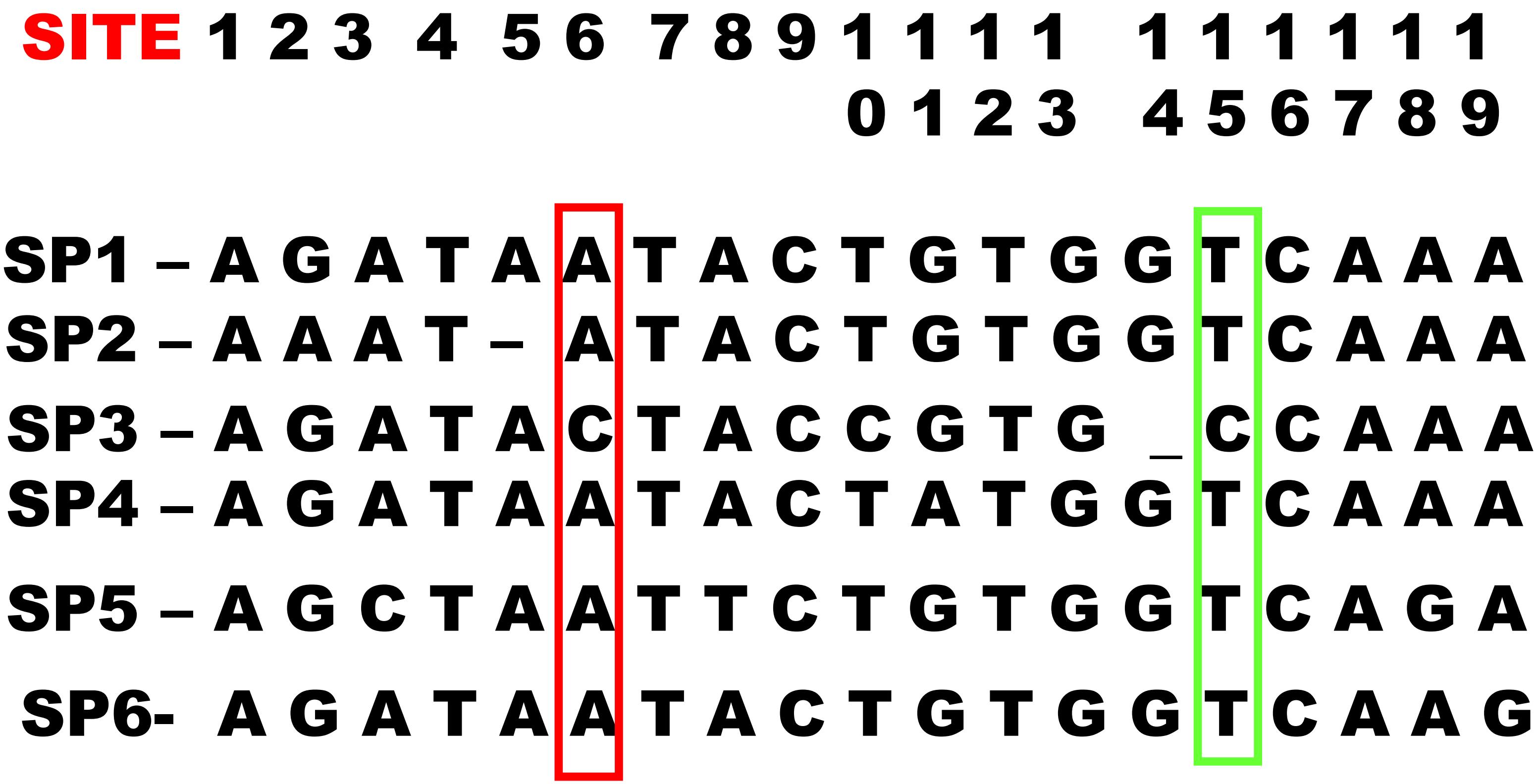
Alinhamentos – as homologias



SITE	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
SP1 –	A	G	A	T	A	A	T	A	C	T	G	T	G	G	T	C	A	A	A
SP2 –	A	A	A	T	–	A	T	A	C	T	G	T	G	G	T	C	A	A	A
SP3 –	A	G	A	T	A	C	T	A	C	C	G	T	G	_	C	C	A	A	A
SP4 –	A	G	A	T	A	A	T	A	C	T	A	T	G	G	T	C	A	A	A
SP5 –	A	G	C	T	A	A	T	T	C	T	G	T	G	G	T	C	A	G	A
SP6-	A	G	A	T	A	A	T	A	C	T	G	T	G	G	T	C	A	A	G

