Multiple Sequence Alignment

Biol4230 Tues, February 20, 2018

Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

Goals of today's lecture:

- Why multiple sequence alignment (MSA)?
  - identify conserved (functional?) positions among related sequences
  - input to evolutionary tree methods
- MSA computational complexity
  - Models for MSA: tree-based, Sum-of-pairs, star
  - "optimal" $O(N^k)$ (k sequences of length N)
  - progressive: $O(k^2N^2)$
  - progressive/iterative ($O(k^2N^2)$)
- Evaluating MSA accuracy
  - BALIBASE
- Phylogenetic alignments – BaliPhy

To learn more:

- Pevsner Bioinformatics Chapter 6 pp 179–212
- Altschul, S. F. and Lipman, D. J. (1989) "Trees, stars, and multiple biological sequence alignment." SIAM J. Appl. Math. 49:197-209.
- Thompson, J. D., Plewniak, F., and Poch, O. (1999) "A comprehensive comparison of multiple sequence alignment programs." Nucleic Acids Res 27:2682-2690
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position- specific gap penalties and weight matrix choice." Nucleic Acids Res 22:4673-4680.
- R. C. Edgar (2004) "MUSCLE: a multiple sequence alignment method with reduced time and space complexity." BMC Bioinformatics 5:113
- Suchard, M. A. and Redelings, B. D. (2006) "BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny" Bioinformatics 22:2047-2048

Overview

- *No multiple alignments without* ***HOMOLOGY***
- Multiple sequence alignments can resolve ambiguous gaps – largely used to specify gap positions
- Many popular programs build successive pair-wise alignments (progressive alignment) – Clustal-W (Clustal-Omega), T-coffee, MUSCLE
- Simple progressive alignment methods fix gaps early, after which they cannot be moved
- Iterative approaches required to adjust gaps
- Tree-based alignments bring a more phylogenetic perspective
- What are appropriate tests – alignments for trees vs alignments for structures?

Algorithms for Pairwise Sequence Comparison

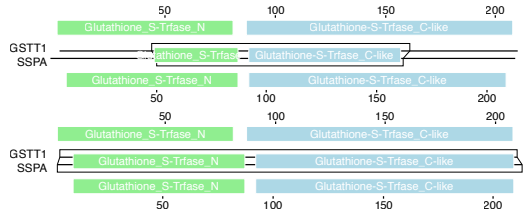
Algorithm	Value calculated	Scoring Matrix	Gap Penalty	Time Required	
Needleman-Wunsch	Global similarity	arbitrary	penalty/gap	$O(n^2)$	Needleman and Wunsch (1970)
Sellers	(global) distance	Unity	penalty/residue	$O(n^2)$	Sellers (1974)
Smith-Waterman	local similarity	$S_{ij} < 0.0$	affine q+rk	$O(n^2)$	Smith and Waterman, 1981; Gotoh 1982
SRCHN	approx. local	$S_{ij} < 0.0$	penalty/gap	O(n)-$O(n^2)$	Wilbur and Lipman (1983)
FASTA	approx. local	$S_{ij} < 0.0$	affine q+rk	$O(n^2)/K$	Lipman and Pearson (1985, 1988)
BLASTP		$S_{ij} < 0.0$	multiple HSP	$O(n^2)/K$	Altschul et al (1990)
BLAST2.0	approx. local	$S_{ij} < 0.0$	affine q+rk	$O(n^2)/K$	Altschul et al (1997)

Local, global, and "glocal" alignments

Local – 26.3% id
E() < 0.00024

Globally similar:

Global – 20.6% id
E() < 10^{-7}

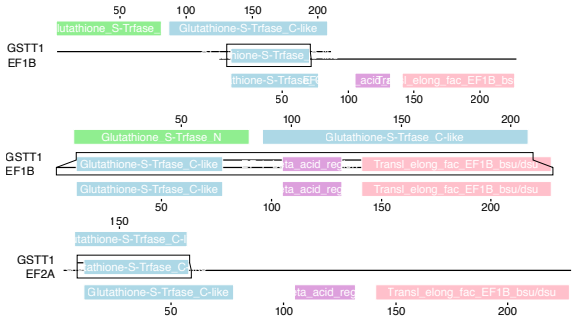


Local – 29.2% id
E() < 9

Locally similar:

Global – 15.3% id
E() < 7000

Glocal – 26.8% id
E() < 0.02



*No multiple alignments without **HOMOLOGY***

Homologs

```
GSTM1_HUMAN  ----------MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKITQ
GSTM3_HUMAN  ----------MSCESSMVLGYWDIRGLAHAIRLLLEFTDTSYEEKRYTCGEAPDYDRSQWLDVKFKLDLDFPNLPYLLDGKNKITQ
GSTP1_HUMAN  ----------MPPYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTV--------ETWQEGSLKASCLYGQLPKFQDGDLTLYQ
GSTA1_HUMAN  ----------MAEKPKLHYFNARGRMESTRWLLAAAGVEFEEKFIKS-------AEDLDKLRNDGYLMFQQVPMVEIDGMKLVQ
HPGDS_HUMAN  ----------MPNYKLTYFNMRGRAEIIRYIFAYLDIQYEDHRIEQ--------ADWP--EIKSTLPFGKIPILEVDGLTLHQ
GSTT1_HUMAN  ----------MGLELYLDLLSQPCRAVYIFAKKNDIPFELRIVDLIK------GQHLSDAFAQVNPLKKVPALKDGDFTLTE
GSTO1_HUMAN  MSGESARSLGKGSAPPGPVPEGSIRIYSMRFCPFAERTRLVLKAKGIRHEVININ-------LKNKPEWFFKKNPFGLVPVLENSQGQLIY
                          :                                  . .                              :* .  .   :

GSTM1_HUMAN  SNAILCYIARKHNL----CGETEEEKIRVDILENQ---------TMDNHMQLGMICYNPEFEKLKP--KYLEELPEKLKLYSEFLGK-----
GSTM3_HUMAN  SNAILRYIARKHNM----CGETEEEKIRVDIIENQ---------VMDFRTQLIRLCYSSDHEKLKP--QYLEELPGQLKQFSMFLGK-----
GSTP1_HUMAN  SNTILRHLGRTLGL----YGKDQQEAALVDMVNDG---------VEDLRCKYISLIYT-NYEAGKD--DYVKALPGQLKPFETLLSQNQGG-
GSTA1_HUMAN  TRAILNYIASKYNL----YGKDIKERALIDMYIEG---------IADLGEMILLLPVCPPEEKDAK--LALIKEKIKNRYFPAFEKVLKSHG
HPGDS_HUMAN  SLAIARYLTKNTDL----AGNTEMEQCHVDAIVDT---------LDDFMSCFPWAEKKQDVKEQMFNELLTYNAPHLMQDLDTYLGG-----
GSTT1_HUMAN  SVAILLYLTRKYKVPDYWYPQDLQARARVDEYLAWQHTTLRRSCLRALWHKVMFPVFLGEPVSPQTLAATLAELDVTLQLLEDKFLQN----
GSTO1_HUMAN  ESAITCEYLDEAYPGKKLLPDDPYEKACQKMILEL---------FSKVPSLVGSFIRSQNKEDYAG---LKEEFRKEFTKLEEVLTN---KK
               :*               .        .
```

Non-homologs

```
GSTM1_HUMAN  ----MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAP--DYDRSQWLNEKFKLGLDFPNLPYLIDGAHKITQSNAILCY
GSTP1_HUMAN  ---MPPYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTV----------ETWQEGSLKASCLYGQLPKFQDGDLTLYQSNTILRH
GSTT1_HUMAN  ----MGLELYLDLLSQPCRAVYIFAKKNDIPFELRIVDLIKG--------QHLSDAFAQVNPLKKVPALKDGDFTLTESVAILLY
NARJ_ECO57   ---MIELVIVSRLLEYPDAALWQHQQEMFEAIAASKNLP-------------KEDAHALGIFLRDLTTMDPLDAQAQYSELFDRG
DYR_BPT4     ---MIKLVFRYSPTKTVDGFNELAFG--------------------------LGDGLPWGRVKKDLQNFKARTEGTIMIMGAKTF
TPIS_RABIT   APSRKFFVGGNWKMNGRKKNLGELITTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKD
                                                                                    .  .

GSTM1_HUMAN  IARKHNL----CGETEEEKIRVD--ILENQTMDNHMQLGMICYNP----EFEKLK-----PKYLEELPEKLKLYSEFLGK-----
GSTP1_HUMAN  LGRTLGL----YGKDQQEAALVD--MVNDGVEDLRCKYISLIYT-----NYEAGK-----DDYVKALPGQLKPFETLLSQNQGG-
GSTT1_HUMAN  LTRKYKVPDYWYPQDLQARARVDEYLAWQHTTLRRSCLRALWHKV----MFPVFLGEPVSPQTLAATLAELDVTLQLLEDKFLQN
NARJ_ECO57   RATSLLLFEHVHGESRDRGQAMVDLLAQYEQHGLQLNSRELPDHLPLYLEYLAQLPQSEAVEGLKDIAPILALLSARLQQRESR-
DYR_BPT4     QSLPTLLP--------GRSHIVVCDLARDYPVTKDGDLAHFYITWEQYITYISGG--------EIQVSSPNAPFETMLDQNSK--
TPIS_RABIT   CGATWVVLG--HSERRHVFGESDELIGQKVAHALSEGLGVIACIGEKLDEREAGITEKVVFEQTKVIADNVKDWSKVVLAYEP--
                        :                 :              :                              :
```

