MUSCLE

a multiple sequence alignment method with reduced time and space complexity

Robert C. Edgar

presented by Chase Jenkins
November 13, 2008

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity.
BMC Bioinformatics 5(1): 113.

Overview

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Overview

- MUSCLE
  - "Multiple sequence comparison by log-expectation"
- Previous paper introducing MUSCLE
  - Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-97
  - Gave an overview of MUSCLE and showed its excellent performance against four alignment benchmarks

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Overview

- This paper
  - More in-depth discussion and analysis of the MUSCLE algorithm
  - Describes techniques to improve the MUSCLE algorithm's biological accuracy and/or computational complexity
  - Introduces "MUSCLE-fast" for high performance uses

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Overview

- Results
  - Speed and accuracy compared with:
    - CLUSTALW
    - Progressive POA
    - MAFFT script FFTNS1
      - *"fastest previously published program known to the author"*

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Overview

- Results
  - Comparison made using four benchmarks:
    - BAliBASE
    - SABmark
    - SMART
    - The author's own benchmark, PREFAB

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Overview

- Results
  - Three variants of MUSCLE tested:
    - MUSCLE (default)
      - Most accurate
    - MUSCLE-fast
      - High throughput
    - MUSCLE-prog
      - A compromise between the two
    - A more detailed description of these will be given later

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Overview

- Conclusions
  - *"MUSCLE offers a range of options that provide improved speed and/or alignment accuracy compared with currently available programs."*
  - MUSCLE is freely available
    - http://www.drive5.com/muscle

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Overview

- Multiple sequence alignment (MSA) algorithms have two characteristics of primary importance to the user (as with most algorithms):
  - Accuracy
    - Specifically, biological accuracy
  - Computational complexity
    - Running time
    - Space requirements

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Implementation

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Implementation

- Basic idea
  - Bring together the concepts of sequence alignment and phylogenetic tree construction
  - The basic strategy is similar to that used by PRRP and MAFFT
    - Build a progressive alignment
    - Apply "horizontal refinement"
      - Extract pairs of profiles from the progressive alignment and re-align them
      - Keep the new alignment only if it improves some objective score

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Basic Algorithm Elements

Implementation

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Basic Algorithm Elements

- Similarity measures
  - A measure calculated for a pair of sequences that estimates their degree of evolutionary divergence
    - Operates under the assumption that the sequences are related

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Basic Algorithm Elements

- Similarity measures
  - Two types of similarity measures are used by MUSCLE:
    - The fractional identity, *D*, calculated from a global alignment of the two sequences
    - *k*-mer (subsequences of length *k*) counting
      - Related sequences should have more common subsequences than unrelated sequences
      - Generally faster than performing a global alignment
      - Relies on *k* not being too large, and the sequences not being too divergent
  - Author has previously shown that *k*-mer counting correlates well with fractional identity

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Basic Algorithm Elements

- Distance measures
  - The distance measure between two sequences, *A* and *B*, should be additive
    - *d(A, B) = d(A, C) + d(C, B)* for any sequences *A*, *B*, *C* that are related
  - The ideal distance measure is the "mutation distance" – the number of mutations that occurred on the evolutionary path between the two sequences

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Basic Algorithm Elements

- Distance measures
  - The fractional identity, *D*, is often used as a similarity measure. For closely related sequences, *1 - D* is then a good approximation of the mutation distance.
    - As we've discussed, *1 - D* would be an exact measure of the mutation distance under the *infinite site model*
    - As sequences diverge, however, there is a higher probability for multiple mutations at a single site, therefore MUSCLE uses the Kimura distance as an estimate of the mutation distance:

$$d_{\text{Kimura}} = -\log_e (1 - D - D^2/5)$$

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Basic Algorithm Elements

- Distance measures
  - When a *k*-mer similarity measure is used, the following equation is used for the distance measure:

$$d_{\text{kmer}} = 1 - F.$$

where *F* is the *k*-mer similarity measure.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Basic Algorithm Elements

- Tree construction
  - Given a triangular distance matrix, a binary tree is constructed by clustering the matrix
    - Clustering is performed using either UPGMA or neighbor-joining
      - Given two clusters (subtrees), the distance to a third cluster is calculated using a method similar to that used by MAFFT, where a weighted mix of the average distance and minimum distance is used