Ascendente comum/extant species

Alinhamentos – tipos de dados

Sequencias de proteinas

Protein-coding DNA (transcribed and translated)

Ribosomal DNA (transcribed)

Non-Coding DNA (control region, introns, pseudogenes)

Alinhamentos – algoritmos

Needleman-Wunsch – alinhamento global –primeira aplicação de programação dinâmica à comparação de sequências

Smith-Waterman – algoritmo de programação dinâmica que asssegura o alinhamento local óptimo para uma dada matrix de substituição - lento

Blast - Basic Local Alignment Search Tool - heuristic approach that approximates the Smith-Waterman algorithm

Clustal – Progressive alignment algorithm

Malign – Progressive alignment algorithm but adds an additional alignment technique beyond clustal, by searching for an optimal guide tree

Alinhamentos – Blast

The BLAST algorithm can be conceptually divided into three stages.

1) BLAST searches for exact matches of a small fixed length W between the query and sequences in the database. For example, given the sequences AGTTAC and ACTTAG and a word length W = 3, BLAST would identify the matching substring TTA that is common to both sequences. By default, W = 11 for nucleic seeds.

2) BLAST tries to extend the match in both directions, starting at the seed. The ungapped alignment process extends the initial seed match of length W in each direction in an attempt to boost the alignment score. Insertions and deletions are not considered during this stage. For our example, the ungapped alignment between the sequences AGTTAC and ACTTAG centered around the common word TTA would be:

```
..AGTTAC..
  |   || |
..ACTTAG..
```

Alinhamentos – Blast

If a high-scoring un-gapped alignment is found, the database sequence is passed on to the third stage.

3) BLAST performs a gapped alignment between the query sequence and the database sequence using a variation of the Smith-Waterman algorithm. Statistically significant alignments are then displayed to the user.

Clustal – Multiple alignments

1 - Comparação das sequencias duas a duas – Pairwise alignment.

2- Uso desta relação de proximidade para
traçar um dendrograma (árvore)

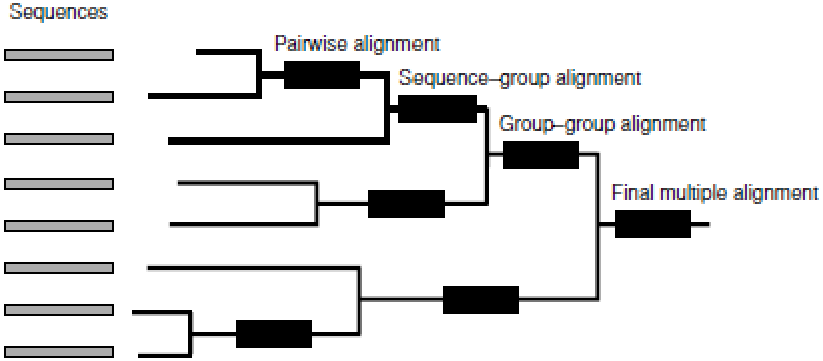
3- uso do dendrograma como guia para realizar
um alinhamento multiplo final (árvore). Começando por alinhar as sequencias com menor custo de alinhamento e assim sucessivamente

Final step – always adjust by eye

Alignments

1) Clustal

Widely used for progressive alignments



2) MAFFT

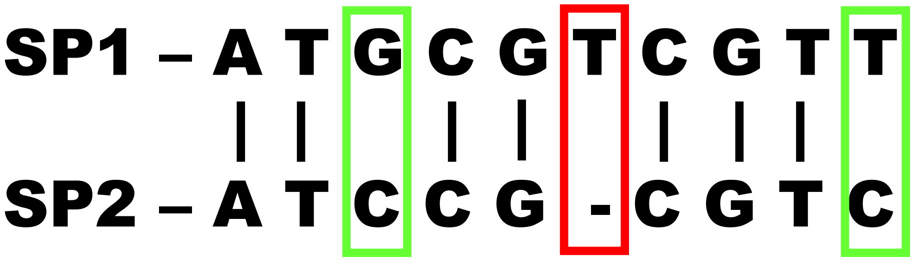
Multiple alignment with iterative refinement and consistency-based scoring approaches

