BIOINFORMATICS

Vol. 14 no. 10 1998
Pages 830–838

Empirical estimation of the reliability of ribosomal RNA alignments

Emmet A. O'Brien and Desmond G. Higgins

Department of Biochemistry, University College, Cork, Ireland

Received on June 11, 1998; revised on September 2, 1998; accepted on September 4, 1998

Abstract

Motivation: *The automatic alignment of rRNA sequences can reproduce manual expert alignments with high, but not perfect, fidelity. We examine the use of empirical methods for the identification of regions of an alignment of a new sequence with an existing large alignment which can confidently be predicted to be correctly aligned.*

Results: *We show how to use a simple jack-knife procedure to derive an estimate of the reliability that is to be expected at each position of a large alignment of eukaryotic rRNA sequences. These reliabilities are then improved using measures that are specific to the input sequence. Regions where the sequence-specific reliability method performs particularly well are identified and seen to correspond with elements in the structure of the rRNA molecules that vary between species in the alignment. We also compare these reliability measures to an algorithmic alignment stability measure.*

Availability: *The software is available free of charge by sending an e-mail message to emmet@chah.ucc.ie.*

Contact: *emmet@chah.ucc.ie*

Introduction

Automatic sequence alignment methods are very widely used tools in sequence analysis. Most methods are based directly or indirectly on the dynamic programming algorithm of Needleman and Wunsch (1970), which finds the optimal alignment between two sequences, given scores for gaps and aligned residues. This method is usually used to find a single best alignment, defined as the alignment giving the best score. The method has been generalized to find the best alignment between a sequence and a set of pre-aligned sequences, often referred to as a profile (Gribskov *et al.*, 1987).

It has proven difficult fully to express dynamic programming results in a probabilistic framework. This is a major advantage of alternative alignment strategies, using hidden Markov models (HMMs) (Krogh *et al.*, 1994). The probability of observing a given alignment, with gaps, can be estimated analytically (Waterman and Vingron, 1994) from random shuffling experiments or from the alignment scores given by a database similarity search (Collins and Coulson, 1990; Pearson, 1998). This gives a guide as to the significance or otherwise of an entire alignment. Some attempts have also been made to generate alignments using probability models in a maximum-likelihood (Thorne *et al.*, 1992) or Bayesian framework (Zhu *et al.*, 1998).

An alternative approach is to examine the significance of each position in the alignment. The method of Mevissen and Vingron (1996) derives a stability measure for each aligned pair of residues in an alignment, which is defined as the difference between the score of the highest-scoring alignment containing that aligned pair and the score of the highest-scoring alignment which does not contain that pair. These measures are then divided into stability classes, from highest to lowest. So, for example, positions are considered uncertain when there is little to choose between the alternative alignments at that position. At the other extreme are positions which are considered of low uncertainty because alternative alignments which do not align these pairs of residues have much poorer alignment scores. Mevisson and Vingron (1996) give some efficient methods for estimating these alignment stability scores and for calibrating these with biological reality using test cases.

In this paper, the specific case of profile alignment, in which a new sequence is aligned against a large existing alignment, is examined. The eukaryotic small subunit ribosomal RNA (SSU rRNA) is used as a test case because of the availability of a large alignment of >1500 sequences (Van de Peer *et al.*, 1997). Other authors have made extensive use of protein structures to derive accurate reference datasets (e.g. Brenner *et al.*, 1998). The method of Mevissen and Vingron (1996) is general and can be used to examine the stability of any alignment based on dynamic programming. In our case, the information contained in the large alignment allows for a simpler and more direct generation of alignment reliability measures. A jack-knife procedure is carried out in which each of the sequences in the existing alignment in turn is removed and realigned with the remainder. For each sequence, those positions which are correctly realigned, as judged by comparison with the original alignment, are noted. Finally, the percentage of the sequences in the alignment that can be aligned correctly at each position is noted and these are used directly as measures of alignment reliability.

In this case, we examine the reliability of an alignment which is, informally, a measure of the accuracy with which each residue in a sequence may be aligned. The stability or uncertainty measure of Mevissen and Vingron measures the degree to which an alignment position is unambiguous. The two measures may correlate well in simple cases, but they may behave differently in others.

In this study, those positions with a reliability score of 95% or more are considered easy to align correctly and predicted to be reliable in future alignments with new sequences, while those with low scores are more difficult and predicted to be less reliable with future alignments. The success of the procedure depends on the coverage of the alignment and will be inaccurate if very aberrant sequences are aligned in the future. However, if the new sequences come from within the phylogenetic scope of the test dataset, then the large number of sequences involved should protect against difficulties.

