MUSCLE: Low-complexity multiple sequence alignment with T-Coffee accuracy

Robert C. Edgar

195 Roque Moraes Drive, Mill Valley, CA 94941, U.S.A.

Received line

ABSTRACT

We describe MUSCLE, a new program for creating multiple alignments of protein sequences. MUSCLE achieves the highest score so far reported on the BAliBASE benchmark, with average accuracy statistically indistinguishable from T-Coffee. MUSCLE aligns 5,000 sequences of average length 350 in 7 minutes on a current desktop computer, requiring less time than all other tested methods, including MAFFT. We also introduce PREFAB, a new multiple alignment benchmark. PREFAB results confirm that MUSCLE and T-Coffee produce, on average, the most accurate alignments, with 6% more positions correctly aligned than ClustalW. Software, source code and test data is freely available at: http://www.drive5.com/muscle.
Contact: bob@drive5.com

INTRODUCTION

Multiple sequence alignments (MSAs) of protein sequences are important in many applications, including phylogenetic tree estimation, structure prediction and critical residue identification. MSA algorithms are typically evaluated on the BAliBASE benchmark (Thompson *et al.*, 1999; Bahr *et al.*, 2001), on which T-Coffee (Notredame *et al.*, 2000) achieves the highest score previously reported. High-throughput methods are becoming increasingly important due to the continuing rapid growth in biological sequence databases; however, the best current methods have relatively high complexity and can typically align no more than a few tens or hundreds of sequences on current desktop computers. Here, we describe MUSCLE (Edgar, 2004c) (multiple sequence comparison by log expectation), a new MSA method that achieves average accuracy equal to T-Coffee. For large numbers of sequences, MUSCLE is faster than the FFTNS-1 script in the MAFFT suite (Katoh *et al.*, 2002), the fastest previously reported method known to the author. We note that several methods, including MUSCLE, have used BAliBASE as training data for gap penalties and other parameters, and therefore introduce an alternative alignment accuracy benchmark test, PREFAB.

MUSCLE ALGORITHM

Input to the algorithm is a set of N unaligned amino acid sequences of typical length L. The evolutionary distance between each pair of sequences is estimated in O(L) time and space by computing the fraction of common k-mers (substrings of length k) in a compressed amino acid alphabet, a measure that we have previously shown to correlate well with fractional identity and to be substantially faster than methods requiring a pairwise alignment (Edgar, 2004a). The resulting distance matrix is clustered using UPGMA (Sneath and Sokal, 1973), giving a binary tree. The tree is then used to construct a progressive alignment (Hogeweg and Hesper, 1984; Feng and Doolittle, 1987) by aligning profiles of the two subtrees at each internal node. This requires a scoring function for a pair of profile positions, analogous to the substitution matrix score for a pair of residues; see e.g. (Mittelman *et al.*, 2003; Edgar and Sjolander, 2004). Let i and j be amino acid types, p_i be the background probability of i, p_{ij} the joint probability of i and j being aligned to each other, f^x_i be the observed frequency of i in column x of the first profile, f^x_G be the observed frequency of gaps in that column at position x in the family. (Similarly for position y in the second profile.) MUSCLE uses the following function:

$$\mathrm{LE}^{xy} = (1-f^x_G)\,(1-f^y_G)\,\log \Sigma_i \Sigma_j f^x_i f^y_j \, p_{ij} / p_i p_j. \qquad (1)$$

This *log-expectation* (LE) score is a modified form of the log-average function previously proposed on theoretical grounds (von Ohsen and Zimmer, 2001). The LE form was discovered to provide a better estimate than other tested functions of the probability that a pair of columns is correctly aligned, as assessed on a database in which columns created by PSI-BLAST (Altschul *et al.*, 1997) were aligned to each other by reference to FSSP (unpublished). Probabilities p_i and p_{ij} are derived from the VTML 240 matrix (Muller *et al.*, 2002). Frequencies f_i must be normalized to sum to one if indels are present (otherwise the logarithm becomes increasingly negative with increasing numbers of gaps even when aligning conserved or similar