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

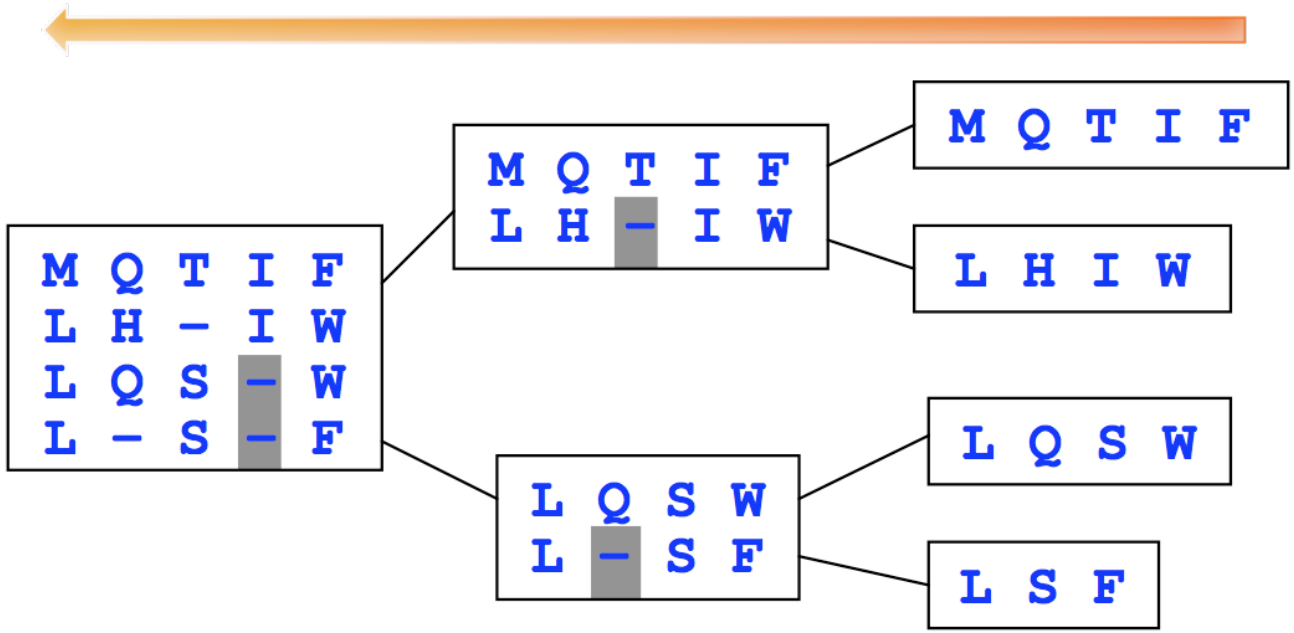
Basic Algorithm Elements

- Progressive alignment
  - In the first two stages of the algorithm, a progressive alignment is constructed for the sequences.
  - This requires a rooted binary tree where each sequence is a leaf node of the tree.
    - Tree is constructed by calculating the distance measures, placing them in a triangular matrix, then clustering the matrix, as previously mentioned.
  - The branching order of the tree is followed from the leaf nodes up to the root. At each internal node, a profile-profile alignment is used to align the existing two subtree alignments, and that new alignment is assigned to that internal node.
  - Progressing in this fashion, an MSA is obtained at the root of the tree.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Basic Algorithm Elements

- Progressive alignment



Implementation

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Basic Algorithm Elements

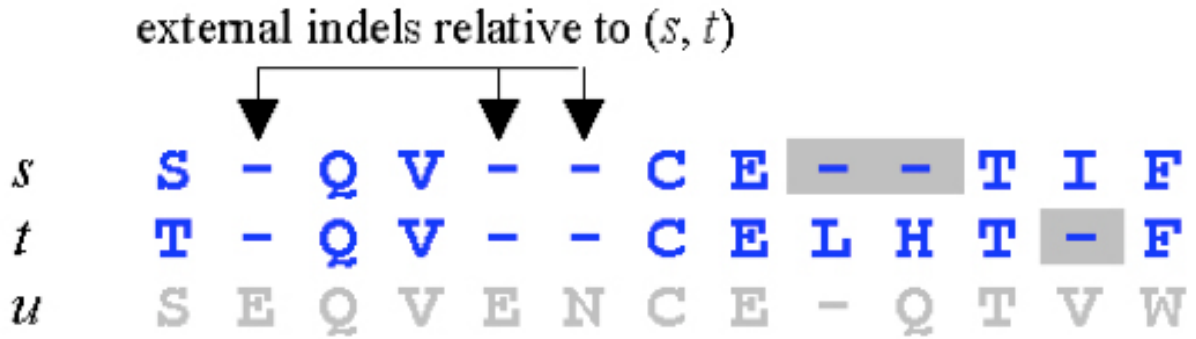
- Objective score
  - In its third stage, the MUSCLE algorithm iteratively attempts to maximize an objective score. The scoring method used by MUSCLE is the **SP (sum-of-pairs) score**.
    - This is the sum of the alignment scores of each pair of sequences.
    - The alignment score of an individual sequence pair is simply the sum of the scores of each aligned residue using a substitution matrix, plus affine gap penalties