The judgement of which positions are correctly aligned is based on the opinions of the experts who created the alignment initially. In practice, new sequences are added to the SSU rRNA alignment by an initial automatic alignment procedure based on sequence identity, which is then refined manually, taking into account the locations of the main secondary structure elements of stems and loops as well as the locations of conserved core regions and variable expansion segments. It is not possible to capture this information easily with simple dynamic programming methods and it is to be expected that there will be some parts of the alignment which will be difficult to duplicate, regardless of the sophistication of the methods. The current empirical approach will reveal these difficult sections automatically.

Methods

The test alignment was taken from the WWW server of the group of Rupert De Wachter (Van de Peer *et al*., 1997) at http://www-rrna.uia.ac.be/ssu/index.html. This is a pre-aligned collection of SSU rRNA molecules taken from all taxonomic groups. All the eukaryotic nuclear sequences were extracted. The sequences of *Babesia bovis* 4 and *Butomus umbellatus* were removed, leaving 1517 aligned sequences, as they were not fully aligned with the rest and the resulting alignment is 5370 characters long, including gaps. Typically, eukaryotic nuclear SSU rRNA sequences are between 1.4 and 2.2 kb in length. The alignment is over twice that length because of the presence of many extra insertions in variable regions of many different sequences. This alignment is our test set and we wish to predict the reliability of alignment of new sequences along its length.

The jack-knife procedure involved taking each of the 1517 sequences from the alignment and creating a profile (Gribskov *et al*., 1987) using the rest. The removed sequence was aligned with the profile using dynamic programming and a set of parameters which were optimized in an earlier study by examination over a wide range of values and a series of 16 test case sequences (O'Brien at *al*., 1998). Affine gap penalties were used, with a gap opening penalty of 6.0 and a gap extension penalty of 0.2 relative to a score of 1 for a match with a full column in the profile and a mismatch score of 0. Position-specific gap penalties were used, such that the penalty for opening a gap in the sequence was inversely related to the frequency of gaps at that position in the profile. This reduces the penalty for addition of gaps in the sequence at positions which have many gaps in the profile. Terminal gaps were penalized solely with a gap extension penalty.

A combination of two types of sequence weights was used for the sequences in the profile. The first is used to counter the effect of over-representation in the profile of certain taxonomic groups, such as fungi and vertebrates. This sampling error distorts the alignment of new sequences. The weighting scheme of Thompson *et al*. (1994), which takes a phylogenetic tree of the sequences and assign weights using the lengths of the branches, is designed to counteract this. These weights are set up in advance using a rooted neighbour-joining tree calculated from the alignment (Saitou and Nei, 1987). As a first approximation, the weight for a sequence is equal to the distance of the sequence from the root of the tree. If the sequence shares a branch with another, it shares the weight. This gives very low weight to closely related groups of sequences and high weight to ones on long branches with no close relatives.

The second type of weights are assigned to all of the sequences in the profile depending on the sequence which is to be aligned. This attempts to give most weight to those sequences which are most closely related to the test sequence (O'Brien *et al*., 1998). The distance (d; mean number of differences per aligned site, ignoring gaps) between the test sequence and each sequence in the profile is calculated from an approximate alignment against a profile in which each sequence is weighted equally. The inverse of each distance ($1/d$) is then used as a weight. These two weights are multiplied together and normalized before the alignment of each sequence in the jack-knife experiment. Weights generated using this combination of schemes are then assigned to each sequence in the alignment, and a profile is calculated from these weights which is specific to the particular input sequence. Dynamic programming alignment of single new sequences with these specific profiles was found to give the most accurate alignments in earlier tests (O'Brien *et al*., 1998).

The sequences are then removed and realigned one at a time. This gives the percentage of the sequences which can be correctly realigned at each position. These values can immediately be used to assign a reliability to each position of a new alignment, based solely on sequence identity, and are referred to here as general reliabilities. Figure 1 shows the

