

MUSCLE: Multiple sequence alignment with improved accuracy and speed

Robert C. Edgar

Dept. of Plant and Microbial Biology, University of California, Berkeley

bob@drive5.com

Abstract

We present MUSCLE, a new program for creating multiple alignments of protein sequences. MUSCLE achieves the highest scores so far reported on four alignment benchmarks: Balibase, PREFAB, SABmark and SMART, achieving accuracy from 1% to 2.5% higher than T-Coffee and execution times that are generally lower than CLUSTALW for typical input data. With options designed for high-throughput applications, MUSCLE gives average accuracy statistically indistinguishable from T-Coffee and is the fastest published method for large numbers of sequences, able to align 5,000 sequences of length 300 in 7 minutes on a desktop computer. MUSCLE is freely available at <http://www.drive5.com/muscle>.

1. Introduction

Multiple alignments of protein sequences are important in many applications, including phylogenetic tree estimation, secondary structure prediction and critical residue identification. Many multiple sequence alignment (MSA) algorithms have been proposed; for a recent review, see [1]. Two attributes of MSA programs are of particular importance to the user: biological accuracy and computational complexity (i.e., time and memory requirements). Complexity is of increasing relevance due to the rapid growth of sequence databases, which now contain enough representatives of larger protein families to exceed the capacity of most current programs. Obtaining biologically accurate alignments is also a challenge, as the best methods sometimes fail to align readily apparent conserved motifs. Here, we present MUSCLE [2], a new MSA package that achieves the highest scores so far reported on four alignment accuracy benchmarks and substantially faster execution times than T-Coffee [3], the most accurate program

previously published. MUSCLE provides alternative options for high-throughput applications. One such option, MUSCLE-p, is the fastest method known to the author for large numbers of sequences. MUSCLE-p gives average accuracy statistically equal to T-Coffee and can align 5,000 sequences of length 300 in 7 minutes on a current desktop computer.

2. MUSCLE algorithm

Distances between sequences are estimated using k -mer counting [4] and clustered using UPGMA [5], giving a binary tree. The resulting binary tree is used to construct a progressive alignment [6, 7]. At each internal node of the tree, profiles of each subtree are aligned by optimizing the log-expectation objective function [2]. 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 [8]. This avoids the conventional (and expensive) step of building an explicit multiple alignment at each internal node of the tree in order to compute the new profile. An alignment (A) at the root node is constructed 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. The fractional identity of each pair of sequences is computed from A and converted to an additive distance estimate by correcting for multiple substitutions at a single site [9]. This distance matrix is clustered by UPGMA, yielding a new tree. The branching orders of the old and new trees are 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 a new alignment of all input sequences. The algorithm may be terminated here; giving the option we call

MUSCLE-p. This method gives accuracy statistically indistinguishable from T-Coffee and is substantially faster than CLUSTALW, especially for large numbers of sequences. The final stage of the algorithm attempts iterative refinement using tree-dependent restricted partitioning [10].

3. Results

We assessed the performance of MUSCLE on four sets of reference alignments: BALiBASE v2 [11], SABmark [12], SMART [13] and our own benchmark, PREFAB [2]. We compared with CLUSTALW [14], the most widely used program at the time of writing, and T-Coffee, which has the best BALiBASE score previously reported. Table 1 gives accuracy scores, where accuracy is defined as the fraction of positions correctly aligned. Table 2 gives execution times required to complete each benchmark in seconds on a 2.5 GHz desktop computer with 1 GB RAM. On test sets containing from 200 to 1,000 sequences, and thus more representative of typical input data, MUSCLE was consistently faster than CLUSTALW (Table 3). On a large test set with 5,000 sequences of average length 350, MUSCLE-p needed 7 minutes. We estimate that CLUSTALW would need approximately one year to complete this test.

Table 1. Benchmark accuracy scores.

Method	PREFAB	BALiBASE	SABmark	SMART
MUSCLE	0.648	0.896	0.430	0.856
MUSCLE-p	0.634	0.883	0.427	0.837
T-Coffee	0.632	0.882	0.424	0.835
CLUSTALW	0.588	0.860	0.404	0.823

Table 2. Benchmark CPU times.

Method	PREFAB	BALiBASE	SABmark	SMART
MUSCLE-p	3,000	52	429	180
MUSCLE	11,000	81	1,500	560
CLUSTALW	15,000	160	210	480
T-Coffee	1,000,000	1,500	5,600	78,000

Table 3. CPU times for 200, 600 and 1,000 sequences.

Method	200 seqs.	600 seqs.	1,000 seqs.
MUSCLE-p	10	47	131
MUSCLE	36	592	5986
CLUSTALW	142	1595	6914
T-Coffee	(aborted)	(aborted)	(aborted)

4. References

- [1] C. Notredame, "Recent progress in multiple sequence alignment: a survey," *Pharmacogenomics*, vol. 3, pp. 131-44, 2002.
- [2] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res*, vol. 32, pp. 1792-1797, 2004.
- [3] C. Notredame, D. G. Higgins, and J. Heringa, "T-Coffee: A novel method for fast and accurate multiple sequence alignment," *J Mol Biol*, vol. 302, pp. 205-17, 2000.
- [4] R. C. Edgar, "Local homology recognition and distance measures in linear time using compressed amino acid alphabets," *Nucleic Acids Res*, vol. 32, pp. 380-5, 2004.
- [5] P. H. A. Sneath and R. R. Sokal, *Numerical taxonomy*. San Francisco: Freeman, 1973.
- [6] D. F. Feng and R. F. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," *J Mol Evol*, vol. 25, pp. 351-60, 1987.
- [7] P. Hogeweg and B. Hesper, "The alignment of sets of sequences and the construction of phyletic trees: an integrated method," *J Mol Evol*, vol. 20, pp. 175-86, 1984.
- [8] R. C. Edgar, "Multiple sequence alignment with reduced time and space complexity," (Submitted), 2004.
- [9] M. Kimura, *The neutral theory of molecular evolution*: Cambridge University Press, 1983.
- [10] M. Hirosawa, Y. Totoki, M. Hoshida, and M. Ishikawa, "Comprehensive study on iterative algorithms of multiple sequence alignment," *Comput Appl Biosci*, vol. 11, pp. 13-8, 1995.
- [11] A. Bahr, J. D. Thompson, J. C. Thierry, and O. Poch, "BALiBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations," *Nucleic Acids Res*, vol. 29, pp. 323-6, 2001.
- [12] I. Van Walle, I. Lasters, and L. Wyns, "Align-m--a new algorithm for multiple alignment of highly divergent sequences," *Bioinformatics*, 2004.
- [13] C. P. Ponting, J. Schultz, F. Milpetz, and P. Bork, "SMART: identification and annotation of domains from signalling and extracellular protein sequences," *Nucleic Acids Res*, vol. 27, pp. 229-32, 1999.
- [14] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res*, vol. 22, pp. 4673-80, 1994.