

Contents

1 Introduction	3
2 The new algorithm	3
3 Benchmark: BAliBASE	4
4 Benchmark: BRaliBase II	5
5 Benchmark: Large datasets	7
References	8

1 Introduction

The alignment algorithms used in the software from CLC bio A/S have some unique features including the option of adjusting the cost of gaps in the end of the alignment to suit the sequences being aligned. Until now, however, the algorithms have also been relatively slow and not as accurate as the leading alignment programs.

The users of our programs have pointed the shortcoming of our alignment algorithms out on a number of occasions. This has led us to put a serious effort into improving our performance. Here we present a note regarding these improvements.

We have now two alignments: A standard algorithm that is 10 times faster than our previous alignment in most scenarios, and an additional alignment that is even faster, but less accurate than our standard algorithm.

Both alignments are available in all the CLC Workbenches.

Alignments of large data sets

On large data sets of sequences that are not too divergent, our new alignment is significantly faster than the standard CLUSTAL W alignment, and around the same speed as the fast CLUSTAL W alignment.

Performing an alignment of 28 HIV genomes, our fast alignment is more than 10 times (55 minutes) faster than the standard CLUSTAL W, cf. table 3 below.

On very divergent data sets, CLUSTAL W is faster than our new alignment.

Protein alignments We have benchmarked our new algorithms on the BALiBASE 3.0 database of accurate protein alignments [Thompson et al., 2005]. In summary, the result is an alignment algorithm that is about 1% more accurate than the latest version of the standards CLUSTAL W on protein alignments.

RNA alignments We have also used the BRaliBase II database of structurally aligned RNA sequences to evaluate the performance on nucleotide sequences. Here, our new algorithm is about 3.5% more accurate than the standard CLUSTAL W.

Our standard algorithm is still a little slower than the standard CLUSTAL W on the fairly divergent alignments in BALiBASE and BRaliBase.

Our fast alignment is as precise and as fast as the standard CLUSTAL W on these data sets.

All tests were made on a Linux Fedora Core 4 Dell D610 laptop with 1 GB ram, and a 2.0 GHz Intel Pentium M processor.

2 The new algorithm

Like most other multiple alignment algorithms, our new algorithm is based on progressive alignment [Feng and Doolittle, 1987]. This method is based on combining sequences into alignments, which can in turn be combined with other sequences or alignments to form larger alignments. The procedure is repeated until all the input sequences have been joined in a single multiple alignment.

The method has the inherent drawback that once two sequences are aligned, there is no way of changing their relative alignment based on the information that additional sequences may contribute later in the process.

It is therefore important to make the best possible alignments early in the procedure, to avoid accumulating

errors. To accomplish this, a tree of the sequences is usually made, which guides the progressive alignment algorithm by following the tree from the leaves to the root when deciding which sequences to align.

For many algorithms, the neighbor joining method or the UPGMA method is used to create the guide tree based on all pairwise distances between the sequences to be aligned. This generally requires that all pairwise alignments are found between the input sequences to determine their distances. For data sets with many sequences, this step may take much longer than the actual progressive alignment.

To overcome this problem of the very time consuming tree construction, we have applied a method of word matching, which can group sequences in a very efficient way, saving much time, without reducing the resulting alignment accuracy significantly.

Our algorithm has two speed settings: 'standard' and 'fast'. The standard method makes a fairly standard progressive alignment using the fast method of generating a guide tree. When aligning two alignments to each other, two matching columns are scored as the average of all the pairwise scores of the residues in the columns. The gap cost is affine, allowing a different cost for the first gapped position and for the consecutive gaps. This ensures that gaps are not spread out too much.

The fast method of alignment uses the same overall method, except that it uses fixpoints in the alignment algorithm based on short subsequences that are identical in the sequences that are being aligned. This allows similar sequences to be aligned much more efficiently, without reducing accuracy very much.

3 Benchmark: BALiBASE

To test the performance of our algorithm on protein alignments, we use the BALiBASE database of accurate alignments [Thompson et al., 2005]. The database is distributed with a program for comparing an alignment to the 'correct' alignment in the database. Table 1 shows the results for our algorithm and for CLUSTAL W, divided on six different categories of alignments from BALiBASE:

RV11 Equi-distant sequences, very divergent (<20% identity).

RV12 Equi-distant sequences, medium to divergent (20–40% identity).

RV20 Families aligned with a highly divergent "orphan" sequence.

RV30 Subgroups with <25% residue identity between groups.