Homology is confusing I: Homology defined Three(?) Ways

- Proteins/genes/DNA that share a common ancestor
- Specific positions/columns in a multiple sequence alignment that have a 1:1 relationship over evolutionary history
  - sequences are *50% homologous* ???
- Specific (morphological/functional) characters that share a recent divergence (clade)
  - bird/bat/butterfly wings are/are not homologous

Multiple alignments (can) place gaps

+1 : match
-1 : mismatch
-2 : gap

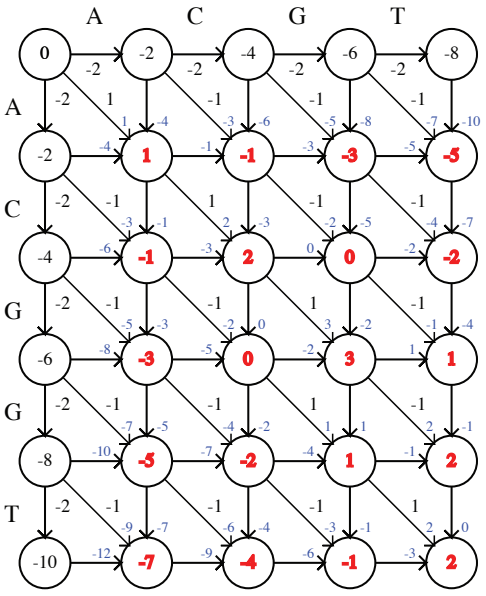
A:A

ACGT:ACGGT

```
1: ACG-T AC-GT
2: ACGGT ACGGT
     +2    +2

1: ACG-T AC-GT
2: ACGGT ACGGT
3: ACCGT ACCGT
     +5    +7
```

Sum of pairs score:
1v2+1v3+2v3



Multiple alignments (can) place gaps
Sum-of-pairs scoring

+1 : match
-1 : mismatch
-2 : gap

```
1: ACG-T AC-GT
2: ACGGT ACGGT
3: ACCGT ACCGT
    +5     +7
```

Sum of pairs score:
1v2+1v3+2v3

```
1: ACG-T AC-GT
2: ACGGT ACGGT
    +2     +2
1: ACG-T AC-GT
3: ACCGT ACCGT
    +0     +2
2: ACGGT ACGGT
3: ACCGT ACCGT
    +3     +3
SP: +5     +7
```


Affine gap penalties:
gap(x) = open+x*extend

- Affine gap penalties consolidate gaps:
  - gap(x) = 0 + 7*x

```
      60        70        80          90         100
GSTM1  FPNLPYLIDGAHKITQSNAILCYIARKHN--LCGETE-EEKIRVDI-LE---NQ-TMD-N
       : ..: : ::   . .:.::  : :::..  :  : . ::   ::. .:   :: :   :
GSTF1  FGQVPALQDGDLYLFESRAICKYAARKNKPELLREGNLEEAAMVDVWIEVEANQYTAALN
              60        70        80        90        100       110
```

  - gap(x) = 11 + 1*x

```
 60        70        80          90        100       110
GSTM1  FPNLPYLIDGAHKITQSNAILCYIARKHN---LCGETEEEKIRVDI-LENQTMDNHMQLG
       : ..: : ::   . .:.::  : :::..   :   . ::   ::. .: ..   .    :.
GSTF1  FGQVPALQDGDLYLFESRAICKYAARKNKPELLREGNLEEAAMVDVWIEVEANQYTAALN
              60        70        80        90        100       110
```

The 3-Sequence Alignment Problem

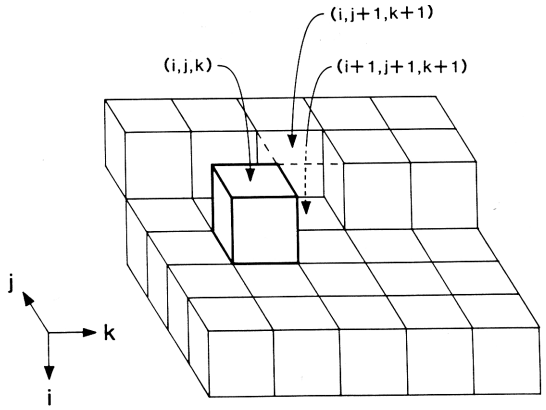
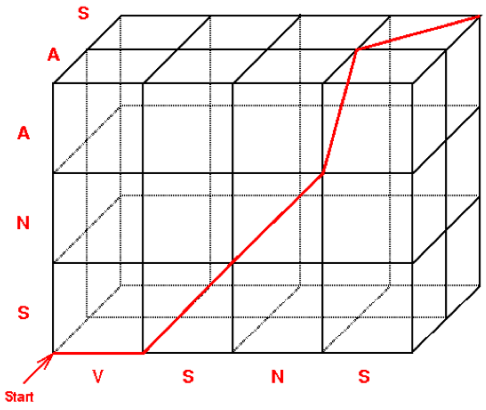


FIG. 1. Illustration of Algorithm 1 at the point where the value of L for the cell (i,j,k) (shown with heavily outlined edges) is calculated. Cells with lightly outlined edges are those for which the array $Q1$ holds the current value of Q just before the assignment to $Q1(j+1,k+1)$. After the assignment, $Q1(j+1,k+1)$ holds the value of Q for the cell $(i,j+1,k+1)$, indicated by dashed edges. As the depiction suggests, one may think of the array $Q1$ as holding the values of Q for a layer of cells with a moving ridge or fault line in it: the dashed cell is about to replace the one below it as part of this layer. (Reproduced from Ref. 2.)

Pairwise alignment:
$O(n^2)$ time
$400 \times 400 = 10^5$

k-wise alignment:
$O(n^k)$ time
$400^{10} = 10^{26}$

The Dynamic Programming Hyperlattice



V S N _ S
_ S N A _
_ _ _ A S

http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/node2.html

Efficiencies for global, close, alignments

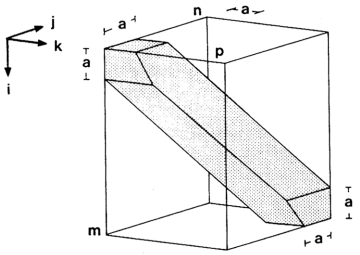


FIG. 3. The cells in the dotted region are included in the matrix calculations in ALIGN3. The region indicated can be imagined as being created by cutting corners of L with two pairs of parallel planes, one perpendicular to the $i-j$ face and the other perpendicular to the $i-k$ face. The upper and lower bounds of j with respect to i are $\min(\lfloor u(i-1)+a \rfloor, n)$ and $\max(\lceil u(i-a)+1 \rceil, 1)$, respectively, where $u=(n-a)/(m-a)$. Similarly, the upper and lower bounds of k with respect to i can be obtained by replacing n with p in the formulas.

Sequence alignment (particularly multiple sequence alignment) is about placing gaps. It is trivial to align K identical length sequences without gaps.

Trees, stars, and multiple alignment

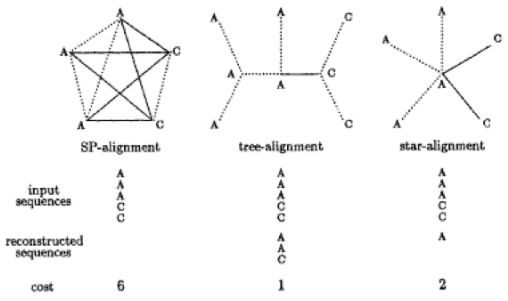
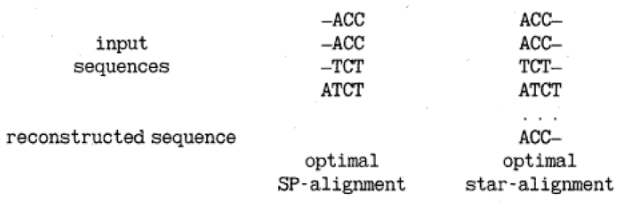


Fig. 1 SP-, tree-, and star-alignments for five, one-letter, input sequences. Pairwise alignments with cost one are indicated by solid lines, and pairwise alignments with cost zero are indicated by dotted lines.