$$g + \lambda e$$

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Basic Algorithm Elements

- Objective score
  - The contribution of gap penalties to the SP for a single pair of sequences is calculated by discarding all indels common to both sequences, then calculating an affine gap penalty for the two sequences



Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

Implementation

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- Three-stage algorithm
  - Stage 1: Draft progressive alignment built
  - Stage 2: Improved progressive alignment built
  - Stage 3: Iterative refinement performed
- An MSA is available at the completion of each stage

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 1**: Draft progressive alignment built
  - Steps:
    - *Similarity measure*
    - *Distance estimate*
    - *Tree construction*
    - *Progressive alignment*

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 1**: Draft progressive alignment built
  - *Similarity measure*
    - A measure of the similarity of each pair of sequences is computed either:
      - Using *k-mer* counting, or
      - By constructing a global alignment of the two sequences and calculating *D*, the "fractional identity"

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 1**: Draft progressive alignment built
  - *Distance estimate*
    - A triangular distance matrix is computed from the pair-wise similarity measures

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 1**: Draft progressive alignment built
  - *Tree construction*
    - A tree is constructed from the distance matrix using either:
      - UPGMA, or
      - Neighbor-joining

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 1**: Draft progressive alignment built
  - *Progressive alignment*
    - A progressive alignment is built by following the branching order of the constructed tree
    - This gives an MSA of all input sequences at the root of the tree

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 2**: Improved progressive alignment built
  - This stage is iterative.
  - Steps:
    - *Similarity measure*
    - *Distance estimate & tree construction*
    - *Tree comparison*
    - *Progressive alignment*

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 2**: Improved progressive alignment built
  - *Similarity measure*
    - A measure of the similarity of each pair of sequences is again computed
    - This time, however, the similarity measure is computed using the fractional identity of the two sequences as aligned in the current MSA

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 2**: Improved progressive alignment built
  - *Distance estimate & tree construction*
    - A tree is constructed by calculating a Kimura distance matrix and applying a clustering method to the matrix
      - again, either UPGMA or neighbor-joining

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 2**: Improved progressive alignment built
  - *Tree comparison*
    - The previous tree (either the tree from Stage 1 or the tree from the previous execution of Stage 2) is compared to the new tree and the internal nodes which change the branching order are identified.
    - If the number of changed internal nodes has not decreased, the tree has been "fully" improved and iteration ceases.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 2**: Improved progressive alignment built
  - *Progressive alignment*
    - A new progressive alignment is built based on the previous alignment and the changed tree nodes
    - The previous alignment is retained for each subtree where there are no changes in the branching order.
      - New alignments are computed only for the changed nodes (if there are any)

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 2**: Improved progressive alignment built
  - *Progressive alignment*
    - When the alignment at the tree's root completes, the algorithm can:
      - Terminate
      - Return to the first step of Stage 2
      - Move on to Stage 3

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 3**: Iterative refinement performed
  - This stage is also iterative.
  - Iterative refinement is performed using partitioning of the tree.
  - Steps:
    - *Choice of bipartition*
    - *Profile extraction*
    - *Re-alignment*
    - *Acceptance / Rejection*

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 3**: Iterative refinement performed
  - *Choice of bipartition*
    - An edge is deleted from the tree, partitioning the tree into two disjoint subtrees
    - For iteration, edges are visited beginning farthest away from the root, progressing closer to the root.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 3**: Iterative refinement performed
  - *Profile extraction*
    - The profile (MSA) of each of the two partitions is extracted from the overall MSA
    - Columns containing no residues (that is, only indels/gaps) are discarded

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 3**: Iterative refinement performed
  - *Re-alignment*
    - The two profiles obtained in the previous step are aligned using profile-profile alignment

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 3**: Iterative refinement performed
  - *Acceptance / Rejection*
    - The **SP score** of the MSA derived from the new profile-profile alignment is calculated.
      - If the score is greater than the previous alignment's score, the new MSA is kept.
      - Otherwise, it is discarded.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Algorithm Overview

- **Stage 3**: Iterative refinement performed
  - *Acceptance / Rejection*
    - If either:
      - All edges have been visited
      - A user-defined number of maximum iterations has been reached

    then the algorithm terminates, otherwise it returns to the first step in Stage 3.

