BIOINFORMATICS

Vol. 14 no. 4 1998
Pages 332–341

Optimization of ribosomal RNA profile alignments

Emmet A. O'Brien, Cedric Notredame[1] and
Desmond G. Higgins

Department of Biochemistry, University College, Cork, Ireland and [1]EMBL-European Bioinformatics Institute, Hinxton, Cambridge CB10 1RQ, UK

Received on October 28, 1997; revised and accepted on January 23, 1998

Motivation*: Large alignments of ribosomal RNA sequences are maintained at various sites. New sequences are added to these alignments using a combination of manual and automatic methods. We examine the use of profile alignment methods for rRNA alignment and try to optimize the choice of parameters and sequence weights.*

Results*: Using a large alignment of eukaryotic SSU rRNA sequences as a test case, we empirically compared the performance of various sequence weighting schemes over a range of gap penalties. We developed a new weighting scheme which gives most weight to the sequences in the profile that are most similar to the new sequence. We show that it gives the most accurate alignments when combined with a more traditional sequence weighting scheme.*

Availability*: The source code of all software is freely available by anonymous ftp from chah.ucc.ie in the directory /home/ftp/pub/emmet, in the compressed file PRNAA.tar.*

Contact*: emmet@chah.ucc.ie, des@chah.ucc.ie*

Introduction

Ribosomal RNA sequences (rRNA) are widely used to estimate the phylogenetic relatedness of groups of organisms (e.g. Sogin *et al.*, 1986; Pawlowski *et al.*, 1996), especially that of the small subunit (SSU rRNA). The SSU rRNA has been sequenced from thousands of different species and large alignments are maintained at several sites (Maidak *et al.*, 1997; Van de Peer *et al.*, 1997). The alignments are large and complex and the addition of new sequences is a demanding task, either for the alignment curators or for individuals who wish to align new sequences with existing aligned sequences. In simple cases, automatic alignment programs such as Clustal W (Thompson *et al.*, 1994a) may be used to align groups of closely related sequences or as a prelude to manual refinement. There may be large stretches of unambiguous alignment with high sequence identity which may be useful for phylogenetic purposes. The fully automated, accurate alignment of rRNA sequences remains a difficult problem, however.

In principle, one can use profile alignment methods (Gribskov, 1987) which use dynamic programming algorithms (Needleman and Wunsch, 1970, Gotoh, 1982) to align a new sequence against an existing 'expert' alignment. For example, one could take an alignment of all SSU rRNA sequences from one of the rRNA collections and one could use this as a guide; aligning each new sequence in turn, treating the large alignment as a profile. This approach has the advantage of simplicity and speed but the final accuracy may be limited by the lack of any ability to use secondary structure information. The RNALIGN approach (Corpet and Michot, 1994) or the stochastic context free grammar approach (Eddy and Durbin, 1994; Sakakibara *et al.*, 1994) provide elegant methods for the alignment of rRNA sequences taking both primary sequence and secondary structure into account. These methods, however, are very demanding in computer resources and cannot deal easily with pseudoknots so that their immediate application to the alignment of SSU rRNA sequences is not trivial.

In this paper, we examine, empirically, the effectiveness of profile alignment methods for the alignment of RNA sequences. We remove test sequences from existing 'expert' alignments and measure the extent to which they can be realigned with the original alignment, automatically. We use the eukaryotic SSU rRNA sequences from Van de Peer *et al.* (1997) as a test case. For a range of test sequences, we measure the number of positions that can be correctly realigned over a range of different parameters (gap opening and gap extension penalties).

Sequence weighting has been shown to increase the reliability of profile alignments using amino acid sequences (Thompson *et al.*, 1994b). This can be used to give less weight to clusters of closely related sequences and increased weight to sequences with no close relatives in order to counteract the effect of unequal sampling across a phylogenetic tree of possible sequences. We examine the effectiveness of one commonly used scheme (Thompson *et al.*, 1994b). We also propose a new weighting scheme which is designed to give increased weight to those sequences in the profile (reference alignment) which are closest (highest sequence identity) to the new sequence being aligned. If a new mammalian sequence is being aligned, for example, it makes most sense to give a high weight to other mammalian sequences and decreasing weights to sequences that are more and more distantly related.

Some sections of SSU rRNA sequences are from regions whose secondary structure is conserved across many species. These conserved, 'core', regions are relatively easy to align with high accuracy but are interspersed with less conserved regions that may be very difficult to align. We empirically determine which regions of the eukaryotic reference alignment can be aligned with high accuracy by a simple jack-knife experiment. We remove each sequence, one at a time, and try to realign it with the rest. It is then a simple matter to count how often each nucleotide of each sequence is correctly realigned. This gives a definition of conserved core regions that is purely empirical and which can be used by users to delimit regions of alignment which can be safely used in phylogenetic research.

Finally, we examine the effect of G+C content of each sequence on the accuracy of alignment. Sequences of high or low G+C may be expected to be more difficult to align than those with more balanced nucleotide compositions.

System and methods

Small subunit ribosomal RNA

An alignment of eukaryotic, nuclear SSU rRNA sequences (that dated May 6, 1997) was obtained from the World Wide Web server at http://www-rrna.uia.ac.be/ssu/index.html (Van de Peer *et al.*, 1997). After removal of columns which consist only of gaps, the two incomplete sequences of *Butomus umbellatus* and the unaligned sequence *Babesia bovis 4* the alignment contains 1517 sequences and is 5370 characters long. Individual sequences vary widely in length, from<1300 nucleotides to >2500.