Altschul, S. F. and Lipman, D. J. (1989) SIAM J. Appl. Math. 49:197-209

Trees, stars, and multiple alignment



Altschul, S. F. and Lipman, D. J. (1989) SIAM J. Appl. Math. 49:197-209

MSA2.1 - "optimal" multiple alignment 5 – 10 sequences

J Comput Biol. 1995 2:459-72.

Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment.
Gupta SK, **Kececioglu JD**, **Schaffer AA**.

The MSA program, written and distributed in 1989, is one of the few existing programs that attempts to find optimal alignments of multiple protein or DNA sequences. The MSA program implements a branch-and-bound technique together with a variant of Dijkstra's shortest paths algorithm to prune the basic dynamic programming graph. We have made substantial improvements in the time and space usage of MSA. The improvements make feasible a variety of problem instances that were not feasible previously. On some runs we achieve an order of magnitude reduction in space usage and a significant multiplicative factor speedup in running time. To explain how these improvements work, we give a much more detailed description of MSA than has been previously available. In practice, MSA rarely produces a provably optimal alignment and we explain why.

Optimal (MSA2.1) vs ClustalW/TCoffee/MUSCLE

N-terminal

```
MSA optimal multiple alignment
GSTA1_HUMAN  MA--EKPKLHYFNARGRMESTRWLLAAAGVEFEEKFIKSAE-------DLDKLRNDGYLMFQQVPMVEIDGMKLVQTRAIL
GSTP1_HUMAN  MP---PYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTVET--------WQEGSLKASCLYGQLPKFQDGDLTLYQSNTIL
GSTM1_HUMAN  MP----MILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKITQSNAIL
GSTM3_HUMAN  MSCESSMVLGYWDIRGLAHAIRLLLEFTDTSYEEKRYTCGEAPDYDRSQWLDVKFKLDLDFPNLPYLLDGKNKITQSNAIL
                     *   **     * **        *                                       *          *   **
                     !     #                                  !           !           !
CLUSTAL 2.0.12 multiple sequence alignment
GSTA1_HUMAN  --MAEKPKLHYFNARGRMESTRWLLAAAGVEFEEKFIKS-------AEDLDKLRNDGYLMFQQVPMVEIDGMKLVQTRAIL
GSTP1_HUMAN  ---MPPYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTV--------ETWQEGSLKASCLYGQLPKFQDGDLTLYQSNTIL
GSTM1_HUMAN  ----MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKITQSNAIL
GSTM3_HUMAN  MSCESSMVLGYWDIRGLAHAIRLLLEFTDTSYEEKRYTCGEAPDYDRSQWLDVKFKLDLDFPNLPYLLDGKNKITQSNAIL
                     *   **     * **        *                                       *          *   **
MUSCLE (3.8) multiple sequence alignment
GSTA1_HUMAN  MAEKPKLH--YFNARGRMESTRWLLAAAGVEFEEKFIKSAE----DLDKLRNDGY---LMFQQVPMVEIDGMKLVQTRAIL
GSTP1_HUMAN  ---MPPYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTV--------ETWQEGSLKASCLYGQLPKFQDGDLTLYQSNTIL
GSTM1_HUMAN  ----MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKITQSNAIL
GSTM3_HUMAN  MSCESSMVLGYWDIRGLAHAIRLLLEFTDTSYEEKRYTCGEAPDYDRSQWLDVKFKLDLDFPNLPYLLDGKNKITQSNAIL
                     *   **     * **        *                                       *          *   **
CLUSTAL FORMAT for T-COFFEE Version_8.69
GSTA1_HUMAN  MAE--KPKLHYFNARGRMESTRWLLAAAGVEFEEKFIKSAEDL-------DKLRNDGYLMFQQVPMVEIDGMKLVQTRAIL
GSTP1_HUMAN  MP---PYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTVE--------TWQEGSLKASCLYGQLPKFQDGDLTLYQSNTIL
GSTM1_HUMAN  MP----MILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKITQSNAIL
GSTM3_HUMAN  MSCESSMVLGYWDIRGLAHAIRLLLEFTDTSYEEKRYTCGEAPDYDRSQWLDVKFKLDLDFPNLPYLLDGKNKITQSNAIL
             *         *   **     * **        *                                       *          *   **
```


Optimal (MSA2.1) vs ClustalW/TCoffee/MUSCLE

center

```
MSA optimal multiple alignment
GSTA1_HUMAN  NYIASKYNLYGKDIKERALIDMYIEGIADLGEMILLLPVCPPEEKDAKLALIKEKIKNRYFPAFEKVLKSH--GQDYLVGNKL
GSTP1_HUMAN  RHLGRTLGLYGKDQQEAALVDMVNDGVEDLRCKYISL-IY-TNYEAGKDDYVKA-LPGQ-LKPFETLLSQNQGGKTFIVGDQI
GSTM1_HUMAN  CYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMI-CYNPEFEKLKPKYLEE-LPEK-LKLYSEFLGKR----PWFAGNKI
GSTM3_HUMAN  RYIARKHNMCGETEEEKIRVDIIENQVMDFRTQLIRL-CYSSDHEKLKPQYLEE-LPGQ-LKQFSMFLGKF----SWFAGEKL
                     *    *    *       *
                                          #
CLUSTAL 2.0.12 multiple sequence alignment
GSTA1_HUMAN  NYIASKYNLYGKDIKERALIDMYIEGIADLGEMILLLPVCPPEEKDAKLALIKEKIKNRYFPAFEKVLKSHGQDYLVGNKL
GSTP1_HUMAN  RHLGRTLGLYGKDQQEAALVDMVNDGVEDLRCKYISLIYT-NYEAGKDDYVKALPGQLKPFETLLSQNQGG-KTFIVGDQI
GSTM1_HUMAN  CYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPEFEKLKPKYLEELPEKLKLYSEFLGK-----RPWFAGNKI
GSTM3_HUMAN  RYIARKHNMCGETEEEKIRVDIIENQVMDFRTQLIRLCYSSDHEKLKPQYLEELPGQLKQFSMFLGK-----FSWFAGEKL
                     *    *    *       *             *                              *
MUSCLE (3.8) multiple sequence alignment
GSTA1_HUMAN  NYIASKYNLYGKDIKERALIDMYIEGIADLGEMILLLPVCPPEE--KDAKLALIKEKIKNRYFPAFEKVL--KSHGQDYLVGNKL
GSTP1_HUMAN  RHLGRTLGLYGKDQQEAALVDMVNDGVEDLRCKYISLIYT-NYEAGKDDYVKALPGQLK-----PFETLLSQNQGGKTFIVGDQI
GSTM1_HUMAN  CYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPEFEKLKPKYLEELPEKLK-----LYSEFL----GKRPWFAGNKI
GSTM3_HUMAN  RYIARKHNMCGETEEEKIRVDIIENQVMDFRTQLIRLCYSSDHEKLKPQYLEELPGQLK-----QFSMFL----GKFSWFAGEKL
                     *    *    *       *             *  *          *          *          *
CLUSTAL FORMAT for T-COFFEE Version_8.69 SCORE=88, Nseq=4, Len=240
GSTA1_HUMAN  NYIASKYNLYGKDIKERALIDMYIEGIADLGEMILLLPVCPPEEKDAKLALIKEKIKNRYFPAFEKVLKSH---------GQDYLVGNKL
GSTP1_HUMAN  RHLGRTLGLYGKDQQEAALVDMVNDGVEDLRCKYISLIYT-N----------YEAGKDDYVKALPGQLKPFETLLSQNQGGKTFIVGDQI
GSTM1_HUMAN  CYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPE----------FEKLKPKYLEELPEKLKLYSEFL----GKRPWFAGNKI
GSTM3_HUMAN  RYIARKHNMCGETEEEKIRVDIIENQVMDFRTQLIRLCYSSD----------HEKLKPQYLEELPGQLKQFSMFL----GKFSWFAGEKL
                     *    *    *       *                        *  *  *       **              *
```