residues). The factor $(1–f_G)$ is the *occupancy* of a column. The occupancy factors are introduced to encourage more highly occupied columns to align. Our profile parameters, which include residue frequencies and gap frequencies (opens, closes and extensions) at each position, together with position-specific gap penalties, allow the profile of a pairwise profile alignment to be computed in O(L) time from the trace-back path and input profiles (Edgar, 2004b). This avoids the conventional step of building an explicit multiple alignment in order to compute the new profile, which is an O(NL) procedure that becomes expensive when $N >> 20$. The final multiple alignment at the root node is recovered in O($NL \log N$) time by storing the trace-back path at internal nodes and traversing the path from each leaf (input sequence) to the root. From this "first draft" multiple alignment, the fractional identity of each pair of sequences is computed and converted to an additive distance estimate by correcting for multiple substitutions at a single site (Kimura, 1983). This gives a new distance matrix which is clustered by UPGMA, yielding a new tree which is, on average, closer to optimal than the first tree produced by *k*-mer counting. The branching orders of the old and new trees are then compared using an O(N) algorithm. Profiles of subtrees having unchanged branching orders are retained, and a progressive alignment over the (possibly empty) set of changed nodes is constructed, yielding the final alignment of the sequences. The time complexity of MUSCLE is $O(N^2L + NL^2)$, space complexity is $O(N^2 + NL + L^2)$.

PREFAB BENCHMARK

A test set in PREFAB is constructed from a pairwise structural alignment. Each sequence is used to query a database, from which high-scoring hits are collected. The two sequences and their hits are combined and aligned by an MSA program. Accuracy is assessed on the original pair alone, by comparison with their structural alignment. We used a set of pairwise structural alignments from (Sadreyev and Grishin, 2003) (data kindly provided by Ruslan Sadreyev). These authors selected 500 families at random from the FSSP database (Holm and Sander, 1998). Within each family, they chose three pairs of structures at random from the sequence identity ranges 0-15%, 15-30% and 30-97%, giving a total of 1,484 pairs. We used each full-chain sequence (not restricted to its aligned region) to make a PSI-BLAST search of the NCBI non-redundant protein sequence database (Pruitt *et al.*, 2003), keeping the locally aligned regions of hits with e-values below 0.01. Hits were filtered to 80% maximum identity (including the query), and 24 selected at random. Finally, the original pair and their remaining hits were combined to make a set of ≤50 sequences. The limit of 50 was arbitrarily chosen to make the test tractable for some of the more resource-intensive methods, in particular T-Coffee (which needed 10 CPU days, as noted in Table 1). Input sets in PREFAB average 49 sequences of length 242.

RESULTS

We compared the accuracy and speed of MUSCLE with T-Coffee, which achieves the highest previously published BAliBASE score; ClustalW, probably the most popular program; and the MAFFT script FFTNS-1, the fastest previously published method known to the author. Tested versions were MUSCLE 2.1, T-Coffee 1.37, ClustalW 1.82 and MAFFT 3.82. Benchmarks were PREFAB version 1.0 and references 1 through 5 in BAliBASE version 2. Alignment quality is measured by Q_P, the number of correctly aligned pairs of residues divided by the number of aligned pairs in the reference alignment. CPU times were measured on a 2.5 GHz Pentium 4 desktop computer. Results are summarized in Table 1.

	BAliBASE		PREFAB	
Method	**Q_P**	**CPU secs**	**Q_P**	**CPU secs**
MUSCLE	0.884	20	0.496	980
T-Coffee	0.882	1,500	0.496	860,000
ClustalW	0.860	170	0.465	18,000
FFTNS-1	0.844	16	0.479	590

Table 1. Summary of benchmark test results.
Here we show Q_P scores averaged over each reference database, together with the CPU time in seconds. T-Coffee required approximately 10 days to complete the PREFAB test, compared with 16 minutes for MUSCLE.

Non-parametric rank tests show MUSCLE and T-Coffee to be indistinguishable in accuracy on these benchmarks, but determine both programs to rank higher than ClustalW and FFTNS-1. For example, MUSCLE is superior to ClustalW on PREFAB with $p=10^{-5}$. To investigate resource requirements for increasing N, we used the Rose sequence generator (Stoye *et al.*, 1998). In agreement with other studies, e.g. (Katoh *et al.*, 2002), we found that T-Coffee is unable to align more than ~100 sequences of typical length on a current desktop computer. ClustalW is able to align a few hundred sequences, with a practical limit around $N \sim 10^3$ where CPU time begins to scale approximately as N^4. Our largest set had 5,000 sequences of average length 350. MUSCLE completed this test in 7 minutes, compared with 10 minutes for FFTNS-1. We project that ClustalW would need approximately one year.