"*Visiting edges in order of decreasing distance from the root has the effect of first re-aligning individual sequences, then closely related groups of sequences.*"

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

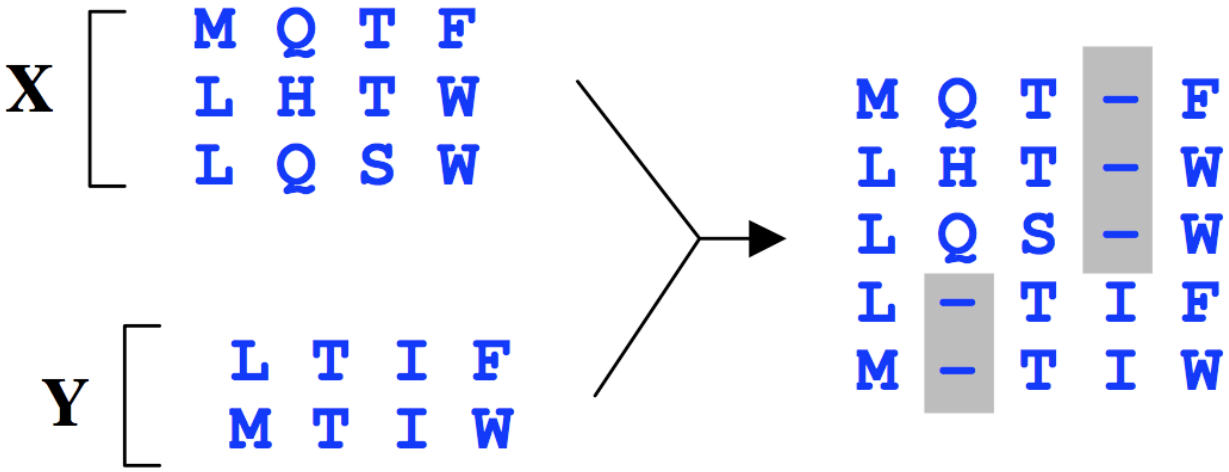
Additional Algorithm Elements

Implementation

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Additional Algorithm Elements

- Profile functions
  - To align profiles, a scoring function must be defined for a pair of profile positions (MSA columns)
    - Analogous to a substitution matrix for residues



Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Additional Algorithm Elements

- Profile functions
  - PSP is the method used by CLUSTALW and MAFFT for this purpose, and it is also used by MUSCLE, as based on PAM 200 and VTML 240
  - Additionally, MUSCLE uses LE – log-expectation, a modified version of log-average (LA), which uses probabilities computed from VTML 240