Sixteen test sequences were removed from and realigned with the reference alignment in order to measure the accuracy with which it was possible to recreate their original alignment. The sequences used were *Drosophila melanogaster, Xenopus laevis, Homo sapiens, Caenorhabditis elegans, Saccharomyces cerevisiae, Oryza sativa, Dictyostelium discoideum, Euglena gracilis, Ammonia beccarii, Physarum polycephalum, Entamoeba histolytica 1, Vahlkampfia lobospinosa, Giardia sp., Naegleria gruberi, Hexamita sp.* and *Trypanosoma brucei*. These sequences were chosen based on a phylogenetic tree of all the sequences in the alignment, in order to give a spread of test cases over a wide range of different positions in the tree. Re-alignment was carried out over a range of gap penalties and using a number of sequence weighting schemes as described below.

Dynamic programming

The reference alignment was converted into a profile (Gribskov *et al.*, 1987) which contains information on the frequency of each residue and gaps at each position. The test sequences were aligned with this using a dynamic programming algorithm (Needleman and Wunsch, 1970). We used Gotoh's algorithm (Gotoh, 1982) and maximized the similarity between the sequence and the profile. A homogenous column in the profile (just one of the four residues), with no gaps will get a score of 1.0 when aligned with the same residue in the test sequence and a score of 0 otherwise. Other columns score in proportion to the frequency of each of the four residue types. In positions in the profile where one or more of the sequences has a gap, gaps were treated as a class of residue for frequency calculations. Other methods have been proposed for generating profiles using the natural logarithms of residue frequencies which may be normalized by overall residue frequencies to give log-odds scores (see Henikoff and Henikoff, 1996 for a review). We carried out some tests using the latter scheme and found that performance was comparable although slightly inferior to that using simple frequencies. Therefore we only present results obtained using the frequencies.

Gap penalties

A range of gap opening and extension penalties were used in alignment generation. For each test sequence and each weighting scheme, a total of 81 alignments were carried out. Gap opening penalties were used ranging from 1 to 9 in increments of 1, and gap extension penalties ranging from 0.1 to 0.9 in increments of 0.1. This range of ratios between gap penalties and residue match scores was chosen as it encompasses values empirically shown to give alignments of biological relevance. Terminal gaps were penalized solely with an extension penalty.

Position-specific gap opening penalties were derived from the frequency of gaps at each position along the alignment. At each position, a value equal to the number of residues (non-gap characters) in the column divided by the number of sequences in the alignment was derived. This value was then multiplied by the gap opening penalty, as taken from the range above, to give a specific gap opening penalty at each position. This gives gap opening penalties which are higher in positions at which residues mostly occur in comparison with positions which are occupied mostly by gaps.

Sequence weighting

By default, each sequence in the existing alignment will have an equal effect on the alignment of new sequences with the profile. If additional information is available concerning the relationships of sequences within the alignment to each other and to the sequence being aligned, this may not be optimal. For example, if a new sequence is identical to a sequence already in the alignment, the correctly aligned position of each residue in the new sequence could be deduced solely from that one identical sequence, and no information concerning the other sequences is necessary. Further, sampling bias can lead to an unequal representation of taxa within the alignment (e.g. there might be very many sequences from some taxa and very few from others), and it is possible to use sequence weighting to correct for this also. Three different weighting schemes

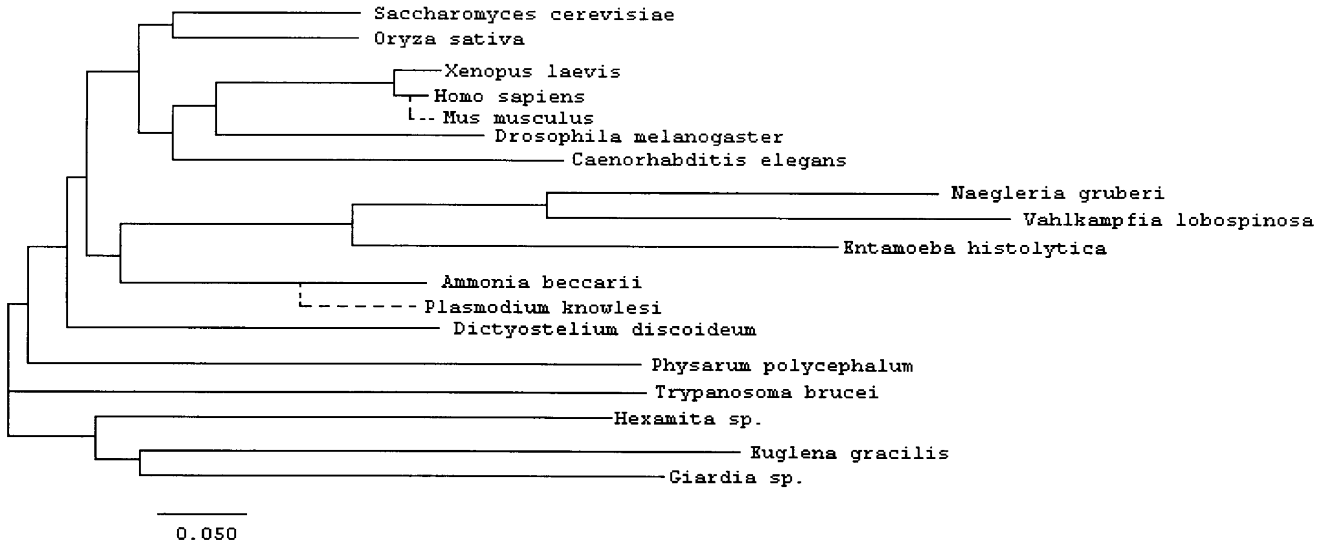

Fig. 1. Tree of the sequences that were used as test cases. The weights for these sequences under different weighting schemes are given in Table 1.

were applied to the sequences in the SSU rRNA alignment, and compared with the default of equal weights.

