.166)</td>
<td>0.999 (0.033)</td>
</tr>
<tr>
<td>Gap1</td>
<td>0.839 (0.222)</td>
<td>1.002 (0.019)</td>
</tr>
<tr>
<td>Gap5</td>
<td>0.749 (0.252)</td>
<td>1.003 (0.020)</td>
</tr>
<tr>
<td>Gap80</td>
<td>0.576 (0.295)</td>
<td>1.007 (0.048)</td>
</tr>
</tbody>
</table>
is obvious that differences in sensitivity are due to true matches falling just one side or the other of this threshold. In an effort to test the impact of this effect on our findings, we have calculated coverage and errors per query (EPQ) for each method over a range of thresholds. Coverage is defined as the total number of true matches found by the 348 models as a fraction of the total number of proteins in the corresponding Pfam families, and can be seen as a measure of sensitivity. EPQ is defined as the average number of additional hits detected by the models, and can be seen as a measure of specificity. These data are shown in Figure 3a.

The plot clearly shows the trade-off between coverage and specificity, with all methods showing increased coverage with an increase in the number of allowed errors. The coverage values are generally well below one as many of the HOMSTRAD families are small in comparison with the corresponding Pfam families against which sensitivity is measured. Again the data clearly show that T-COFFEE and CLUSTALW derived models perform essentially identically to the HOMSTRAD derived models, and that this performance is independent of the chosen e-value threshold. This justifies the choice of an arbitrary blanket threshold of 0.01 for all subsequent sensitivity analysis.

As already discussed, an important consideration when analyzing these findings is that the numbers of sequences in the alignments are small. HMMs built from only two unaligned sequences are able to model those two sequences relatively well, but this capability will diminish with increasing numbers of sequences. Thus it is important to demonstrate that the above results are not unduly biased by the low sequence number, and to this end we have calculated coverage and EPQ values for the 40 alignments with six or more sequences. These results are plotted in Figure 3b. Obviously, the relatively low number of alignments has a bearing on the significance of the results. However, the data show that even in this smaller subset of HOMSTRAD families, the models derived from the sequence-based alignments match the performance of the structure-based models.

For the reasons outlined above, we would expect the surprisingly high sensitivity of models built from unaligned sequences to fall dramatically as the number of sequences increases. This is indeed shown by the data in Figure 4b. In contrast, there is no clear correlation between relative model sensitivity and number of sequences for the T-COFFEE and CLUSTALW based models which remain very close to one across the range studied. The intermediate quality Gap1 and Gap5 model sets show small decreases in performance as the number of sequences increases.

The finding that structural alignments do not increase the sensitivity of derived models is most clearly illustrated by the data in Figure 4a. Even below the so-called twilight zone of 30% sequence identity, the T-COFFEE and CLUSTALW derived models perform very similarly to those derived from structure alignments. This is despite the observed trend of similarity to structure alignment in Figure 2a which shows that the alignment quality changes significantly over this range. Average sequence length does little to alter the average performance of any of the models as shown in Figure 4c, despite a general reduction in the similarity to structural alignments.

**CONCLUSIONS**

Structural alignments are an invaluable resource. For example, secondary structure prediction, phylogeny and inference of functional features of proteins all rely on using the highest possible quality alignments. The work of others has shown substantial improvements in align-
Fig. 4. Average sensitivity of profile HMMs normalized against those built from HOMSTRAD alignments. The data are categorized by (a) average sequence identity within an alignment, (b) average number of sequences, and (c) average sequence length.

ment quality when structural data are incorporated (Shi et al., 2001) and the work presented in this study shows that sequence methods alone are not able to reproduce the multiple sequence alignments derived from structural superposition. In addition, algorithms which make use of structural information in the form of position specific scoring matrices, such as FUGUE (Shi et al., 2001) and 3DPSSM (Fischer et al., 1999), outperform purely sequence-based methods in independent tests such as CAFASP (Fischer et al., 1999) and LIVEBENCH (Bujnicki et al., 2001).

However, the data presented above show that the performance of profile HMMs derived from multiple sequence alignments is relatively tolerant of the alignment quality. The best sequence methods (here CLUSTALW and T-COFFEE, but others have been found to perform similarly well; Thompson et al., 1999b), provide very adequate alignments for profile HMM creation in a large number of real cases. Even when the aligned sequences are mutually distant (<=40% sequence identity) it seems little performance is to be gained by extensive tweaking of these alignments, either manually or by the incorporation of added quality information such as structural alignment. In contrast, profile HMMs generated from poorly aligned sequences do suffer a performance decrease, although even essentially unaligned sequences are able to produce models which find nearly 60% of the homologues detected by those derived from structure alignments, at least in the case of small numbers of sequences. We expect that these results will also apply to other related methods such as profiles.

One possible reason for the apparent tolerance of profile HMMs to relatively poor quality alignments is that the quality difference we measure between the sequence-based alignments and those from structure comes mainly from poorly aligned residues in loop regions. In contrast to highly conserved regions, for instance in secondary structure, loop regions are likely to contribute little to the signal of the model.

So, in which areas should one focus time and effort to increase the number of homologues detected by profile HMMs? The most obvious area is threshold curation. Indeed, when we repeat the sensitivity measurements but choosing family specific thresholds such that no false positives appear above the cut-off, we find that the sensitivity differences narrow further (see Table 3). Again, T-COFFEE and CLUSTALW alignments produce models which perform essentially identically to the structure alignment, with the Gap1, Gap5 and Gap80 based models finding 93%, 86% and 74% of the HOMSTRAD hits.

The second, and perhaps most fruitful strategy is an iterative approach to seed alignment generation. The incorporation of distant homologues into the seed alignment, and then re-searching using the new profile HMM is a strategy commonly employed to improve families in the Pfam database, and this iterative approach also forms the basis of PSI-BLAST (Altschul et al., 1997), SAM-T98 (Karplus et al., 1998) and SUPERFAMILY (Gough et al.,
Table 3. Relative mean (standard deviation) sensitivity of profile HMMs built from sequence alignments, normalized against those from structural alignment. Family specific thresholds are chosen to maximize the number of true positives and eliminate false positives (as determined by Pfam).

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-COFFEE</td>
<td>1.002 (0.121)</td>
</tr>
<tr>
<td>CLUSTALW</td>
<td>0.996 (0.128)</td>
</tr>
<tr>
<td>Gap1</td>
<td>0.930 (0.212)</td>
</tr>
<tr>
<td>Gap5</td>
<td>0.857 (0.227)</td>
</tr>
<tr>
<td>Gap80</td>
<td>0.736 (0.282)</td>
</tr>
</tbody>
</table>

In light of the findings presented above, this approach is likely to be far more effective than improvement of alignment quality in boosting the performance of profile HMMs for homologue detection.

ACKNOWLEDGEMENTS

We are grateful to Richard Durbin, Paul de Bakker and Sean Eddy for helpful comments.

REFERENCES