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

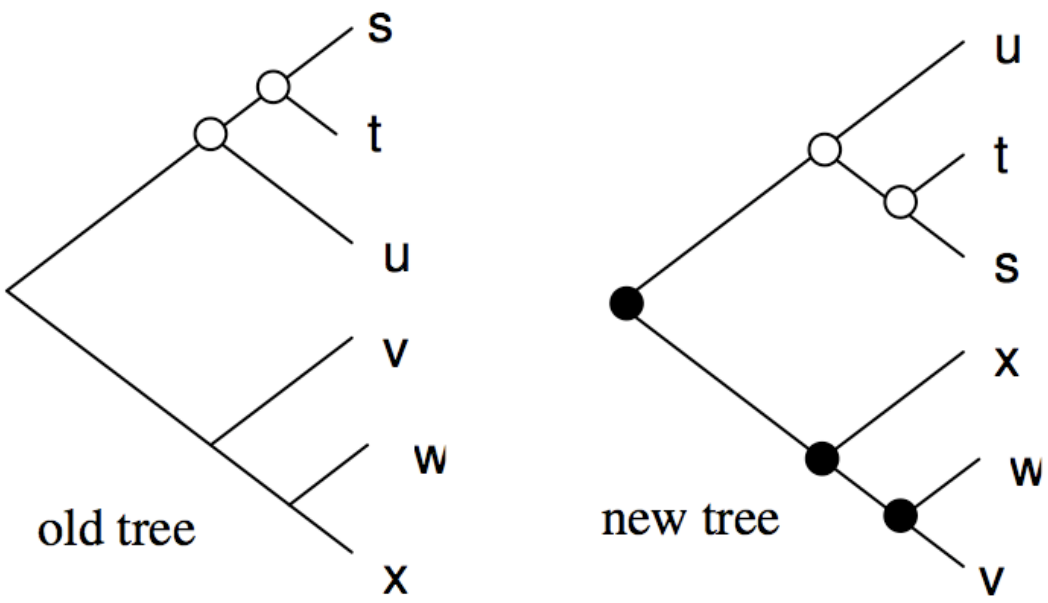
Additional Algorithm Elements

- Tree comparison
  - In progressive alignment, two trees with the same branching order and the same sequences at their leaves produce identical alignments
  - When a new tree is constructed in Stage 2, the subtrees that are unchanged do not have to have their alignments re-computed.
    - Progressive alignments only have to be computed for the changed subtrees, yielding an alignment at the root that would be identical to the alignment constructed without retaining previous alignments

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Additional Algorithm Elements

- Tree comparison



Implementation

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Additional Algorithm Elements

- Sequence weighting
  - Sequences should be weighted to correct for biased sampling from a family of related proteins
  - No consensus on how this is best done
  - MUSCLE allows multiple methods of sequence weighting:
    - None (sequences not weighted)
    - Henikoff
    - PSI-BLAST
    - CLUSTALW's GSC
    - The three-way method

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Additional Algorithm Elements

- Sequence weighting
  - Author found weighting to yield a small improvement in accuracy against the benchmarks (1% against BAliBASE)
    - Little difference between the weighting methods (excluding none)
    - CLUSTALW's GSC reduces the algorithm's complexity, so it is used by default

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Defaults, optimizations, and complexity

Implementation

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Defaults, optimizations, and complexity

- There are a variety of options in MUSCLE for improving accuracy or improving complexity
- The algorithm's default options provide the most accurate results, as tested against the four benchmarks

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Results

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Results

- MUSCLE provides a number of options that trade-off accuracy and time-space complexity
- Three sets of options were tested:
  - Full **MUSCLE**, including all three stages with default options
  - **MUSCLE-prog**: Stages 1 and 2 only with default options
  - **MUSCLE-fast**: Stage 1 only using the fastest options

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Results

Table 2: Complexity of MUSCLE. Here we show the big-O asymptotic complexity of the elements of MUSCLE as a function of *L*, the typical sequence length, and *N*, the number of sequences, retaining the highest-order terms in *N* with *L* fixed and vice versa.

Step	O(Space)	O(Time)
K-mer distance matrix	$N^2 + L$	N^2L
UPGMA	N^2	N^2
Progressive (one iteration)	$L_P^2 = NL + L^2$	$L_P^2 = N^2 + L^2$
Progressive (root alignment)	$NL_P = N^2 + NL$	$NL_P \log N = N^2 \log N + NL \log N$
Progressive (*N* iterations + root)	$N^2 + NL + L^2$	$N^3 + NL^2$
Refinement (one edge)	$NL_P + L_P^2 = N^2 + L^2$	$N^2L_P + L_P^2 = N^3 + L^2$
Refinement (*N* edges)	$N^2 + L^2$	$N^4 + NL^2$
TOTAL	$N^2 + L^2$	$N^4 + NL^2$

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Results

Table 3: Accuracy scores. The average accuracy, measured by the Q score, is reported for each method on each set of reference alignments.

Method	PREFAB	BAliBASE	SABmark	SMART
MUSCLE	0.648	0.896	0.430	0.856
MUSCLE-prog	0.634	0.883	0.427	0.837
FFTNS1	0.619	0.844	0.376	0.815
MUSCLE-fast	0.616	0.849	0.396	0.813
CLUSTALW	0.588	0.860	0.404	0.823
POA-blast	0.577	0.839	0.352	0.788
POA	0.576	0.834	0.280	0.797

Note that the methods are ranked from highest to lowest scores *using the author's benchmark, PREFAB*

The full MUSCLE algorithm, as well as MUSCLE-prog, still outperform the other methods for all four benchmarks for accuracy

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Results

Table 4: CPU times. The total CPU time required to create all alignments in each test set, measured in seconds on a 2.5 GHz Pentium 4 desktop computer.