Exemplo de alinhamento pairwise

SP1 – ATGCGTCGTT

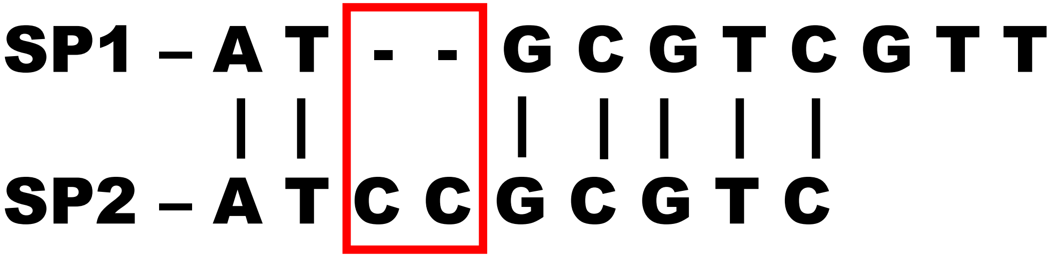
SP2 – ATCCGCGTC

Alinhamento 1



```
SP1 – A T G C G T C G T T
      | |   | |   | | |
SP2 – A T C C G - C G T C
```

Alinhamento 2



```
SP1 – A T - - G C G T C G T T
      | |     | | | | |
SP2 – A T C C G C G T C
```

Exemplo

Alinhamento 1

```
SP1 – A T G C G T C G T T
      | |   | |   | | |
SP2 – A T C C G - C G T C
```

Alinhamento 2

```
SP1 – A T - - G C G T C G T T
      | |     | | | | |
SP2 – A T C C G C G T C
```

D=s+wg	w=1	w=3
D=2+w1	D=3	*D=5*
D=0+w2	*D=2*	D=6

Generalizando

$$D = \min(Y + \sum_{1}^{n} W_k Z_k)$$

Y é o numero de substituições

K varia de 1 a n (n é maior gap do alinhamento)

Z é o número de gaps de tamanho K

$$S = \max(X - \sum_{1}^{n} W_k Z_k)$$

X é o número posições semelhantes entre as sequencias

Gap penalties

Número de gaps de tamanho k

$$S = \max(X - \sum_{1}^{n} W_k Z_k)$$

Penalidades para gaps de tamanho k

$$W_k = a + bk \qquad W_k = a + b * \ln k$$

a – gap open penalty

b – gap extension penalty

k – gap size

Numerical example - melhor

$S = \max(X - \Sigma W_k Z_k)$

$w_k = a + bk$

$a=3 \quad b=0.4$

Alinhamento 1

```
SP1 – A T G C G T C - - G T T G
      | |   | |   |     | |   |
SP2 – A T C C G - C A A G T C G
```

$S=8-((3+0.4*1)*1)+((3+0.4*2)*1)=0.8$

Numerical example

$$S = \max(X - \Sigma W_k Z_k)$$

$$w_k = a + bk$$

$a=3$ $b=0.4$

Alinhamento 2

```
SP1 – A T  -  -  G C  -  -  G T C G T T G
      | |        |  |       | | | |
SP2 – A T  C  C  G C  A  A  G T C G
```

$$S=8-((3+0.4*2)*2)=0.4$$

DNA weight matrix – Substitution matrices

IUB – Default. Xs and Ns are treated as matches to any IUB ambiguity symbol. All matches score 1.9; all mismatches for IUB symbols score 0.