DISCUSSION

MUSCLE achieves average accuracy equal to the most accurate previous method (T-Coffee) with execution times comparable or better than the fastest previous method (FFTNS-1). MUSCLE consistently achieves both higher

accuracy and substantially shorter execution times than the most widely used method (ClustalW). MUSCLE can align thousands of protein sequences in minutes, enabling high-throughput, high-accuracy applications. High accuracy is obtained by a purely progressive algorithm without a subsequent refinement stage. This is possibly due to the use of the log-expectation profile function (Equation 1) rather than the sum-of-pairs heuristic used by several other programs, including ClustalW and MAFFT. We found UPGMA clustering to give slightly better benchmark results than neighbor-joining, despite the expectation that neighbor-joining will tend to give a better estimate of the evolutionary tree; see e.g. (Felsenstein, 2004). This may be explained by assuming that in a set of related profiles, the two that can be aligned with fewest errors is the pair with fewest differences, even if they are not evolutionary neighbors. It is therefore plausible that nearest-neighbor clustering can give a more accurate progressive alignment than the true evolutionary tree. High speed is achieved by combining several techniques. Compared with ClustalW, the most important of these is the reduced cost of guide tree construction. The initial distance measure is computed by *k*-mer counting, which is $O(N^2L)$. In contrast, the distance matrix computation in CustalW is $O(N^2L^2)$ due to the creation of alignments for every pair of input sequences. With $L \sim 10^2$, combined with an additional factor of ~10 due to the complication of dynamic programming compared with *k*-mer counting, this typically gives a three orders of magnitude reduction in the time cost of computing distances[1]. Clustering of the distance matrix is performed by an $O(N^2)$ implementation of UPGMA, a significantly lower time complexity than the neighbor-joining implementation in ClustalW, which is $O(N^4)$. ClustalW 1.8 was trained on BAliBASE (Julie Thompson, personal communication), as was MUSCLE 2.1. While the use of BAliBASE for both training and assessment is questionable, we find that BAliBASE in fact gives similar rankings of MSA methods to PREFAB, with the notable exception of ClustalW. Despite a high rank on BAliBASE, ClustalW scores lower on PREFAB than all programs we tried, other than POA (Lee *et al.*, 2002) (average Q_P=0.41). On average, ClustalW aligns 6% fewer PREFAB positions correctly than MUSCLE and 3% fewer than FFTNS-1. This suggests that ClustalW, which incorporates several heuristics and hence a relatively large number of parameters, may be over-tuned to BAliBASE. MUSCLE and PREFAB are freely available at: http://www.drive5.com/muscle.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17): 3389-402.

Bahr, A., Thompson, J.D., Thierry, J.C. and Poch, O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res* **29**(1): 323-6.

Edgar, R.C. (2004a) Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Res* **32**(1): 380-5.

Edgar, R.C. (2004b) Low complexity multiple sequence alignment. (Submitted).

Edgar, R.C. (2004c) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*: (in press).

Edgar, R.C. and Sjolander, K. (2004) A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics* (in press).

Felsenstein, J. (2004) *Inferring Phylogenies*. Sunderland, Massachusetts, Sinauer Associates.

Feng, D.F. and Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* **25**(4): 351-60.

Hogeweg, P. and Hesper, B. (1984) The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *J Mol Evol* **20**(2): 175-86.

Holm, L. and Sander, C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res* **26**(1): 316-9.

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**(14): 3059-66.

Kimura, M. (1983) *The neutral theory of molecular evolution*, Cambridge University Press.

Lee, C., Grasso, C. and Sharlow, M.F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**(3): 452-64.

Mittelman, D., Sadreyev, R. and Grishin, N. (2003) Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. *Bioinformatics* **19**(12): 1531-9.

Muller, T., Spang, R. and Vingron, M. (2002) Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol Biol Evol* **19**(1): 8-13.

Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**(1): 205-17.

Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2003) NCBI Reference Sequence project: update and current status. *Nucleic Acids Res* **31**(1): 34-7.

Sadreyev, R. and Grishin, N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* **326**(1): 317-36.

Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical taxonomy*. San Francisco, Freeman.

Stoye, J., Evers, D. and Meyer, F. (1998) Rose: generating sequence families. *Bioinformatics* **14**(2): 157-63.

Thompson, J.D., Plewniak, F. and Poch, O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15**(1): 87-8.

von Ohsen, N. and Zimmer, R. (2001) Improving profile-profile alignment via log average scoring. *Algorithms in Bioinformatics, First International Workshop, WABI 2001*. O. Gascuel and B. M. E. Moret. Berlin, Springer-Verlag. **2149:** 11-26.

[1] The *–quicktree* option of ClustalW gives ~3× faster distance calculation, but is sometimes less accurate and remains $O(N^2L^2)$.