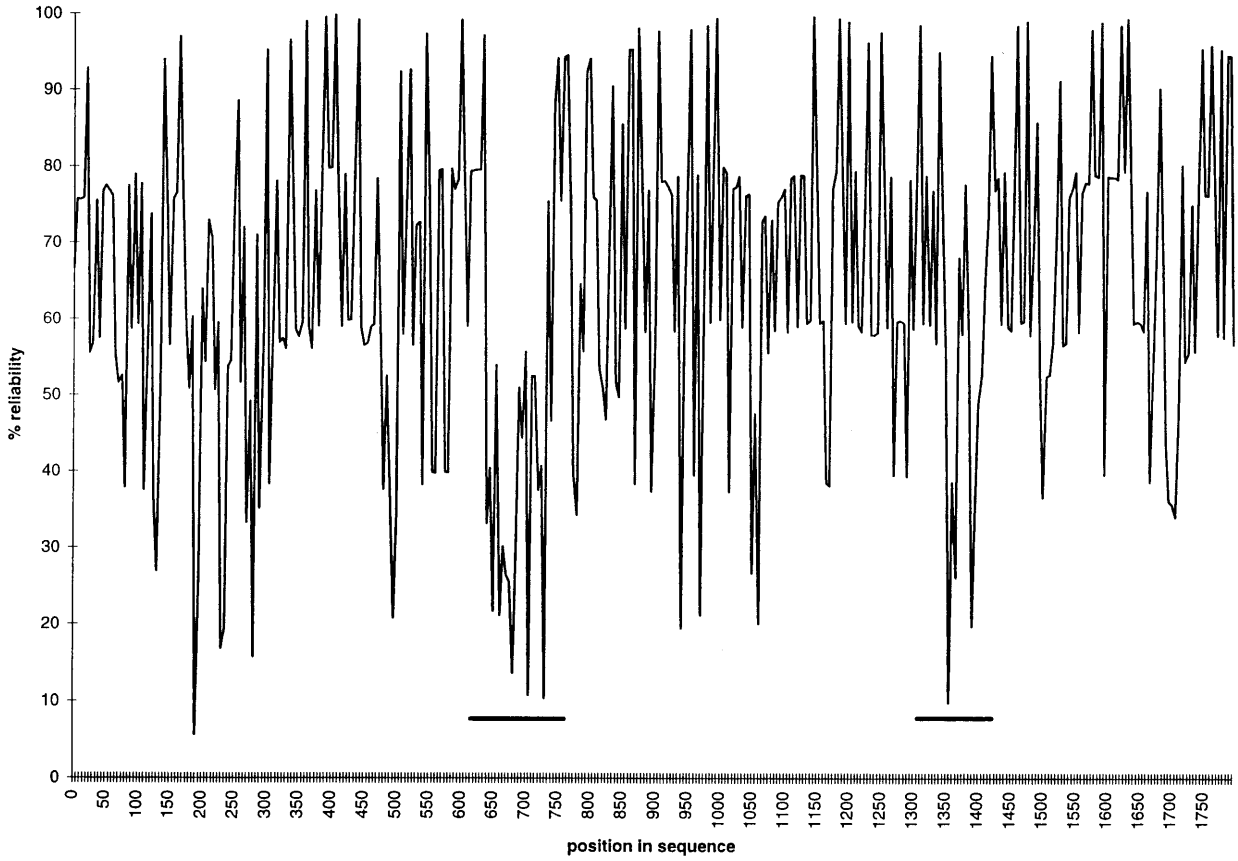

Fig. 1. Graph of general reliability values for the profile plotted along the nuclear SSU rRNA sequence of *Saccharomyces cerevisiae.* The horizontal scale shows the nucleotide position. The two horizontal lines indicate the expansion regions in the sequence.

values of this general reliability measure plotted along the SSU rRNA sequence of *Saccharomyces cerevisiae*. The values range from <10% in regions which are almost impossible to align correctly up to 100% in regions which can always be aligned accurately. Regions of lower reliability between approximately positions 620–750 and 1320–1390, marked by horizontal lines in the figure, correspond to the sections of the molecule which are most variable in length between the species in the alignment.

These general reliabilities are the same regardless of which sequence is being aligned to the large profile. In practice, some sequences will be easier to align than others. Some sequences will come from taxonomic groups which are very well represented in the profile and whose alignment reliability might be more accurately predicted if this information could be used. This is addressed by basing the reliability estimates on only those sequences which are most closely related to the test sequence, as measured by the distance between it and each sequence of the profile to give sequence-specific reliabilities.

The sequence-specific reliability profile is generated by considering only some subset of the sequences which are closely related to the new sequence. In the extreme case, one could use just the single most closely related sequence from the profile. This would use the information from the most relevant part of the underlying phylogenetic tree at the cost of having low generality. Two intermediate solutions were tried. First the distribution of distances (mean numbers of differences per site, ignoring gap positions) was taken between the new sequence and the sequences from the profile. The mean distance was then taken, those sequences having less than the mean distance being more closely related to the test sequence. Finally, reliability profiles were calculated based on each of two subsets of the sequences: (i) those sequences having less than the mean distance (mean-specific reliabilities) or (ii) those sequences having less than the mean distance minus one standard deviation (1 SD-specific reliabilities).