The first weighting scheme, referred to as tree-based weights, is based on a phylogenetic tree of the sequences in the alignment. A neighbour-joining tree (Saitou and Nei, 1987) of all the sequences in the profile was generated using the DNADIST and NEIGHBOR programs of the PHYLIP package (Felsenstein, 1989). Weights were then derived from the branch lengths as described by Thompson *et al.* (1994b). These weights are then normalized to have a mean of 1.0. This gives a total weight for the profile equal to that where each sequence is weighted equally, which is necessary in order to keep the effects of changing gap penalties congruent across the different schemes. The general effect of these tree-based weights is to downweight sequences with many close relatives in order to prevent the more densely populated regions of the tree exerting a disproportionate effect on the alignment of sequences from other regions of the tree.

The second weighting scheme is based on the level of similarity between the sequence being aligned and each individual sequence in the alignment, and is referred to as identity-based weighting. The new sequence is first aligned with the profile using equal weights. A distance is then calculated between the new sequence and each other sequence in the alignment equal to the mean number of differences per site in this initial approximate alignment. This is percent difference divided by 100 and there is no correction for multiple hits or unequal rates of transition and transversion. The reciprocal of this distance is used as a weight for each sequence and these are again normalized to give a mean of 1.0. This weighting scheme has the effect of upweighting sequences more similar to the sequence being added relative to those that are more distantly related. The upweighting effect increases as the sequences become more similar to the sequence being aligned. The third scheme is a combination of these weighting schemes, in which the weight derived for each sequence based on branch lengths is multiplied by the weight derived from sequence identities, and the values are again renormalized. This scheme is referred to as combination weights.

Table 1 shows the values given by the various weighting schemes for the case shown in the example tree in Figure 1. The tree-based weights are independent of the new sequence that is to be added, being derived wholly from the structure of the existing data. Weights are calculated using the method of Thompson *et al.* (1994b), which are then renormalized to give a mean of 1, leaving the values shown. The identity-based weights are derived by taking the distance of each sequence in the tree from the new sequence, defined as the mean number of differences per aligned pair of residues, ignoring any pairs with a gap in either sequence. The reciprocals of these values are renormalized around 1 to give the figures shown. For the final set of combination weights, the product is taken of the weights in each of the preceding columns and again renormalized to give a mean of 1.

Table 1. The weights assigned to the sequences in the test tree shown in Figure 1 when the sequences *Mus musculus* and *Plasmodium gallinaceae* were added

	a	b	c	d	e	f
Ammonia beccarii	1.000	0.746	0.273	1.256	0.379	0.991
Caenorhabditis elegans	1.000	0.974	0.289	1.008	0.522	1.038
Dictyostelium discoideum	1.000	0.875	0.250	1.049	0.406	0.968
Drosophila melanogaster	1.000	0.727	0.349	1.054	0.470	0.809
Entamoeba histolytica	1.000	1.194	0.225	0.984	0.500	1.241
Euglena gracilis	1.000	1.519	0.198	0.809	0.557	1.298
Giardia sp.	1.000	1.340	0.193	0.773	0.481	1.094
Hexamita sp.	1.000	1.266	0.206	0.854	0.484	1.141
Homo sapiens	1.000	0.411	10.628	1.053	8.088	0.456
Naegleria gruberi	1.000	1.212	0.204	0.942	0.459	1.205
Oryza sativa	1.000	0.511	0.390	1.235	0.370	0.667
Physarum polycephalum	1.000	1.435	0.205	0.856	0.547	1.298
Saccharomyces cerevisiae	1.000	0.516	0.377	1.302	0.361	0.708
Trypanosoma brucei	1.000	1.488	0.211	0.846	0.583	1.329
Vahlkampfia lobospinosa	1.000	1.398	0.196	0.889	0.508	1.313
Xenopus laevis	1.000	0.383	1.798	1.082	1.278	0.438

Columns represent the following schemes: (a) equal sequence weights, (b) tree-based sequence weights, (c) identity-derived weights for each sequence for the alignment of *Mus musculus,* (d) identity-derived sequence weights for each sequence for the alignment of *Plasmodium gallinaceae,* (e) combination of tree and identity-derived weights for *Mus musculus,* (f) combination of tree and identity-derived weights for *Plasmodium gallinaceae*

For each of the three defined weighting schemes and the default of equal weights, alignments were generated using position-specific gap-opening penalties across the range of gap extension penalties and base gap opening penalties described above. This procedure was repeated for each of the test sequences. The number of residues correctly placed in each alignment was determined by comparison with the sequence as originally aligned in the reference alignment, and this was then divided by the total number of residues in the sequence to give a percentage score for the alignment. From the scores for the alignments across the range of gap opening and gap extension penalties for each test case, the gap penalties giving the best performance across all or most of the test cases were obtained.