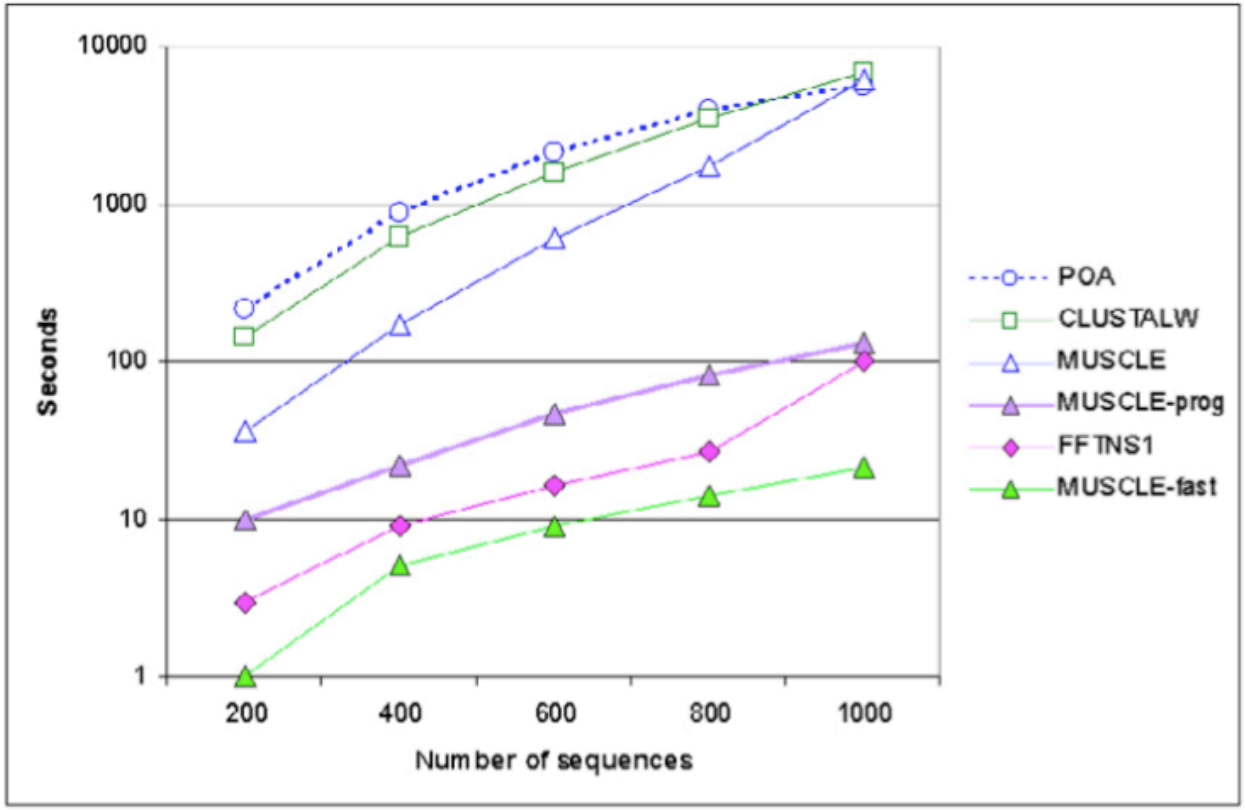
Method	PREFAB	BAliBASE	SABmark	SMART
MUSCLE-fast	540	11	45	30
FFTNS1	720	16	70	46
MUSCLE-prog	3,000	52	429	180
MUSCLE	11,000	81	1,500	560
POA-blast	11,000	90	290	670
CLUSTALW	15,000	160	210	480
POA	24,000	130	380	880

Note, again, that the methods are ranked from highest to lowest scores *using the author's benchmark, PREFAB*

MUSCLE-fast outperforms all the others for speed in all four benchmarks, however, other methods (e.g. CLUSTALW) outperform the full MUSCLE algorithm in two of the four benchmarks – of particular note is SABmark, where CLUSTALW significantly outperforms MUSCLE, while their accuracy scores are quite similar (see previous slide)

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Results



Execution time vs. number of input sequences

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Conclusions

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

Conclusions

- *"MUSCLE demonstrates improvements in accuracy and reductions in computational complexity…"*
  - **My thoughts**:
    - We must note that there is still typically a trade-off between the two, as the data shows.
    - MUSCLE's various options make it a very powerful tool, and its accuracy at its default settings add to its strength.
    - However, even the benchmarks demonstrate that other methods, such as CLUSTALW, can rival its processing time, and MUSCLE's high throughput configuration, MUSCLE-fast, is outperformed in accuracy by half of the other methods.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5(1): 113.

MUSCLE

a multiple sequence alignment method with reduced time and space complexity

Robert C. Edgar

presented by Chase Jenkins
November 13, 2008

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity.
BMC Bioinformatics 5(1): 113.