Table 1. Performance of empirical reliability measures where len is the length of the test sequence, general are reliability values derived from the entire alignment, mean are reliability values derived from a subset of sequences nearer than mean distance to input sequence, 1 SD are reliability values derived from sequences more than one standard deviation nearer to the input sequence than the mean, coverage is the percentage of residues per sequence accurately predicted at 95% accuracy threshold, EPS is the errors per site, i.e. the percentage of residues per sequence incorrectly predicted at 95% accuracy threshold, and no. is the number of sequences used to generate reliability values

	len	General		Mean			1 SD		
		Coverage	EPS	No.	Coverage	EPS	No.	Coverage	EPS
A.beccarii	1714	53.2	1.9	956	75.9	5.8	161	76.0	6.7
C.elegans	1760	49.5	4.2	995	69.9	10.0	86	71.9	7.2
D.discoideum	1872	48.3	2.2	995	67.5	7.6	140	70.5	8.1
D.melanogaster	1996	45.8	1.4	935	67.9	3.2	167	68.5	3.4
E.histolytica	1947	45.7	2.7	979	62.6	9.1	121	64.2	10.3
E.gracilis	2305	34.7	6.3	989	48.4	12.9	110	50.5	15.8
Giardia sp.	1440	57.1	8.7	1080	67.3	17.8	15	29.6	34.3
Hexamita sp.	1550	53.3	7.8	1072	68.0	18.9	99	69.4	19.7
H.sapiens	1869	50.2	0.2	822	76.9	0.3	97	75.2	0.1
N.gruberi	2019	40.0	6.7	1063	54.1	14.0	109	59.2	18.2
O.sativa	1812	51.8	0.2	635	81.6	0.2	210	85.9	0.5
P.polycephalum	1964	44.3	3.8	1033	59.2	12.5	120	60.9	15.9
S.cerevisiae	1798	52.3	0.2	707	81.0	0.8	189	85.9	1.1
T.brucei	2251	39.5	2.4	985	51.9	9.0	84	43.4	0.9
V.lobospinosa	2137	38.4	5.9	998	52.1	13.7	100	53.3	18.3
X.laevis	1825	51.3	0.4	826	65.6	0.4	114	77.8	14.0
Mean		47.2	3.4	942	65.6	8.5	120	65.1	10.9

This gives three possible reliability profiles for the alignment of a given test sequence with the profile. In order to compare these, the performance with 16 test sequences taken from the large alignment, for which correct reference alignments exist, was evaluated. These are the same 16 test cases that were used by O'Brien *et al*. (1998) to optimize the parameters for alignment and the sequence names are shown in Table 1. Each of these test sequences was aligned to the large alignment with default parameters and weights.

One measure of success in using these measures is then how many of the positions in the sequence are correctly predicted to be aligned, as a percentage of the total number of positions in each sequence. This measure is referred to as the coverage and gives an indication of the sensitivity of the method. The coverage gives only a partial representation of the difference in performance between the reliability measures as it does not consider false positives, i.e. those positions predicted by the reliability profile to be correctly aligned which are actually incorrect. The measure used for this is errors per site (EPS), which measures false positives—those positions incorrectly predicted to be aligned—as a percentage of the total number of positions in the sequence.

These measures were evaluated for each of the three methods at various reliability thresholds as described by Brenner *et al*. (1998). The coverage and EPS were measured for each test case for those positions predicted with 5% reliability or greater, for those predicted with 10% reliability or greater, and so on in steps of 5% to the positions predicted with 95% reliability or greater. The mean values, taken across all 16 test cases, for coverage and EPS at each threshold were plotted for each of the three reliability methods. The balance between coverage and EPS allows the potential utility of the different reliability profiles to be judged.