Implementation

Programs were developed and/or run on DEC Alpha workstations running DEC UNIX. All new code was written in the C programming language and is freely available by anonymous FTP (login as anonymous to chah.ucc.ie and transfer the compressed tar archive PRNAA.tar). The code is not designed for portability and users will have to down load their own rRNA alignments and build their own profiles; a JAVA version of the programs is being developed which will be used to provide future access to all the methods via the Internet.

Results

The performance of a set of weights was judged by its efficacy across the range of gap opening and gap extension penalties used. The peak score and the range of gap penalties giving a comparable score were taken into account in making this judgement (Table 2). For scoring purposes, each residue is counted as distinct, and is only considered correctly aligned if it is in the same position as the same residue in the reference sequence. The score for a sequence is counted as the percentage of the total number of residues in the sequence that have been correctly realigned.

The main results are presented in Table 2. In the first column, the percentage accuracy of alignment scores are given for each of the 16 test cases. These scores are the best obtained across the range of gap opening and extension penalties with no sequence weights. The scores are low and range from 43% (*Euglena*) up to 88% (*Oryza*). The addition of position specific gap penalties has a dramatic effect. The scores all increase by about 10–15% which represents an improvement of several hundred residues in the original sequences that have been correctly aligned. The use of sequence weights yields further improvements, although not as dramatically as this. It should be noted that an improvement in score of just 1% is the equivalent of 20 residues in a molecule of 2000 nucleotides. We only give the peak scores from across the full range of gap opening and extension penalties. These were all obtained with a gap opening penalty of between 5.0 and 7.0 and a gap extension penalty of either 0.1 or 0.2.

Table 2.The highest % identity between the reference alignment and the realigned sequence obtained using each of the weighting schemes

	a	b	c	d	e
A.beccarii	71.65	84.19	83.66	84.05	83.96
C.elegans	69.26	83.98	83.98	86.99	87.84
D.discoideum	64.42	78.95	78.95	79.59	79.06
D.melanogaster	70.14	82.72	82.97	81.11	84.02
E.histolytica	55.68	73.50	74.83	75.04	78.17
E.gracilis	43.12	60.22	60.22	60.22	61.08
Giardia sp.	55.00	73.89	73.96	76.81	77.29
Hexamita sp.	56.13	73.10	73.61	78.39	77.16
H.sapiens	79.88	91.01	92.88	91.49	92.30
N.gruberi	50.37	63.60	63.74	67.81	67.86
O.sativa	88.85	97.08	97.13	96.69	97.35
P.polycephalum	53.62	65.02	64.66	68.64	67.52
S.cerevisiae	86.71	93.94	94.55	93.38	94.10
T.brucei	47.62	62.86	63.39	64.77	65.04
V.lobospinosa	46.23	56.20	55.69	56.20	58.96
X.laevis	82.47	93.59	95.18	94.25	95.07

(a) Fixed gap penalties and equal sequence weights, (b) position-specific gap penalties and equal sequence weights, (c) position-specific gap penalties and identity based weights, (d) position-specific gap penalties and tree-based weights, (e) position-specific gap penalties and combination weights. The underlined values are the absolute maximum scores obtained for each sequence

Table 3. Alignment percentage accuracy scores for various weighting schemes and gap penalties

Gap extension penalty	*Trypanosoma brucei* gap opening penalty									*Vahlkampfia lobospinosa* gap opening penalty								
	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
(a)																		
0.1	46	47	47	47	47	47	47	47	47	45	46	46	46	46	46	46	46	46
0.2	32	32	33	33	33	33	33	33	33	31	33	32	32	32	32	32	32	32
0.3	16	17	17	17	16	16	16	16	16	17	16	16	16	16	16	16	16	16
0.4	13	12	12	12	12	12	12	12	12	10	12	12	12	12	12	12	12	12
0.5	4	4	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6
0.6	2	2	2	2	2	2	2	2	2	5	5	5	5	5	5	5	5	5
0.7	2	2	2	2	2	2	2	2	2	4	4	4	4	4	4	4	4	4
0.8	1	1	1	1	1	1	1	1	1	3	3	3	3	3	3	3	3	3
0.9	1	1	1	1	1	1	1	1	1	3	3	3	3	3	3	3	3	3
(b)																		
0.1	58	59	60	62	62	61	61	61	61	51	53	55	*56*	*56*	*56*	*56*	*56*	*56*
0.2	58	59	61	62	62	62	62	62	62	50	53	54	55	55	55	55	55	55
0.3	59	60	62	*63*	*63*	*63*	*63*	*63*	*63*	51	53	55	55	55	55	55	55	55
0.4	58	59	60	61	*63*	*63*	*63*	*63*	*63*	51	54	54	55	55	55	55	55	55
0.5	58	59	60	61	*63*	*63*	*63*	*63*	*63*	51	53	54	55	55	55	55	55	55
0.6	58	59	61	62	*63*	*63*	*63*	*63*	*63*	51	53	54	55	55	55	55	55	55
0.7	58	60	61	62	62	62	62	62	62	52	53	54	55	55	55	55	55	55
0.8	57	60	61	62	62	62	62	62	62	51	54	54	55	55	55	55	55	55
0.9	57	60	61	62	61	61	61	61	61	51	54	54	55	55	55	55	55	55

Cont....