Optimal (MSA2.1) vs ClustalW/TCoffee/MUSCLE

C-terminal

```
MSA optimal multiple alignment
GSTA1_HUMAN  SRADIHLVELLYYVEELDSSLISSFPLLKALKTRISNLPTVKKFLQPGSPRKPPMDEKSLEEARKIFRF
GSTP1_HUMAN  SFADYNLLDLLLIHEVLAPGCLDAFPLLSAYVGRLSARPKLKAFLASPEYVNLPINGNGKQ--------
GSTM1_HUMAN  TFVDFLVYDVLDLHRIFEPKCLDAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGNK----
GSTM3_HUMAN  TFVDFLTYDILDQNRIFDPKCLDEFPNLKAFMCRFEALEKIAAYLQSDQFCKMPINNKMAQWGNKPVC-
                *      *             ** *     *                   *
CLUSTAL 2.0.12 multiple sequence alignment
GSTA1_HUMAN  SRADIHLVELLYYVEELDSSLISSFPLLKALKTRISNLPTVKKFLQPGSPRKPPMDEKSLEEARKIFRF
GSTP1_HUMAN  SFADYNLLDLLLIHEVLAPGCLDAFPLLSAYVGRLSARPKLKAFLASPEYVNLPINGNGKQ--------
GSTM1_HUMAN  TFVDFLVYDVLDLHRIFEPKCLDAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGNK----
GSTM3_HUMAN  TFVDFLTYDILDQNRIFDPKCLDEFPNLKAFMCRFEALEKIAAYLQSDQFCKMPINNKMAQWGNKPVC-
                *      *             ** *     *                   *
MUSCLE (3.8) multiple sequence alignment
GSTA1_HUMAN  SRADIHLVELLYYVEELDSSLISSFPLLKALKTRISNLPTVKKFLQPGSPRKPPMDEKSLEEARKIFRF
GSTP1_HUMAN  SFADYNLLDLLLIHEVLAPGCLDAFPLLSAYVGRLSARPKLKAFLASPEYVNLPINGN----GKQ----
GSTM1_HUMAN  TFVDFLVYDVLDLHRIFEPKCLDAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGNK----
GSTM3_HUMAN  TFVDFLTYDILDQNRIFDPKCLDEFPNLKAFMCRFEALEKIAAYLQSDQFCKMPINNKMAQWGNKPVC-
                *      *             ** *     *                   *
CLUSTAL FORMAT for T-COFFEE Version_8.69 SCORE=88, Nseq=4, Len=240
GSTA1_HUMAN  SRADIHLVELLYYVEELDSSLISSFPLLKALKTRISNLPTVKKFLQPGSPRKPPMDEKSLEEARKIFRF
GSTP1_HUMAN  SFADYNLLDLLLIHEVLAPGCLDAFPLLSAYVGRLSARPKLKAFLASPEYVNLPINGNGK--------Q
GSTM1_HUMAN  TFVDFLVYDVLDLHRIFEPKCLDAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGN----K
GSTM3_HUMAN  TFVDFLTYDILDQNRIFDPKCLDEFPNLKAFMCRFEALEKIAAYLQSDQFCKMPINNKMAQWGNK-PVC
                *      *             ** *     *                   *
```


Clustal/Clustal-W (hundreds of sequences Clustal-Omega (10,000's of sequences)

Thompson JD, **Higgins DG**, **Gibson TJ**.
Nucleic Acids Res. 1994 Nov 11;22(22):4673-80

CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.

The sensitivity of the commonly used progressive multiple sequence alignment method has been greatly improved for the alignment of divergent protein sequences. Firstly, individual weights are assigned to each sequence in a partial alignment in order to down-weight near-duplicate sequences and up-weight the most divergent ones. Secondly, amino acid substitution matrices are varied at different alignment stages according to the divergence of the sequences to be aligned. Thirdly, residue-specific gap penalties and locally reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure. Fourthly, positions in early alignments where gaps have been opened receive locally reduced gap penalties to encourage the opening up of new gaps at these positions. These modifications are incorporated into a new program, CLUSTAL W which is freely available.

Clustal - successive pairwise strategy

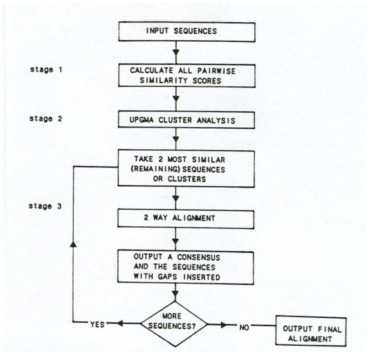


Fig. 1. Flow chart of the multiple alignment strategy described in MATERIALS AND METHODS, section **b**. The core of the multiple alignment process takes place in stage three. At each step, in stage three, two clusters consisting of one or more sequences each are aligned. After each alignment a consensus, which will be used to represent the two aligned clusters in future alignments, is stored and gaps are inserted in the original sequences at appropriate positions. When all sequences have been aligned, the process is complete and the full alignment is outputted.

A comprehensive comparison of multiple sequence alignment programs – BALIBASE

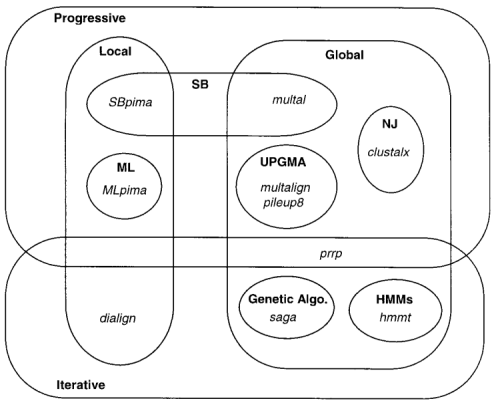


Figure 1. Schematic showing the relation between the different alignment programs and algorithms.

Thompson et al, (1999) NAR 27:2682-2690

BaliBase (1) datasets

Table 1. BAliBASE reference sets, showing the number of alignments in each set

Reference	Short (<100 residues)	Medium (200–300 residues)	Long (>400 residues)
Reference 1: equidistant sequences of similar length			
V1 (<25% identity)	7	8	8
V2 (20–40% identity)	10	9	10
V3 (>35% identity)	10	10	8
Reference 2: family versus orphans	9	8	7
Reference 3: equidistant divergent families	5	3	5
Reference 4: N/C-terminal extensions	12		
Reference 5: insertions	12		

Dimensions:
1. Overall similarity
2. Evolutionary topology (orphans vs equidistant)
3. differences in length

Thompson et al, (1999) NAR 27:2682-2690

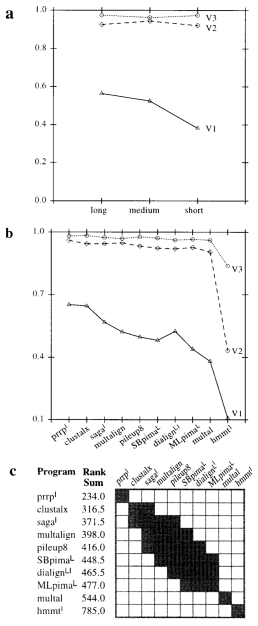


Figure 2. (**a**) SPS for reference 1, showing the median score in each category. (**b**) Median SPS for the programs in reference 1, categories V1, V2 and V3. Programs are displayed in the order of the Friedman test, with the highest scoring program on the left. (**c**) Results of the Friedman rank test to compare the performance of the programs in reference 1 ($S = 9$, $N = 81$, test statistic = 106.9). For each test alignment, the programs are assigned a rank between 1 and 10 (with 1 indicating the highest scoring program). The ranks are then summed over all alignments. Thus, a lower rank sum indicates that a program tends to achieve higher scores. The programs are listed in rank sum order. The grey boxes indicate that the two corresponding programs cannot be differentiated using the Friedman test ([alpha] = 5%). L, local alignment program; I, iterative alignment program.

Reference 1 - equidistant, similar length

SPS – sum of pairs score
CS – column score

Thompson et al, (1999) NAR 27:2682-2690

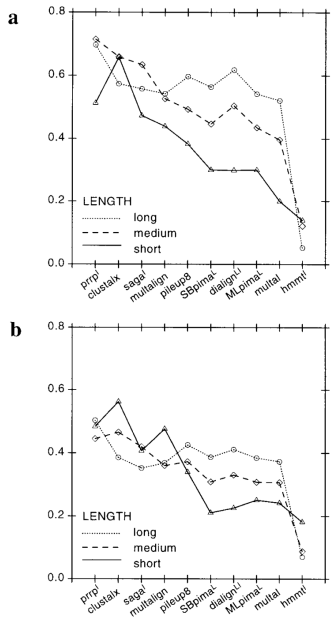


Figure 3. Median SPS for the programs in reference 1 for length categories short, medium and long. Only scores in category V1 are shown. (**a**) Scores based on the core blocks only; (**b**) scores based on the full-length alignment. L, local alignment program; I, iterative alignment program.