Finally, the algorithmic stability method of Mevissen and Vingron (1996) was used to generate position-specific stability estimates for some of the test case alignments used above. This method does not give directly comparable reliability measures. The stability scores were divided into 20 equally spaced bins ranging from most to least stable, and the numbers of true positives found in each of these bins were

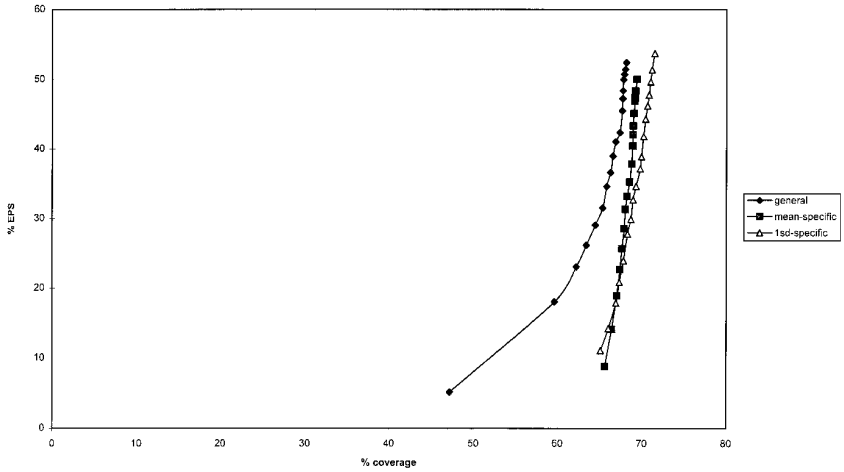

Fig. 2. Graphs of coverage against EPS for the three empirical reliability methods. Values on the horizontal axis represent the covera ge, or percentage of the residues in the sequence that are correctly predicted by the method. Values on the vertical axis represent th e EPS, the percentage of those residues in the sequence that are incorrectly predicted to be correctly aligned.

compared with the distribution of reliability scores from the general and sequence-specific reliabilities in the *S.cerevisiae* test case. These comparisons are difficult as the different methods produce different types of scores and are based on different principles. The results are indicative of their relative behaviours rather than being definitively comparable.

Results

The main results are shown in Table 1 which illustrates the sensitivity and specificity of each method at a threshold of 95%. The sensitivity is measured by the coverage, which is defined as the percentage of the residues in a sequence that are predicted accurately by the method. The specificity is measured by EPS; this is the percentage of the residues in the sequence that are predicted incorrectly by the method. Similar measures have previously been used by Brenner *et al.* (1998) to analyse the performance of various methods for similarity searches. A position is considered 'predicted' to be correct if the reliability value at that position is greater than some arbitrary threshold: 95% in this case. The first columns give the numbers using the general reliability profile; this profile uses the same set of reliability measures for all sequences. A total of 47.2% of the residues in each sequence are predicted by the general reliabilities. This is accompanied by an average of 3.4% EPS.

The use of sequence-specific reliability scores causes a dramatic increase in the number of positions correctly identified. The coverage rises to 65.6%. This is, however, accompanied by a corresponding increase in the numbers of positions erroneously predicted as correctly aligned, which rise to a mean of 8.5% of the residues in the sequence. The performance of the sequence-specific scores is strongly dependent on the input sequence. Those sequences from very densely populated parts of the underlying phylogenetic tree are well predicted with very few false positives. Specifically, the sequences of *H.sapiens*, *O.sativa*, *S.cerevisiae* and *X.laevis* all have >1400 correctly predicted positions with only 0.1–0.8% EPS.

The final columns from Table 1 show the use of sequence-specific reliability scores derived from just a selected group of sequences that are very closely related to the test sequence. These use only those sequences with a distance of less than the mean distance from the test sequence to all the profile sequences, minus 1 SD. Here there is a slight decrease in the sensitivity, to 65.1% coverage, and a further increase in the EPS to 10.9%. With some input sequences such as *Giardia*, this cut-off leads to the reliability map being generated from an extremely small number of sequences and in consequence the predictions of the method are not of use.

Figure 2 gives a broader illustration of the performance of the various reliability methods, with coverage plotted on the horizontal axis and EPS on the vertical. The threshold at which a residue was deemed 'predicted'—the percentage reliability value that was necessary to consider that position in the alignment correctly predicted by the reliability method—was set at intervals of 5%, from 5 to 95%, and the coverage and EPS calculated for each test case at each value. Each point represents the average values of coverage and EPS for the 16 test cases at a given reliability threshold. Reducing the