Table 3. *Continued*

Gap extension penalty	*Trypanosoma brucei* gap opening penalty									*Vahlkampfia lobospinosa* gap opening penalty								
(c)																		
0.1	58	59	60	62	62	61	61	61	61	51	53	55	*56*	*56*	*56*	*56*	*56*	*56*
0.2	59	59	61	62	62	62	62	62	62	50	53	54	55	55	55	55	55	55
0.3	59	60	62	*63*	*63*	*63*	62	62	62	51	53	54	55	55	55	55	55	55
0.4	58	58	60	61	*63*	*63*	62	62	62	51	54	54	55	55	55	55	55	55
0.5	58	59	60	61	*63*	*63*	62	62	62	51	54	54	55	55	55	55	55	55
0.6	58	59	61	61	*63*	*63*	*63*	*63*	*63*	51	53	54	55	55	55	55	55	55
0.7	58	60	61	62	62	62	62	62	62	52	53	54	55	55	55	55	55	55
0.8	57	60	61	62	62	62	61	61	61	51	54	54	55	55	55	55	55	55
0.9	57	60	61	62	62	62	61	61	61	51	54	54	55	55	55	55	55	55
(d)																		
0.1	62	*63*	*65*	*65*	*64*	*64*	*64*	*64*	*64*	51	53	55	*56*	*56*	*56*	*56*	*56*	*56*
0.2	61	*63*	*64*	*64*	*64*	*64*	*64*	*64*	*64*	50	53	54	55	55	55	55	55	55
0.3	60	*63*	*63*	*64*	*64*	*64*	*64*	*64*	*64*	51	53	55	55	55	55	55	55	55
0.4	61	62	*63*	*64*	*64*	*64*	*64*	*64*	*64*	51	54	55	55	55	55	55	55	55
0.5	60	61	62	*64*	*64*	*64*	*64*	*64*	*64*	51	53	54	55	55	55	55	55	55
0.6	59	61	62	*63*	*63*	*63*	*63*	*63*	*63*	51	53	54	55	55	55	55	55	55
0.7	59	61	61	62	61	61	61	61	61	52	53	54	55	55	55	55	55	55
0.8	59	61	61	62	61	61	61	61	61	51	54	54	55	55	55	55	55	55
0.9	59	61	61	61	61	61	61	61	61	51	54	54	55	55	55	55	55	55
(e)																		
0.1	62	*64*	*64*	*65*	*65*	*65*	*65*	*65*	*65*	*56*	*58*	*58*	*58*	*58*	*58*	*58*	*58*	*58*
0.2	60	62	*63*	*64*	*64*	*64*	*64*	*63*	*63*	*56*	*57*	*58*	*58*	*58*	*58*	*58*	*58*	*58*
0.3	61	62	*63*	*64*	*64*	*64*	*64*	*64*	*64*	55	*57*	*57*	*57*	*57*	*57*	*57*	*57*	*57*
0.4	60	62	62	*64*	*64*	*64*	*64*	*64*	*64*	55	*57*	*57*	*57*	*57*	*57*	*57*	*57*	*57*
0.5	60	60	62	*63*	*64*	*64*	*64*	*64*	*64*	55	*57*	*58*	*58*	*58*	*58*	*58*	*58*	*58*
0.6	59	61	*63*	*63*	*63*	62	62	*63*	*63*	55	*57*	*57*	*57*	*57*	*57*	*57*	*57*	*57*
0.7	59	61	62	*63*	*63*	*63*	*63*	*63*	*63*	55	*58*	*58*	*58*	*58*	*58*	*58*	*58*	*58*
0.8	59	61	62	62	62	62	62	62	62	55	*57*	*57*	*57*	*57*	*57*	*57*	*57*	*57*
0.9	59	62	62	62	62	62	62	62	62	55	*57*	*57*	*57*	*57*	*57*	*57*	*57*	*57*

Italics represent those regions at or above the highest score attainable with equal sequence weights. Underlining represents the highest score attained across all the different parameters. Parameter sets are: (a) fixed gap penalties and equal sequence weights, (b) position-specific gap penalties and equal sequence weights, (c) position-specific gap penalties and identity based sequence weights, (d) position-specific gap penalties and tree-derived sequence weights, (e) position-specific gap penalties and weights derived from combination of tree-based and identity-based weights.

In nine out of the 16 test cases, the single best alignment score generated across the ranges of gap penalties was obtained using the combined weights (the last column of Table 2). In three of the remaining cases, tree-based weights give the best performance (column c). The identity weights give the highest score in three cases, and *Ammonia beccarii* is aligned most accurately with equal weights. Both identity-based and tree-based methods of sequence weighting are shown to improve over equal weights in most cases, with the combination of both these weights giving the best overall performance.

Two examples are shown in detail in Table 3. Here the scores for all values of gap opening and gap extension penalties are given for each weighting scheme for just two of the test cases: *Vahlkampfia lobospinosa* and *Trypanosoma brucei*. In both cases, the results with uniform gap penalties, shown in row (a), are very poor and depend strongly on the exact value of the parameters. There is a huge improvement in row (b) where the values for position specific gap penalties are shown. Here, the values are much higher than in row (a) and there is almost no dependence on the exact values chosen for the gap penalties. In the case of *Vahlkampfia* there is no noticeable difference between the use of tree-based or identity-based weights [the results are shown in rows (c), (d) and (b)]. Use of the combined weighting scheme, as seen in row (e), gives a consistent improvement, showing increase of 2% across the entire range of gap penalties. In the case of *Trypanosoma* the relative performance of each weighting scheme is more distinct. In comparing identity weights to equal weights in this case, there is improvement for some values of gap penalty. The effect of using tree-based weights is to produce improvement across a larger range of gap penalties, particularly for gap extension penalties <0.3. The combination of the two weighting schemes again shows a synergistic effect, with a further increase visible across the range of gap penalties.