RV40 Sequences with N/C-terminal extensions.

RV50 Internal insertions.

The results are also divided between pairwise accuracy (SP) and total column accuracy (TC), and between the entire alignment and the core regions only.

Our new algorithm on the standard setting is about 1% better than CLUSTAL W as an average over the six alignment categories and the various accuracy is measures. The algorithm is generally a little slower than CLUSTAL W on these alignments.

On the fast setting, our new algorithm is still slightly better than CLUSTAL W on average. The speedup is not very large, however. This is because the quick progressive alignment method is only used when the algorithm finds enough fixpoints between two sequences or alignments to be relatively sure that they are correct. In the case of BALiBASE, the data are fairly divergent and the quick method is not used that often by the algorithm.

Table 1: Top: data set sizes for BALiBASE 3.0. Below: the result of the benchmark on these data. The accuracy scores are sum of pairs (SP) and total column accuracy (TC).

BALiBASE 3.0 data			
Data set	Number of alignments	Sequences per alignment	Average seq. length
RV11	38	6.87	241
RV12	44	9.00	303
RV20	41	45.59	245
RV30	30	63.17	261
RV40	49	27.63	466
RV50	15	29.20	327

CLC bio, standard						CLC bio, fast					
Data set	Time	SP	TC	Core		Data set	Time	SP	TC	Core	
				SP	TC					SP	TC
RV11	1.03s	55.4%	31.5%	66.2%	43.5%	RV11	1.03s	54.2%	29.8%	64.7%	40.6%
RV12	2.41s	83.1%	64.8%	90.5%	77.5%	RV12	1.17s	81.8%	62.4%	89.3%	74.6%
RV20	7.62s	86.5%	32.9%	92.0%	45.5%	RV20	3.13s	86.3%	32.6%	91.9%	45.1%
RV30	11.73s	73.4%	33.8%	83.1%	50.8%	RV30	4.97s	73.3%	33.0%	83.2%	50.4%
RV40	34.37s	70.0%	29.9%	82.5%	42.0%	RV40	29.21s	69.6%	28.1%	81.9%	39.5%
RV50	8.67s	71.2%	30.8%	82.9%	45.9%	RV50	5.04s	70.4%	30.0%	82.0%	44.3%
Avg.	10.97s	73.2%	37.3%	82.9%	50.9%	Avg.	7.42s	72.6%	36.0%	82.2%	49.1%

CLUSTAL W v. 1.83, standard						CLUSTAL W v. 1.83, fast					
Data set	Time	SP	TC	Core		Data set	Time	SP	TC	Core	
				SP	TC					SP	TC
RV11	0.28s	55.1%	28.6%	64.7%	39.6%	RV11	0.17s	55.1%	29.9%	64.0%	40.7%
RV12	0.89s	84.0%	67.9%	90.3%	78.9%	RV12	0.37s	84.9%	67.9%	91.2%	79.4%
RV20	10.13s	87.5%	32.6%	92.3%	44.2%	RV20	1.66s	88.0%	34.8%	92.6%	47.0%
RV30	17.84s	72.6%	32.2%	81.2%	47.6%	RV30	2.94s	74.0%	34.4%	82.6%	51.1%
RV40	11.49s	69.6%	28.9%	79.2%	39.6%	RV40	4.93s	67.3%	26.9%	76.3%	36.5%
RV50	6.34s	70.6%	30.5%	79.9%	43.9%	RV50	1.69s	71.5%	30.9%	81.1%	42.6%
Avg.	7.83s	73.2%	36.8%	81.3%	49.0%	Avg.	1.96s	73.4%	37.5%	81.3%	49.6%

4 Benchmark: BRaliBase II

To test the accuracy of our alignment algorithms on nucleotide data, we use the BRaliBase II database of structural RNA alignments, made specifically for benchmarking alignment methods [Gardner et al., 2005]. The results of the alignments are shown in table 2. The data consists of six groups of alignments:

SRP A set of SRP RNA alignments.

U5 A set of U5 RNA alignments.

g2intron A set of group II intron alignments.

rRNA A set of 5S ribosomal RNA alignments.

Table 2: Top: data set sizes for BRaliBase II. Below: the result of the benchmark on these data. The accuracy scores are sum of pairs (SP) and total column accuracy (TC). No core regions were identified in the data sets.