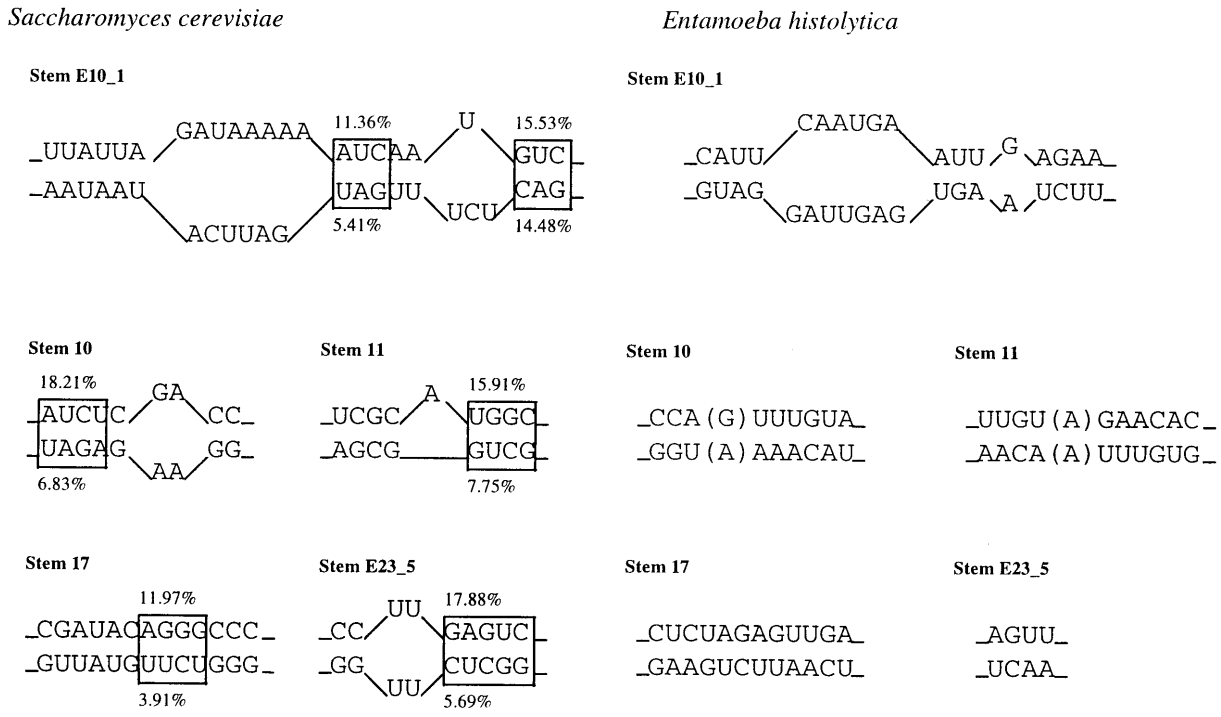

Fig. 3. Differences between reliabilities for five sample stems in the sequence *Saccharomyces cerevisiae*. Regions where sequence-specific reliabilities are >10% higher than general reliabilities are marked with boxes. The figure above each box represents the mean increase in reliability for the residues contained. The figure below each box is the increase in the percentage of residues at these positions that form a secondary structure pair. The structures of the same five stems in *Entamoeba histolytica* are shown for the purpose of comparison.

reliability threshold from 95% (the lowest and leftmost point in each data series) leads to a steep rise in EPS with little corresponding rise in coverage, except in the case of the general reliabilities, which are rising from a much lower initial coverage. This indicates that 95% gives an appropriate balance between sensitivity and specificity. At lower thresholds, the percentage error per site rises rapidly: for thresholds below 60%, more than half the values predicted by any of the methods are not correctly aligned. The mean-specific reliabilities perform better than the general reliability values—higher coverage and lower EPS—for all the thresholds investigated. The 1 SD-specific reliabilities perform slightly worse on average than the mean-specific reliabilities.

Each position of the general and sequence-specific reliability profiles from *S.cerevisiae* was compared. Any positions showing a large difference between the two profiles indicate regions which are not stable for the alignment of sequences from the profile as a whole, but where the yeast sequence and its relatives might be aligned more accurately or vice versa. It is in these regions, in particular, that the sequence-specific profiles might be most useful. Some short pieces of the *S.cerevisiae* sequence are shown in Figure 3 that come from five different stems in the predicted secondary structures of the sequences. Regions on the *S.cerevisiae* stems which showed a difference of 10% or more between the sequence-specific and general reliabilities were marked. The mean difference in reliability for all residues within the marked regions is indicated above them in each case. In these positions, there is also distinct variation in the percentages of residues paired, which is marked below the boxes. In each case, the variation is also an increase in the sequence-specific values compared to the general values. For comparison, the same stems are shown from one of the other test cases from a very different part of the underlying phylogenetic tree of the sequences. The boxed sections all show differences in either base pairing or sequence, or both.