The values of gap opening and gap extension penalties giving the maximum scores for each test case are given in Table 4. These are the optimum parameters when using the combined weighting scheme with position specific gap penalties. They all fall in a very narrow range.

Table 4. Gap opening and extension penalties giving optimum alignment scores for each test case using combined weights

	Gap opening	Gap extension
A.beccarii	6.0	0.2
C.elegans	6.0	0.1
D.discoideum	6.0	0.1
D. melanogaster	5.0	0.2
E.histolytica	6.0	0.1
E.gracilis	6.0	0.1
Giardia sp.	7.0	0.1
Hexamita sp.	5.0	0.2
H.sapiens	6.0	0.1
N.gruberii	6.0	0.1
O.sativa	6.0	0.1
P.polycephalum	6.0	0.2
S.cerevisiae	6.0	0.2
T.brucei	6.0	0.1
V.lobospinosa	6.0	0.1
X.laevis	6.0	0.2

In order to tell which sections of the reference alignment may be reliably aligned, each of the 1517 sequences in turn was removed from the alignment and re-aligned with the remaining sequences. Each column of the original, reference alignment was scored depending on what percentage of the residues in it can be realigned in the correct positions. Figure 2 shows the estimated secondary structure of the *Saccharomyces cerevisiae* nuclear SSU rRNA with those positions from the full alignment which can be realigned with ≥95% accuracy marked in black and those which realign with <95% accuracy in grey. Stems forming pseudoknots are not displayed in this representation. This is a conservative estimate of the regions that may be reliably aligned as there are some positions that are not found in this molecule and sequences from some taxonomic groupings may be aligned almost perfectly.

Figure 3 shows the accuracy with which each sequence can be realigned compared to its original alignment as a function of G+C content. The re-alignment accuracy is greatest for sequences with average G+C contents (~50%). As expected, sequences with extreme nucleotide compositions (very high or very low G+C content) tend to be less easy to align accurately. High levels of a particular nucleotide increase the chance that a residue in the sequence being aligned may align with the wrong column in the profile. The test cases cover a range of G+C content from 38.4% (*Entamoeba histolytica*) to 68.5% (*Giardia sp.*).

Discussion

The generation of alignments under various parameters shows that position-specific gap opening penalties have a very strong positive effect on the accuracy with which alignments can be generated. Fixed gap penalties perform extremely poorly, particularly at high values of gap extension penalty. This corresponds to situations in which the long gaps that occur in virtually all sequences in certain regions of the alignment, which correspond to long insertions in a few sequences, are penalized very heavily and do not occur in an alignment giving an optimum score. Experimentation with position-specific gap extension penalties did not give any further improvement.

Sequence weighting can have a further positive effect on alignment quality. Both weighting schemes based on sequence identity and those based on the tree structure and branch lengths are seen to have generally positive effects. As expected, the tree-based weights are seen to perform at their best in the case of sequences which are quite distant from the main taxa, with few or no close relatives, such as *Hexamita*, and to be of least benefit to alignment quality with sequences which have many close relatives such as *O.sativa*. With identity-based weights the greatest positive effects are seen in sequences within highly represented taxa such as *S.cerevisiae.*

These two weighting schemes have opposing effects on the values of the sequence weights in the case of sequences aligning into densely populated regions of the tree, and so the net result of combining them, in cases such as *S.cerevisiae*, may not perform any better than either of the weighting schemes used individually. The examples given (Table 3) indicate that there are cases where tree-based and identity-based weights show a synergistic effect when combined, the combination outperforming either of the schemes applied individually. The combined weights give the best result in more than half of the test cases, and the average difference between the score generated with the combined weights and the overall best score is substantially less than the difference between the scores from any of the other weighting schemes and the overall best score in each case. This synergy is seen to occur most strongly in sequences which are distant from the main bulk of the alignment and therefore more difficult to align correctly. Those which are located in highly repre-