Reference 1 - equidistant, similar length
V1 – < 25% identity

Thompson et al, (1999) NAR 27:2682-2690

Appropriate MSA "gold" standards

Published online 4 January 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 7* **2145–2153**
doi:10.1093/nar/gkp1196

Quality measures for protein alignment benchmarks

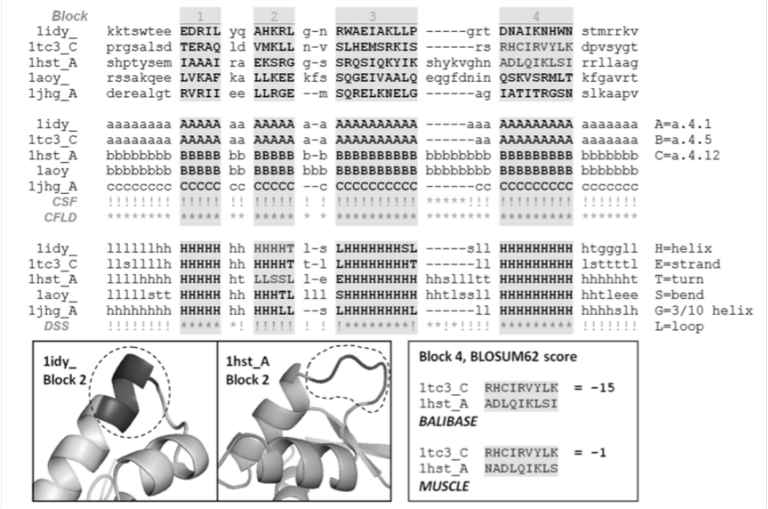
Robert C. Edgar*

Multiple protein sequence alignment methods are central to many applications in molecular biology. These methods are typically assessed on benchmark datasets including BALIBASE, OXBENCH, PREFAB and SABMARK, which are important to biologists in making informed choices between programs. In this article, annotations of domain homology and secondary structure are used to define new measures of alignment quality and are used to make the first systematic, indepen- dent evaluation of these benchmarks. ...

Questionable alignments are found in all benchmarks, especially in BALIBASE where 87% of sequences have unknown structure, 20% of columns contain different folds according to SUPERFAMILY and 30% of 'core block' columns have conflicting secondary structure according to DSSP. A careful analysis of current protein multiple alignment benchmarks calls into question their ability to determine reliable algorithm rankings.

Edgar, R. C. *Nucleic Acids Res* **38,** 2145–2153 (2010).

Appropriate MSA "gold" standards



Edgar, R. C. *Nucleic Acids Res* **38,** 2145–2153 (2010).

T-coffee

T-Coffee: A novel method for fast and accurate multiple sequence alignment.
Notredame C, Higgins DG, Heringa J.

J Mol Biol. 2000 Sep 8;302(1):205-17

We describe a new method (T-Coffee) for multiple sequence alignment that provides a dramatic improvement in accuracy with a modest sacrifice in speed as compared to the most commonly used alternatives. The method is broadly based on the popular progressive approach to multiple alignment but avoids the most serious pitfalls caused by the greedy nature of this algorithm. With T-Coffee we pre-process a data set of all pair-wise alignments between the sequences. This provides us with a library of alignment information that can be used to guide the progressive alignment. Intermediate alignments are then based not only on the sequences to be aligned next but also on how all of the sequences align with each other. This alignment information can be derived from heterogeneous sources such as a mixture of alignment programs and/or structure superposition. Here, we illustrate the power of the approach by using a combination of local and global pair-wise alignments to generate the library. The resulting alignments are significantly more reliable, as determined by comparison with a set of 141 test cases, than any of the popular alternatives that we tried. The improvement, especially clear with the more difficult test cases, is always visible, regardless of the phylogenetic spread of the sequences in the tests.

Molecular Systems Biology 7; Article number 537; doi:10.1038/msb.2011.70
Citation: *Molecular Systems Biology* 7: 537

www.molecularsystemsbiology.com

CORRESPONDENCE

The self-assessment trap: can we all be better than average?

Table I Break out of 57 surveyed papers in which the authors assess their own methods

Number of performance metrics	Total number of studies surveyed	Authors' method is the best in all metrics and all data sets	Authors' method is the best in most metrics and most data sets
1	25	19	6
2	15	13	2
3	7	4	3
4	4	1	3
5	4	1	3
6	2	1	1

Note that we did not find any self-assessment paper where the presented method was not top ranked in at least one metric or data set. The survey was conducted over a large pool of scientific peer-reviewed papers selected as follows. First, a Google Scholar search using the keywords 'computational biology method assessment' was conducted. When papers with comparisons of methods were

Table 2 T-Coffee compared with other multiple sequence alignment methods

Method	Cat1 (81)	Cat2 (23)	Cat3 (4)	Cat4 (12)	Cat5 (11)	Total1 (141)	Total2 (141)	Significance
Dialign	71.0	25.2	35.1	74.7	80.4	61.5	57.3	11.3[a]
ClustalW	78.5	32.2	42.5	65.7	74.3	66.4	58.6	26.2[a]
Prrp	78.6	32.5	50.2	51.1	82.7	66.4	59.0	36.9[a]
T-Coffee	**80.7**	**37.3**	**52.9**	**83.2**	**88.7**	**72.1**	**68.7**	

Method indicates the name of the method evaluated. T-Coffee is the protocol CLE in Table 1. Total1 gives the average accuracy across all the 141 alignments. Total2 is the average accuracy across the five BaliBase categories (unweighted). The last column shows the percentage of times that T-Coffee is outperformed by each other protocol. The statistical significance of the improvement of T-Coffee over each method is shown by

[a] (P<0.001). The Table layout is otherwise similar to that of Table 1.

Notredame et al (2000) JMB 302:205-217

MUSCLE - rapid multiple alignment

MUSCLE: multiple sequence alignment with high accuracy and high throughput.
Edgar RC.
Nucleic Acids Res. 2004 Mar 19;32(5):1792-7

We describe MUSCLE, a new computer program for creating multiple alignments of protein sequences. Elements of the algorithm include fast distance estimation using kmer counting, progressive alignment using a new profile function we call the log-expectation score, and refinement using tree-dependent restricted partitioning. The speed and accuracy of MUSCLE are compared with T-Coffee, MAFFT and CLUSTALW on four test sets of reference alignments: BAliBASE, SABmark, SMART and a new benchmark, PREFAB. MUSCLE achieves the highest, or joint highest, rank in accuracy on each of these sets. Without refinement, MUSCLE achieves average accuracy statistically indistinguishable from T-Coffee and MAFFT, and is the fastest of the tested methods for large numbers of sequences, aligning 5000 sequences of average length 350 in 7 min on a current desktop computer. The MUSCLE program, source code and PREFAB test data are freely available at http://www.drive5. com/muscle.

T-Coffee (top) vs MUSCLE (bot)

```
YES_XIPHE   MGCvrSKEaKgPAlKYqpdNsnvvPvSahlgHYGpeptimg
YES_AVISY   --------------dKgPAmKYrtdNtpePiSshvsHYGsd
YES_CHICK   ------MGCikSKEdKgPAmKYrtdNtpePiSshvsHYGsd
YES_HUMAN   ------MGCikSKEnKsPAiKYrpeNtpePvStsvsHYGae
YES_MOUSE   ------MGCikSKEnKsPAiKYtpeNlteP--vSpsasHYG

YES_XIPHE   MGCvrSKEaKgPAlKYqpdNsnvvPvSahlgHYGpeptimg
YES_AVISY   --------dKgPAmKYrtdNtp-ePiSshvsHYGsdssqat
YES_CHICK   MGCikSKEdKgPAmKYrtdNtp-ePiSshvsHYGsdssqat
YES_HUMAN   MGCikSKEnKsPAiKYrpeNtp-ePvStsvsHYGaepttvs
YES_MOUSE   MGCikSKEnKsPAiKYtpeNlt-ePvSpsasHYGvehatva
```

Figure 1. Motifs misaligned by a progressive method. A set of 41 sequences containing SH2 domains (44) were aligned by the progressive method T-Coffee (above), and by MUSCLE (below). The N-terminal region of a subset of five sequences is shown. The highlighted columns (upper case) are conserved within this family but are misaligned by T-Coffee. It should be noted that T-Coffee aligns these motifs correctly when given these five sequences alone; the problem arises in the context of the other sequences. Complete alignments are available at http://www.drive5.com/muscle.