Figure 4 shows the distribution of reliability scores across all 1798 positions of the *S.cerevisiae* sequence for the general and sequence-specific reliability scores and the algorithmic method of Mevissen and Vingron (1996). Each division represents an interval of 5% reliability, or one-twentieth of the range of stability scores. The distinction between general and sequence-specific reliabilities is very clear. The sequence-specific reliability method shows a much clearer

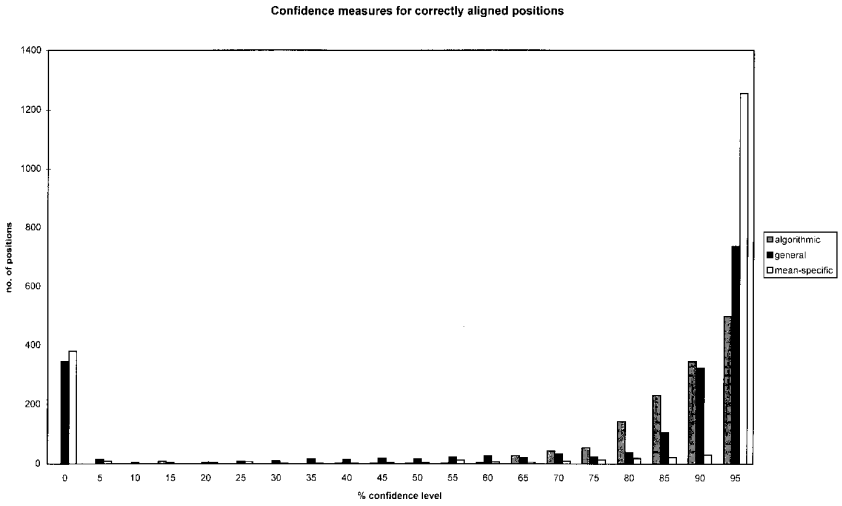

Fig. 4. Graph of the number of true positives in the optimum alignment of *Saccharomyces cerevisiae* identified as such with different degrees of confidence by three reliability methods. The divisions along the horizontal axis represent ranges of 5% reliability from 0–5% to 95–100% for the empirical reliability methods, and division of stabilities into 20 bins for the algorithmic method. Black bars represent the algorithmic stability method of Mevissen and Vingron (1996), white bars the empirical reliability method using the entire alignment, and grey bars the sequence-specific reliability method.

division between those positions that it unambiguously identifies as correct and those that it clearly does not, whereas with the general reliability method there is a large number of aligned positions which cannot be confidently identified as either correct or incorrect.

The algorithmic stability method assigns almost all positions in the alignment to high-stability classes. The division of positions among the higher stability classes shows even less of a distinction between those which would be deemed correct at a 95% confidence level, and those which would not than the general reliabilities. As the distribution of the stability scores between classes is based on relative scores within the alignment algorithm and has not been calibrated against any external measure of the accuracy of the alignment, this is a rough qualitative result rather than a definitive statement of relative merit.

Discussion

The method for estimating reliability of an alignment that is presented here depends on the availability of reference alignments containing many aligned sequences. This is not often the case in practice, but will become increasingly so as more and more sequences are accumulated for different sets of homologous sequences. When such large alignments do exist, it seems obvious that the most powerful method of determining which positions may be reliably aligned will be by reference to the reference alignments themselves and not by using general scoring schemes which are derived from many positions from many alignments. Parameters that are derived specifically from a given alignment have been shown to give significant improvement over a simple default profile (O'Brien *et al.*, 1998) and this indicates that it is possible, to a large extent, to reproduce some of the factors taken into account by expert manual alignment, such as base pairing, by careful adjustment of the parameters of an alignment algorithm based solely on sequence identity.

The general reliability measure in this case predicts that 939 positions from the profile can be correctly aligned at the 95% threshold, which is less than two-thirds of the length of even the shortest sequence in the alignment. It is clear that in some sequences the number of false-positive predictions, positions identified as correctly aligned which are not in fact so, is <1%. The number of false positives varies widely and the possible utility of this measure in identifying a reliable core of conserved residues depends on how closely related the sequence is to the rest of the alignment. The mean-specific reliability measure gives an average increase in coverage of 18.4%, and a definite increase is visible in every test case. The corresponding mean increase in EPS is 5.1% with a maximum increase of 11.1% (*Hexamita*). The 1 SD-specific weights give a slight further increase in coverage on average, but in some cases where the number of sequences used to generate the reliability map is too low the behaviour

of this method is unpredictable, such as the dramatic drop in coverage in the case of *Giardia*. The use of thresholds lower than 95% is contraindicated by the results, which show an increase in the number of positions predicted with a decrease of threshold that is mostly made up of false positives. This is particularly the case with the mean-specific reliability measures.