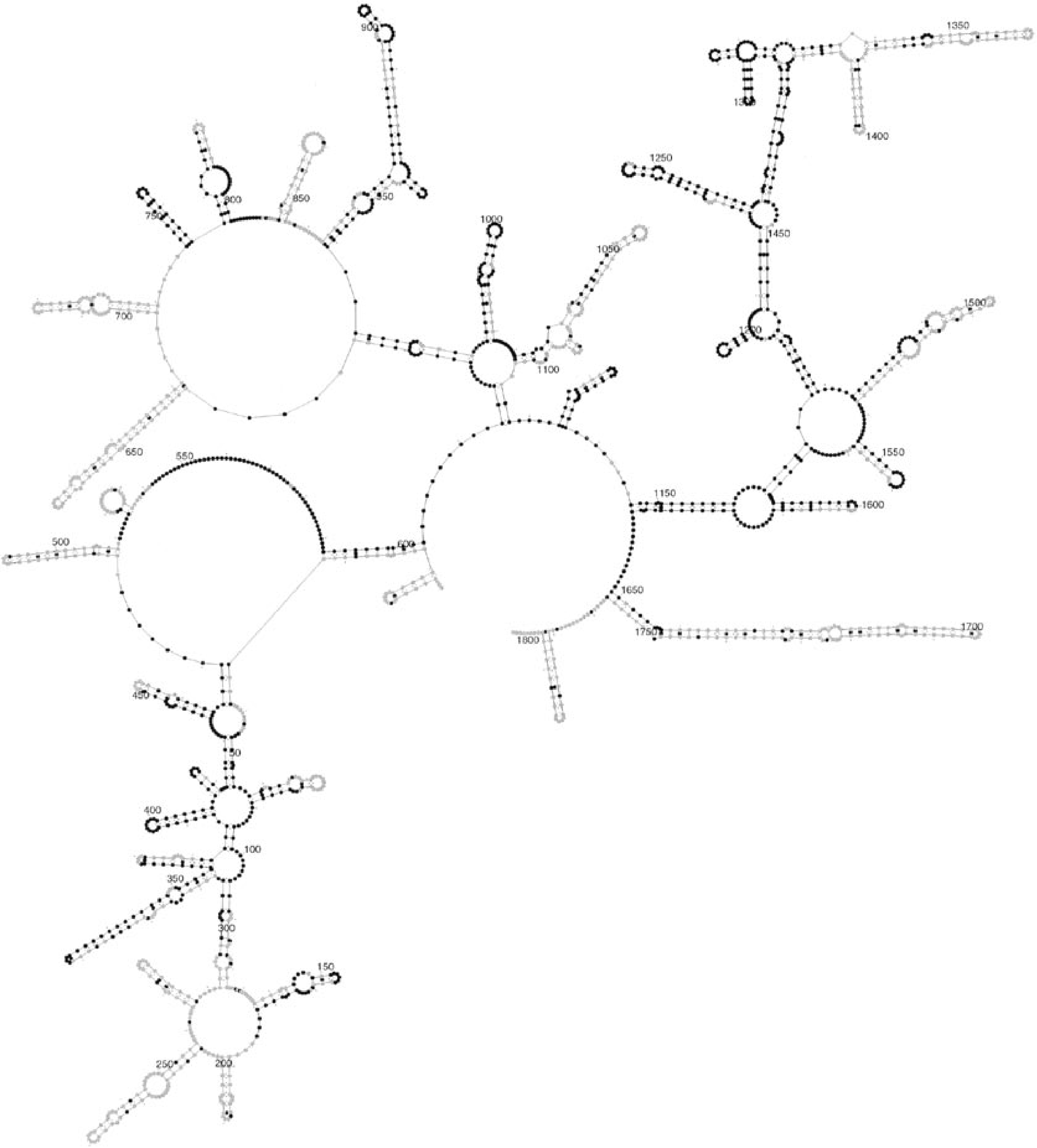

Fig. 2. Secondary structure of *Saccharomyces cerevisiae* SSU rRNA with stable regions indicated in black., generated using the ESSA program (Chetouani *et al.*, 1997).