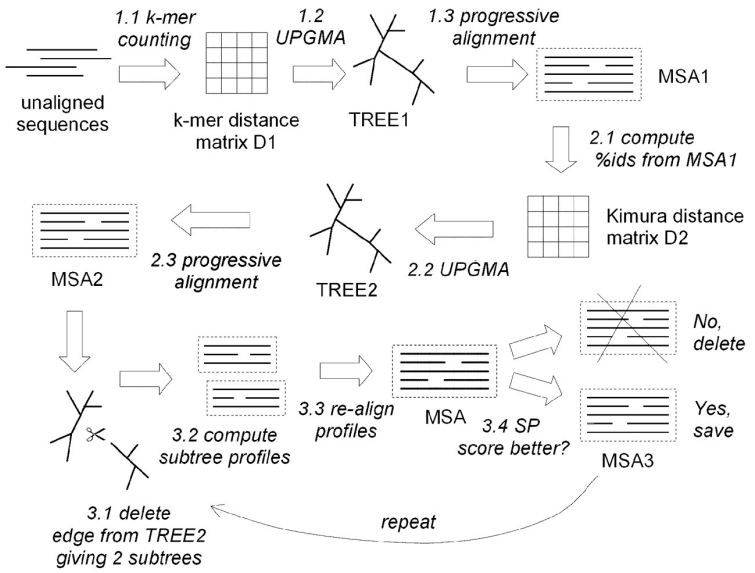


Figure 2. This diagram summarizes the flow of the MUSCLE algorithm. There are three main stages: Stage 1 (draft progressive), Stage 2 (improved progressive) and Stage 3 (refinement). A multiple alignment is available at the completion of each stage, at which point the algorithm may terminate.

Muscle performance (PREFAB)

Table 4. *Q* scores and times on PREFAB

Method	All	0–20%	20–40%	40–70%	70–100%	CPU
MUSCLE	0.645	0.473	0.813	0.937	0.980	1.7×10^4
MUSCLE-p	0.634	0.460	0.802	0.942	0.985	2.0×10^3
T-Coffee	0.615	0.464	0.795	0.935	0.976	1.0×10^6
NWNSI	0.615	0.448	0.772	0.930	0.939	1.4×10^4
FFTNS1	0.591	0.423	0.756	0.931	0.938	1.0×10^3
CLUSTALW	0.563	0.382	0.732	0.916	0.930	3.3×10^4

The average *Q* score for each method over all PREFAB alignments (All), and the total CPU time in seconds are given. The remaining columns show average *Q* scores on subsets in which the structure pairs fall within the given pairwise identity ranges. Note that T-Coffee required 10 CPU days to complete the test, compared with <5 h for MUSCLE and 30 min for MUSCLE-p.

MUSCLE – BaliBase performance

Table 1. BAliBASE scores and times

Method	*Q*	TC	CPU
MUSCLE	0.896	0.747	97
MUSCLE-p	0.883	0.727	52
T-Coffee	0.882	0.731	1500
NWNSI	0.881	0.722	170
CLUSTALW	0.860	0.690	170
FFTNS1	0.844	0.646	16

Table 3. BAliBASE TC scores on subsets

Method	Ref1	Ref2	Ref3	Ref4	Ref5
MUSCLE	0.815	0.574	0.577	0.627	0.902
MUSCLE-p	0.795	0.558	0.550	0.598	0.891
T-Coffee	0.780	0.573	0.510	0.751	0.903
NWNSI	0.788	0.514	0.514	0.742	0.859
CLUSTALW	0.782	0.579	0.470	0.542	0.638
FFTNS1	0.732	0.496	0.350	0.451	0.831

Average *Q* and TC scores for each method on BAliBASE are shown, together with the total CPU time in seconds. Align-m aborted on two alignments; average scores on the remainder were *Q* = 0.852 and TC = 0.670, requiring 2202 s.

Optimal (MSA2.1) vs ClustalW/TCoffee/MUSCLE

N-terminal

```
MSA optimal multiple alignment
GSTA1_HUMAN  MA--EKPKLHYFNARGRMESTRWLLAAAGVEFEEKFIKSAE-------DLDKLRNDGYLMFQQVPMVEIDGMKLVQTRAIL
GSTP1_HUMAN  MP---PYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTVET--------WQEGSLKASCLYGQLPKFQDGDLTLYQSNTIL
GSTM1_HUMAN  MP----MILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKITQSNAIL
GSTM3_HUMAN  MSCESSMVLGYWDIRGLAHAIRLLLEFTDTSYEEKRYTCGEAPDYDRSQWLDVKFKLDLDFPNLPYLLDGKNKITQSNAIL
                       *   **     * **        *                                 *          *   **
                       !     #                                 !            !           !
CLUSTAL 2.0.12 multiple sequence alignment
GSTA1_HUMAN  --MAEKPKLHYFNARGRMESTRWLLAAAGVEFEEKFIKS-------AEDLDKLRNDGYLMFQQVPMVEIDGMKLVQTRAIL
GSTP1_HUMAN  ---MPPYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTV--------ETWQEGSLKASCLYGQLPKFQDGDLTLYQSNTIL
GSTM1_HUMAN  ----MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKITQSNAIL
GSTM3_HUMAN  MSCESSMVLGYWDIRGLAHAIRLLLEFTDTSYEEKRYTCGEAPDYDRSQWLDVKFKLDLDFPNLPYLLDGKNKITQSNAIL
                       *   **     * **        *                                 *          *   **
MUSCLE (3.8) multiple sequence alignment
GSTA1_HUMAN  MAEKPKLH--YFNARGRMESTRWLLAAAGVEFEEKFIKSAE----DLDKLRNDGY---LMFQQVPMVEIDGMKLVQTRAIL
GSTP1_HUMAN  ---MPPYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTV--------ETWQEGSLKASCLYGQLPKFQDGDLTLYQSNTIL
GSTM1_HUMAN  ----MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKITQSNAIL
GSTM3_HUMAN  MSCESSMVLGYWDIRGLAHAIRLLLEFTDTSYEEKRYTCGEAPDYDRSQWLDVKFKLDLDFPNLPYLLDGKNKITQSNAIL
                       *   **     * **        *                                 *          *   **
CLUSTAL FORMAT for T-COFFEE Version_8.69
GSTA1_HUMAN  MAE--KPKLHYFNARGRMESTRWLLAAAGVEFEEKFIKSAEDL-------DKLRNDGYLMFQQVPMVEIDGMKLVQTRAIL
GSTP1_HUMAN  MP---PYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTVE--------TWQEGSLKASCLYGQLPKFQDGDLTLYQSNTIL
GSTM1_HUMAN  MP----MILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAHKITQSNAIL
GSTM3_HUMAN  MSCESSMVLGYWDIRGLAHAIRLLLEFTDTSYEEKRYTCGEAPDYDRSQWLDVKFKLDLDFPNLPYLLDGKNKITQSNAIL
             *         *   **     * **        *                                 *          *   **
```