BRaliBase II data			
Data set	Number of alignments	Sequences per alignment	Average seq. length
SRP	93	5	300
U5	109	5	118
g2intron	92	5	80
rRNA	89	5	117
tRNA	98	5	72
tRNA pair	118	2	76

CLC bio, standard				CLC bio, fast			
Data set	Time	SP	TC	Data set	Time	SP	TC
SRP	0.71s	86.4%	75.0%	SRP	0.27s	86.2%	74.7%
U5	0.09s	81.5%	66.6%	U5	0.06s	81.5%	66.7%
g2intron	0.05s	76.8%	63.9%	g2intron	0.04s	76.8%	63.9%
rRNA	0.09s	93.5%	87.5%	rRNA	0.06s	93.4%	87.2%
tRNA	0.04s	90.8%	83.2%	tRNA	0.03s	90.8%	83.2%
tRNA pair	0.00s	71.9%	71.5%	tRNA pair	0.01s	71.7%	71.3%
Average	0.16s	83.5%	74.6%	Average	0.08s	83.4%	74.5%

CLUSTAL W v. 1.83, standard				CLUSTAL W v. 1.83, fast			
Data set	Time	SP	TC	Data set	Time	SP	TC
SRP	0.20s	87.3%	76.4%	SRP	0.13s	87.4%	76.8%
U5	0.03s	79.5%	65.1%	U5	0.03s	79.7%	65.5%
g2intron	0.02s	72.7%	61.5%	g2intron	0.02s	72.3%	60.0%
rRNA	0.04s	93.3%	86.8%	rRNA	0.03s	93.6%	87.2%
tRNA	0.02s	86.9%	76.1%	tRNA	0.01s	84.5%	75.4%
tRNA pair	0.01s	60.1%	59.8%	tRNA pair	0.01s	60.1%	59.8%
Average	0.05s	80.0%	71.0%	Average	0.04s	79.6%	70.8%

tRNA A set of tRNA alignments.

tRNA pair Pairwise alignments of tRNAs.

Again, our algorithms are a little slower than CLUSTAL W, but are 3.5% more accurate. The data sets are relatively small, so it takes very little time to construct each alignment.

5 Benchmark: Large datasets

A special challenge for alignment algorithms is that of aligning very large data sets. To benchmark our performance on such data, we chose three alignments from public databases:

HIV 28 HIV genomes, all from different subtypes (<http://hiv-web.lanl.gov>).

CA The full alignment of all the 410 carbonic anhydrase proteins from PFAM (PF00194) [Bateman et al., 2004].

KT The full alignment of all the 220 potassium transporter proteins from PFAM (PF02705).

For the protein data sets with many sequences, our algorithms are much faster than CLUSTAL W, both in the fast and standard setting. For the HIV data set, our fast algorithm is faster than CLUSTAL W.

The accuracies with our algorithms is higher than for CLUSTAL W on these three tests, but the data sets may not specifically have been made with alignment accuracy in mind.

Table 3: Top: data set sizes for the large data set benchmark. Below: the result of the benchmark on these data. Again, the accuracy scores are sum of pairs (SP) and total column accuracy (TC). No core regions were identified in the data sets.

Large data sets		
Data set	Sequences	Average seq. length
HIV	28	9,005
CA	410	190
KT	220	594

HIV			
Program	Time	SP	TC
CLC bio, standard	5,235s	97.5%	84.0%
CLC bio, fast	287s	97.4%	84.0%
CLUSTAL W, standard	3,625s	97.6%	84.0%
CLUSTAL W, fast	715s	97.4%	84.0%

Carbonic anhydrases			
Program	Time	SP	TC
CLC bio, standard	78s	90.2%	26.0%
CLC bio, fast	62s	82.8%	26.0%
CLUSTAL W, standard	247s	79.7%	0.0%
CLUSTAL W, fast	31s	82.4%	0.0%

Potassium transporters			
Program	Time	SP	TC
CLC bio, standard	255s	85.3%	57.0%
CLC bio, fast	98s	84.8%	57.0%
CLUSTAL W, standard	759s	85.3%	57.0%
CLUSTAL W, fast	65s	85.1%	57.0%

References

- [Bateman et al., 2004] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res.*, 32(Database issue):D138–D141.
- [Feng and Doolittle, 1987] Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, 25(4):351–360.
- [Gardner et al., 2005] Gardner, P. P., Wilm, A., and Washietl, S. (2005). A benchmark of multiple sequence alignment programs upon structural rnas. *Nucleic Acids Res.*, 33(8):2433–2439.
- [Thompson et al., 2005] Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). Balibase 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, 61(1):127–136.