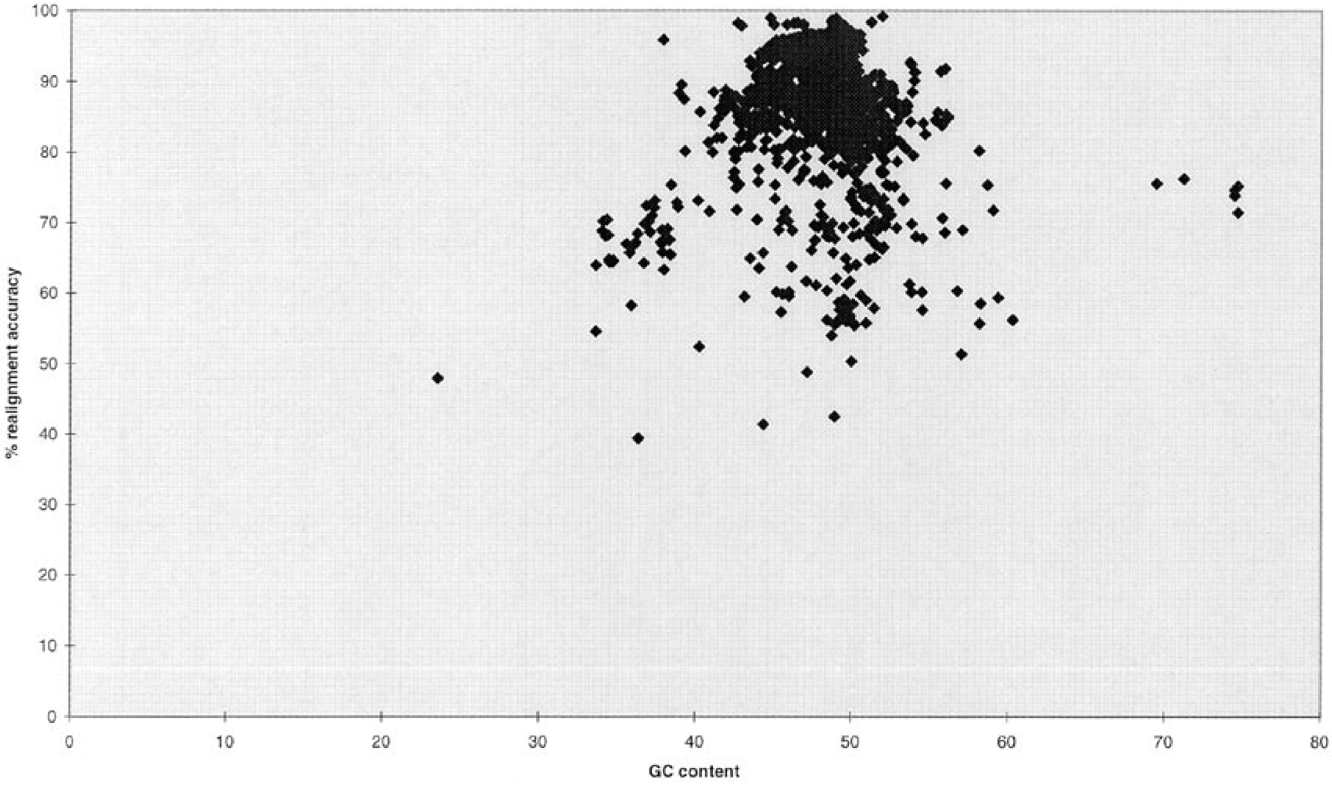

Fig. 3. Graph of percentage of sequence correctly re-aligned against G+C content for each of the 1517 sequences in the reference alignm ent.

sented taxa do not show such strong effects from any of the weighting schemes, but these tend to be those sequences which have the best alignments initially.

We have shown how to improve the accuracy of alignment of rRNA sequences using some simple methods. It is quite possible that alignments of 100% accuracy will not be possible due to the existence of errors introduced manually into the reference alignment. Nonetheless, we can already see that some sequences may be aligned with >95% accuracy (*Oryza* and *Xenopus*), and across the entirety of the alignment 89.84% of all residues can be realigned correctly. Some sequences are still disappointing and this can partially be explained by very biased G+C content (e.g. *Giardia*). Others come from poorly sampled parts of the overall Eukaryote phylogenetic tree and these will become easier to align as new sequences are added. Nonetheless, it may be difficult for users to evaluate the quality of a new alignment. We provide one, extremely simple method for choosing regions of the overall alignment that can be reliably aligned in almost all cases. This covers about half of the positions in any given molecule and provides a selection of sites which can be reliably chosen for phylogenetic purposes. This site selection can be fine-tuned by looking at regions which may be reliably aligned in specific taxa.

Finally, it is very obvious that these methods could benefit from some consideration of secondary structure, which could be used for evaluation of alignments or as part of the alignment process. We are investigating the use of genetic algorithms to optimize the quality of profile alignments where secondary structure is considered (Notredame *et al.*, 1997). We will use a genetic algorithm to optimize the quality function of Corpet and Michot (1994) but based on profiles rather than pairs of sequences.

Acknowledgements

The authors thank Richard Durbin for suggesting the use of the 1/d weights. We also thank Manolo Gouy for his help with rRNA sequences in general. This work was supported by a grant (BIO4-CT95–0130) from the EU Biotechnology programme.

References

Chetouani,F.,Monestie,P.,Thebault,P.,Gaspin,C. and Michot,B. (1997) ESSA: an integrated and interactive computer tool for analysing RNA secondary structure. *Nucleic Acids Res.*, **25**, 2514–3522.

Corpet,F. and Michot,B. (1994) RNAlign program: alignment of RNA sequences using both primary and secondary structures. *Comput. Applic. Biosci.*, **10**, 389–399.

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

Felsenstein,J. (1989) *Cladistics*, **5,** 164–166.

Gotoh,O. (1982) *J. Mol. Biol.*, **162**, 705.

Gotoh,O. (1995) A weighting system and algorithm for aligning many phylogenetically related sequences. *Comput. Applic. Biosci.*, **11**, 543–551.

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

Henikoff,J. and Henikoff,S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput. Applic. Biosci.*, **12**, 135–143.

Luthy,R., Xenarios,I., and Bucher,P. (1994) Improving the sensitivity of the sequence profile method. *Protein Sci.*, **3**, 139–146.

Maidak,B., Olsen,G.,Larsen,N., Overbeek,R.,McCaughey,M. and Woese,C. (1997) The Ribosomal Database Project (RDP). *Nucleic Acids Res.*, **25**, 109–111.

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.

Neefs,J.-M., Van de Peer,Y., Hendriks,L. and De Wachter,R. (1990) Database on the structure of small subunit ribosomal RNA. *N ucleic Acids Res.*, **18**, 2237–2217.

Notredame,C., O'Brien,E.A. and Higgins,D.G. (1997) RAGA: RNA alignment by genetic algorithm. *Nucleic Acids Res.*, **25**, 4570–4580.

Pawlowski,J., Bolivar,I., Fahrni,J.F., Cavalier-Smith,T. and Gouy,M. (1996) Early origin of Foraminifera suggested by SSU rRNA gene sequences. *Mol. Biol. Evol.*, **13**, 445–450.

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.

Sogin,M.,Elwood,H, and Gunderson,J. (1986) Evolutionary diversity of eukaryotic small-subunit rRNA genes. *Proc. Natl Acad. Sci. USA*, **83**, 1383–1387.

Thompson,J., Higgins,D. and Gibson,T. (1994a) CLUSTAL W: improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

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

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.