The method of Mevissen and Vingron (1996) is an elegant and general method for predicting alignment stability that has been shown to correlate well with information from structure-based alignments of protein sequences. It has the advantage of being applicable to any alignment scoring scheme and to any sequences where conventional dynamic programming is used. In the current case, it has the disadvantage of not directly yielding estimates of how accurate an alignment position (a pair of aligned residues or a residue aligned with a gap character) may be, but of indicating which positions are unambiguous with respect to the scoring scheme—those positions for which alternative alignment configurations would receive a much inferior score. This can end up giving high support to incorrectly aligned positions simply because the alternative alignments get even lower scores.

In practice, with protein alignments, the method of Mevissen and Vingron gives a clear distinction between alignment positions which are believable and those which are dubious, but rRNA may be more difficult. rRNA sequences contain runs of the same residue or repeats of short motifs (Hancock and Dover, 1990). This is especially true of sequences such as those of *Plasmodium falciparum* which have very biased nucleotide compositions (very A+T rich). These are more difficult to align in the first place (O'Brien *et al*., 1998) and will generate alignments of low stability because of the many alignments scoring equally that are possible in regions with runs of the same residue. It is unclear which method for calculating alignment reliability is most appropriate to this situation.

Direct comparisons of the current method and that of Mevissen and Vingron with regard to the numbers of true and false positives are unfair and difficult because of the different bases of the two schemes. The numbers in the top bin of the stability method were counted, but performed poorly, finding far fewer true positives than our method. This is inconclusive as the top bin of the stability is not unequivocally comparable to the 95% reliability level of our method. The histogram that compares the three schemes (general reliability, mean-specific reliability, and Mevissen and Vingron) does show some qualitative differences. The empirical methods are seen to give a much clearer distinction between those positions that are unambiguously identified as correct and those that are not.

The current scheme is unwieldy from the point of view of quickly generating reliability scores to order for particular alignments. For large reference alignments, however, all of the time-consuming computing has been done in advance and one merely needs to store the reliability profiles; then the scores are very simple and fast to use. We made no attempt to compare the scheme with methods that directly use probabilities such as those based on HMMs or stochastic context-free grammars (Eddy and Durbin, 1994; Sakakibara *et al*., 1994). The jack-knife approach is simple and of general applicability, and could easily be adapted to such methods. HMMs can directly yield probability estimates for every aligned position. It would be interesting to compare these with the estimates derived from this work.

Acknowledgement

This work was supported by a grant (BIO4-CT95-0130) from the EU Biotechnology programme.

References

Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.

Collins,J.F. and Coulson,A.F.W. (1990) Significance of protein sequence similarities. *Methods Enzymol*., **183**, 474–487.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*., **22**, 2079–2088.

Gribskov,M., McLachlan,A. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.

Hancock,J.M. and Dover,G.A. (1990) Compensatory slippage in the evolution of ribosomal RNA *genes. Nucleic Acids Res*., **18**, 5949–5954.

Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology: applications to protein modelling. *J. Mol. Biol*., **235**, 1501–1531.

Mevissen,H. and Vingron,M. (1996) Quantifying the local reliability of a sequence alignment. *Protein Eng*., **9**, 127–132.

Needleman,S. and Wunsch,C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*., **48**, 443–453.

O'Brien,E., Notredame,C. and Higgins,D. (1998) Optimisation of ribosomal RNA profile alignments. *Bioinformatics*, **14**, 332–341.

Pearson,W.R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol*., **276**, 71–84.

Saitou,N. and Nei,M. (1987) The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*., **4**, 406–425.

Sakakibara,Y., Brown,M., Hughey,R., Mian,I.S., Sjolander,K., Underwood,R.C. and Haussler,D. (1994) Stochastic context-free grammars for tRNA modelling. *Nucleic Acids Res*., **22**, 5112–5120.

Thompson,J., Higgins,D. and Gibson,T. (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Applic. Biosci*., **10**, 19–29.

Thorne,J.L., Kishino,H. and Felsenstein,J. (1992) Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol*., **34**, 3–16.

Van de Peer,Y., Jansen,J., De Rijk,P. and De Wachter,R. (1997) Database on the structure of small ribosomal subunit RNA. *Nucleic Acids Res*., **25**, 111–116.

Waterman,M. and Vingron,M. (1994) Rapid and accurate estimates of statistical significance for sequence database searches. *Proc. Natl Acad. Sci. USA*, **91**, 4625–4628.

Zhu,J., Liu,J.S. and Lawrence,C.E. (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, **14**, 25–39.