Optimal (MSA2.1) vs ClustalW/TCoffee/MUSCLE

center

```
MSA optimal multiple alignment
GSTA1_HUMAN  NYIASKYNLYGKDIKERALIDMYIEGIADLGEMILLLPVCPPEEKDAKLALIKEKIKNRYFPAFEKVLKSH--GQDYLVGNKL
GSTP1_HUMAN  RHLGRTLGLYGKDQQEAALVDMVNDGVEDLRCKYISL-IY-TNYEAGKDDYVKA-LPGQ-LKPFETLLSQNQGGKTFIVGDQI
GSTM1_HUMAN  CYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMI-CYNPEFEKLKPKYLEE-LPEK-LKLYSEFLGKR----PWFAGNKI
GSTM3_HUMAN  RYIARKHNMCGETEEEKIRVDIIENQVMDFRTQLIRL-CYSSDHEKLKPQYLEE-LPGQ-LKQFSMFLGKF----SWFAGEKL
                     *    *    *       *
                                          #
CLUSTAL 2.0.12 multiple sequence alignment
GSTA1_HUMAN  NYIASKYNLYGKDIKERALIDMYIEGIADLGEMILLLPVCPPEEKDAKLALIKEKIKNRYFPAFEKVLKSHGQDYLVGNKL
GSTP1_HUMAN  RHLGRTLGLYGKDQQEAALVDMVNDGVEDLRCKYISLIYT-NYEAGKDDYVKALPGQLKPFETLLSQNQGG-KTFIVGDQI
GSTM1_HUMAN  CYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPEFEKLKPKYLEELPEKLKLYSEFLGK-----RPWFAGNKI
GSTM3_HUMAN  RYIARKHNMCGETEEEKIRVDIIENQVMDFRTQLIRLCYSSDHEKLKPQYLEELPGQLKQFSMFLGK-----FSWFAGEKL
                     *    *    *       *                *                          *
MUSCLE (3.8) multiple sequence alignment
GSTA1_HUMAN  NYIASKYNLYGKDIKERALIDMYIEGIADLGEMILLLPVCPPEE--KDAKLALIKEKIKNRYFPAFEKVL--KSHGQDYLVGNKL
GSTP1_HUMAN  RHLGRTLGLYGKDQQEAALVDMVNDGVEDLRCKYISLIYT-NYEAGKDDYVKALPGQLK-----PFETLLSQNQGGKTFIVGDQI
GSTM1_HUMAN  CYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPEFEKLKPKYLEELPEKLK-----LYSEFL----GKRPWFAGNKI
GSTM3_HUMAN  RYIARKHNMCGETEEEKIRVDIIENQVMDFRTQLIRLCYSSDHEKLKPQYLEELPGQLK-----QFSMFL----GKFSWFAGEKL
                     *    *    *       *                *  *          *          *           *
CLUSTAL FORMAT for T-COFFEE Version_8.69 SCORE=88, Nseq=4, Len=240
GSTA1_HUMAN  NYIASKYNLYGKDIKERALIDMYIEGIADLGEMILLLPVCPPEEKDAKLALIKEKIKNRYFPAFEKVLKSH---------GQDYLVGNKL
GSTP1_HUMAN  RHLGRTLGLYGKDQQEAALVDMVNDGVEDLRCKYISLIYT-N----------YEAGKDDYVKALPGQLKPFETLLSQNQGGKTFIVGDQI
GSTM1_HUMAN  CYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPE----------FEKLKPKYLEELPEKLKLYSEFL----GKRPWFAGNKI
GSTM3_HUMAN  RYIARKHNMCGETEEEKIRVDIIENQVMDFRTQLIRLCYSSD----------HEKLKPQYLEELPGQLKQFSMFL----GKFSWFAGEKL
                     *    *    *       *                      *  *  *      **              *
```


Optimal (MSA2.1) vs ClustalW/TCoffee/MUSCLE

C-terminal

```
MSA optimal multiple alignment
GSTA1_HUMAN  SRADIHLVELLYYVEELDSSLISSFPLLKALKTRISNLPTVKKFLQPGSPRKPPMDEKSLEEARKIFRF
GSTP1_HUMAN  SFADYNLLDLLLIHEVLAPGCLDAFPLLSAYVGRLSARPKLKAFLASPEYVNLPINGNGKQ--------
GSTM1_HUMAN  TFVDFLVYDVLDLHRIFEPKCLDAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGNK----
GSTM3_HUMAN  TFVDFLTYDILDQNRIFDPKCLDEFPNLKAFMCRFEALEKIAAYLQSDQFCKMPINNKMAQWGNKPVC-
               *      *             ** *     *                   *
CLUSTAL 2.0.12 multiple sequence alignment
GSTA1_HUMAN  SRADIHLVELLYYVEELDSSLISSFPLLKALKTRISNLPTVKKFLQPGSPRKPPMDEKSLEEARKIFRF
GSTP1_HUMAN  SFADYNLLDLLLIHEVLAPGCLDAFPLLSAYVGRLSARPKLKAFLASPEYVNLPINGNGKQ--------
GSTM1_HUMAN  TFVDFLVYDVLDLHRIFEPKCLDAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGNK----
GSTM3_HUMAN  TFVDFLTYDILDQNRIFDPKCLDEFPNLKAFMCRFEALEKIAAYLQSDQFCKMPINNKMAQWGNKPVC-
               *      *             ** *     *                   *
MUSCLE (3.8) multiple sequence alignment
GSTA1_HUMAN  SRADIHLVELLYYVEELDSSLISSFPLLKALKTRISNLPTVKKFLQPGSPRKPPMDEKSLEEARKIFRF
GSTP1_HUMAN  SFADYNLLDLLLIHEVLAPGCLDAFPLLSAYVGRLSARPKLKAFLASPEYVNLPINGN----GKQ----
GSTM1_HUMAN  TFVDFLVYDVLDLHRIFEPKCLDAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGNK----
GSTM3_HUMAN  TFVDFLTYDILDQNRIFDPKCLDEFPNLKAFMCRFEALEKIAAYLQSDQFCKMPINNKMAQWGNKPVC-
               *      *             ** *     *                   *
CLUSTAL FORMAT for T-COFFEE Version_8.69 SCORE=88, Nseq=4, Len=240
GSTA1_HUMAN  SRADIHLVELLYYVEELDSSLISSFPLLKALKTRISNLPTVKKFLQPGSPRKPPMDEKSLEEARKIFRF
GSTP1_HUMAN  SFADYNLLDLLLIHEVLAPGCLDAFPLLSAYVGRLSARPKLKAFLASPEYVNLPINGNGK--------Q
GSTM1_HUMAN  TFVDFLVYDVLDLHRIFEPKCLDAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGN----K
GSTM3_HUMAN  TFVDFLTYDILDQNRIFDPKCLDEFPNLKAFMCRFEALEKIAAYLQSDQFCKMPINNKMAQWGNK-PVC
               *      *             ** *     *                   *
```

Trees, stars, and multiple alignment

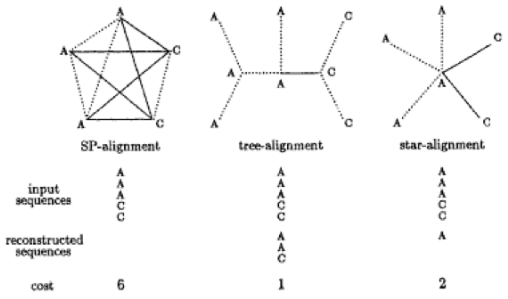


Fig. 1 SP-, tree-, and star-alignments for five, one-letter, input sequences. Pairwise alignments with cost one are indicated by solid lines, and pairwise alignments with cost zero are indicated by dotted lines.

Altschul, S. F. and Lipman, D. J. (1989) SIAM J. Appl. Math. 49:197-209

Phylogenetic alignment I: BAli-Phy

Summary: BAli-Phy is a Bayesian posterior sampler that employs Markov chain Monte Carlo to explore the joint space of alignment and phylogeny given molecular sequence data. Simultaneous estimation eliminates bias toward inaccurate alignment guide-trees, employs more sophisticated substitution models during alignment and automatically utilizes information in shared insertion/deletions to help infer phylogenies.

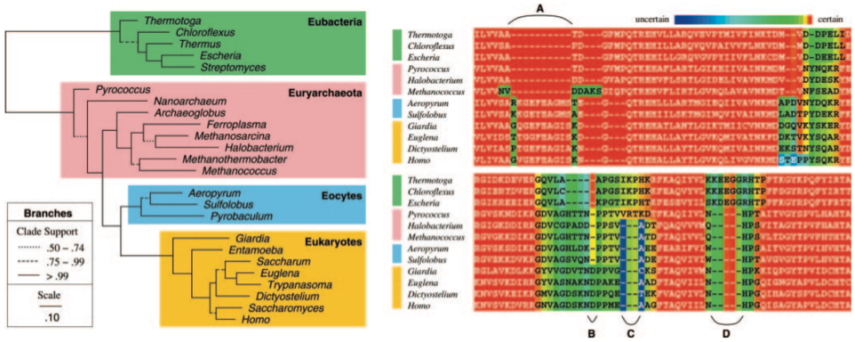


Fig. 1. Maximum a posteriori topology for 24 EF-1α/Tu sequences across the Tree of Life (left) and two separate portions of the alignment uncertainty (AU) plot (right) for a subsample of these sequences. Branch lengths equal posterior mean estimates and line-style depicts partition credibility. In the AU plot, well-resolved entries have a red background, whereas less certain entries have backgrounds tending towards violet based on an approximate probability that each entry is homologous with a residue at the root in each column. Four different types of topologically informative insertion/deletion events (A, B, C and D) are highlighted.

Suchard and Redelings (2006) Bioinfo. 22:2047

Class of Multiple Sequence Alignment Algorithm Affects Genomic Analysis
Blackburne and Whelan (2012) Mol. Biol. Evol. 30:642-653

Multiple sequence alignment (MSA) is the heart of comparative sequence analysis. Recent studies demonstrate that MSA algorithms can produce different outcomes when analyzing genomes, including phylogenetic tree inference and the detection of adaptive evolution. These studies also suggest that the difference between MSA algorithms is of a similar order to the uncertainty within an algorithm and suggest integrating across this uncertainty. In this study, we examine further the problem of disagreements between MSA algorithms and how they affect downstream analyses. We also investigate whether integrating across alignment uncertainty affects downstream analyses. We address these questions by analyzing 200 chordate gene families, with properties reflecting those used in large-scale genomic analyses. We find that newly developed distance metrics reveal two significantly different classes of MSA methods (MSAMs). The similarity-based class includes progressive aligners and consistency aligners, representing many methodological innovations for sequence alignment, whereas the evolution-based class includes phylogenetically aware alignment and statistical alignment. We proceed to show that the class of an MSAM has a substantial impact on downstream analyses. For phylogenetic inference, tree estimates and their branch lengths appear highly dependent on the class of aligner used. The number of families, and the sites within those families, inferred to have undergone adaptive evolution depend on the class of aligner used. Similarity-based aligners tend to identify more adaptive evolution. We also develop and test methods for incorporating MSA uncertainty when detecting adaptive evolution but find that although accounting for MSA uncertainty does affect downstream analyses, it appears less important than the class of aligner chosen. Our results demonstrate the critical role that MSA methodology has on downstream analysis, highlighting that the class of aligner chosen in an analysis has a demonstrable effect on its outcome.

Phylogenetic alignment predicts less selection

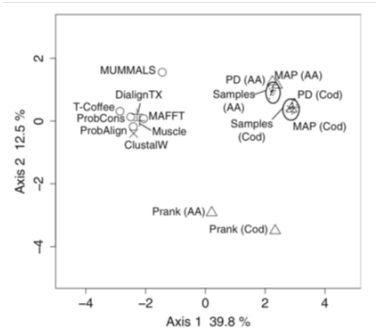


FIG. 1. PCoA plot of mean alignment distances (d_{evol}) for alignments made across 200 data sets from The Adaptive Evolution Database. "PD" and "MAP" refer to the BAli-Phy posterior decoding and maximum a posteriori summary alignments. "Samples" refers to the 20 samples taken from each BAli-Phy run.

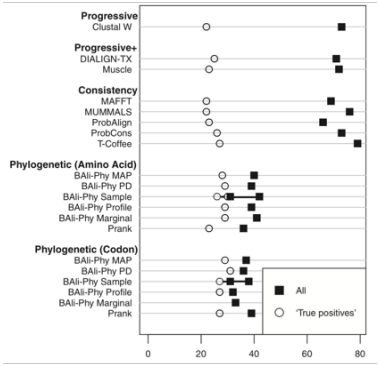


FIG. 3. Total number of families (out of 200) inferred to be under adaptive evolution (P 0:05) found, and the number of families that agree with the BAli-Phy Marginal Codon estimate (putative "true positives," see text).

Blackburne and Whelan (2012) Mol. Biol. Evol 30:642

Summary

- *No multiple alignments without **HOMOLOGY***
- Multiple sequence alignments can resolve ambiguous gaps – largely used to specify gap positions
- Produce profiles, better structural alignments, alignments for phylogenies
- "Classic" programs build successive pairwise alignments (progressive alignment) – Clustal-W, T-coffee, MUSCLE
- Simple progressive alignment methods fix gaps early, after which they cannot be moved
- Iterative approaches required to adjust gaps
- What are appropriate tests – alignments for trees vs alignments for structures?

Multiple Alignments for Phylogenies (for protein coding genes)

- Always align protein sequences
- Use protein alignments to drive DNA codon alignments (all gaps between codons)

```
GTM1_HUMAN ----MPMILG YWDIRGLAHA IRLLLEYTDS SYEEKKYTMG DAPDYDRSQW
GTM2_HUMAN ----MPMTLG YWNIRGLAHS IRLLLEYTDS SYEEKKYTMG DAPDYDRSQW
GTM3_HUMAN MSCESSMVLG YWDIRGLAHA IRLLLEFTDT SYEEKRYTCG EAPDYDRSQW
GTM4_HUMAN ----MSMTLG YWDIRGLAHA IRLLLEYTDS SYEEKKYTMG DAPDYDRSQW
GTM5_HUMAN ----MPMTLG YWDIRGLAHA IRLLLEYTDS SYVEKKYTMG DAPDYDRSQW
GTM1_MOUSE ----MPMILG YWNVRGLTHP IRMLLEYTDS SYDEKRYTMG DAPDFDRSQW
GTM2_MOUSE ----MPMTLG YWDIRGLAHA IRLLLEYTDT SYEDKKYTMG DAPDYDRSQW
GTM3_MOUSE ----MPMTLG YWNTRGLTHS IRLLLEYTDS SYEEKRYVMG DAPNFDRSQW
GTM5_MOUSE MSSKS-MVLG YWDIRGLAHA IRMLLEFTDT SYEEKRYICG EAPDYDRSQW

GTM1_HUMAN   ---------- --ATGCCCAT GATACTGGGG TACTGGGACA TCCGCGGGCT
GTM2_HUMAN   ---------- --ATGCCCAT GACACTGGGG TACTGGAACA TCCGCGGGCT
GTM3_HUMAN   ATGTCGTGCG AGTCGTCTAT GGTTCTCGGG TACTGGGATA TTCGTGGGCT
GTM4_HUMAN   ---------- --ATGTCCAT GACACTGGGG TACTGGGACA TCCGCGGGCT
GTM5_HUMAN   ---------- --ATGCCCAT GACTCTGGGG TACTGGGACA TCCGTGGGCT
GTM1_MOUSE   ---------- --ATGCCTAT GATACTGGGA TACTGGAACG TCCGCGGACT
GTM2_MOUSE   ---------- --ATGCCTAT GACACTAGGT TACTGGGACA TCCGTGGGCT
GTM3_MOUSE   ---------- --ATGCCTAT GACACTGGGC TATTGGAACA CCCGCGGACT
GTM5_MOUSE   ATGTCATCCA AGTCT---AT GGTTCTGGGT TACTGGGATA TCCGCGGGCT
```

First exam sample questions– 2 hours, collab Due Monday, Feb. 26, 5:00 PM

1. Statistical estimates based on sequence shuffling on the fasta.bioch web site typically shows the expectation value as E(10,000).
   a) What does E(10,000) mean?
   b) Since only two sequences are being compared, why does it make sense to present E(10,000)? What E() context would be more appropriate?
2. In the similarity searching exercise, you were asked to find the highest scoring non-homolog in the search.
   a) Assuming the statistical estimates are accurate, what should the Expect (E()-value) be for the highest scoring non-homolog (approx.)?
   b) are all sequences with scores worse than the highest scoring non-homolog non-homologous?
3. Expectation values -.
   a) What is the range of Expect values (smallest and largest) in a database search of the human proteome, with 44,000 proteins?
   b) Expect values are corrected by the size of the database for a single query; E()<0.001 means that a score this good would occur less than once in 1000 searches by chance. What Expect threshold should you choose if you wanted a 1% (0.01) chance of getting a similarity score by chance after a large scale genome analysis that required 10,000 searches?
   c) What kinds of errors might occur because you adjusted the Expect threshold to the value you chose in part (b)?

First exam sample questions– 2 hours, collab Due Monday, Feb. 26 at 5:00 PM

4. A Pfam annotation suggests that a domain with model length 200 aligns in two places to a 150 residue protein. One location has (seq_start,seq_end) = (1,60), with (hmm_start,hmm_end) = (11,70), while the other location has (seq_start, seq_end)=(61,150) and (hmm_start, hmm_end) = (111,200).
   a) Do these mappings of domain regions make biological sense? Why or why not?
   b) Give an explanation for the annotation that makes biological sense.
   c) Give an explanation for the annotation that suggests some kind of artifact.
5. What is the expectation (E()) for a pairwise alignment with a score of 45 bits between two average length proteins (400 aa) in a search of the human proteome (44,000 proteins)
   a) If the 45 bit score were produced by a 200 residue alignment, what is the expected percent identity (approximately) and what scoring matrix should be used?
   b) If the score were produced by a 50 residue alignment, what would be the best scoring matrix and expected percent identity?
6. Why would raising the gap penalty improve the E()-value for very closely related sequences, but reduce significance (increase the E()-value) for distantly related sequences?