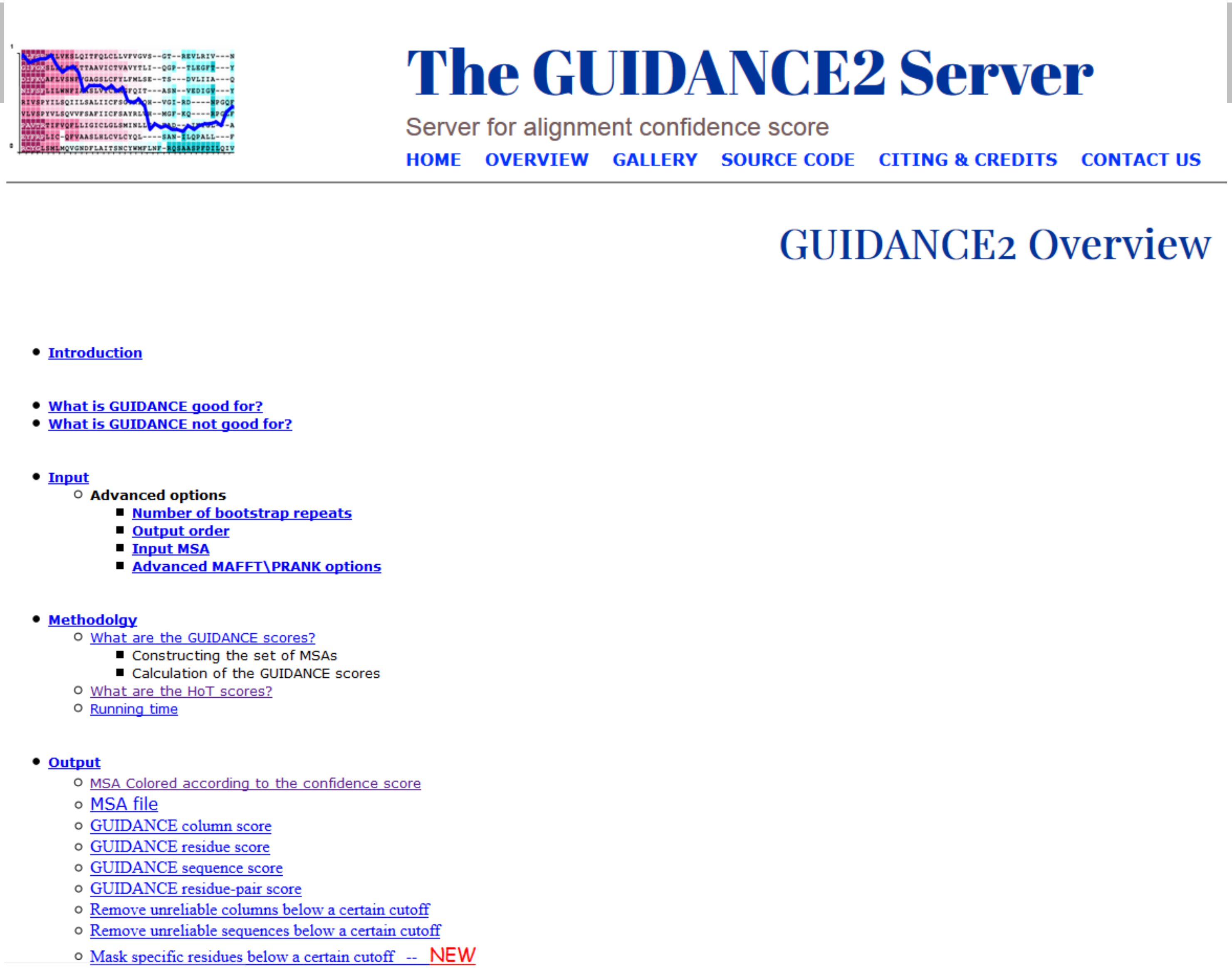
ClustalW 1.6 – Matches score 1 and mismatches score 0.
All matches for IUB symbols also score 0

MAFFT

*L-INS-i (probably most accurate; recommended for <200 sequences; iterative refinement method incorporating local pairwise alignment information):

*G-INS-i (suitable for sequences of similar lengths; recommended for <200 sequences; iterative refinement method incorporating global pairwise alignment information):

*E-INS-i (suitable for sequences containing large unalignable regions; recommended for <200 sequences):



The GUIDANCE2 Server

Server for alignment confidence score

GUIDANCE2 Overview

- Introduction

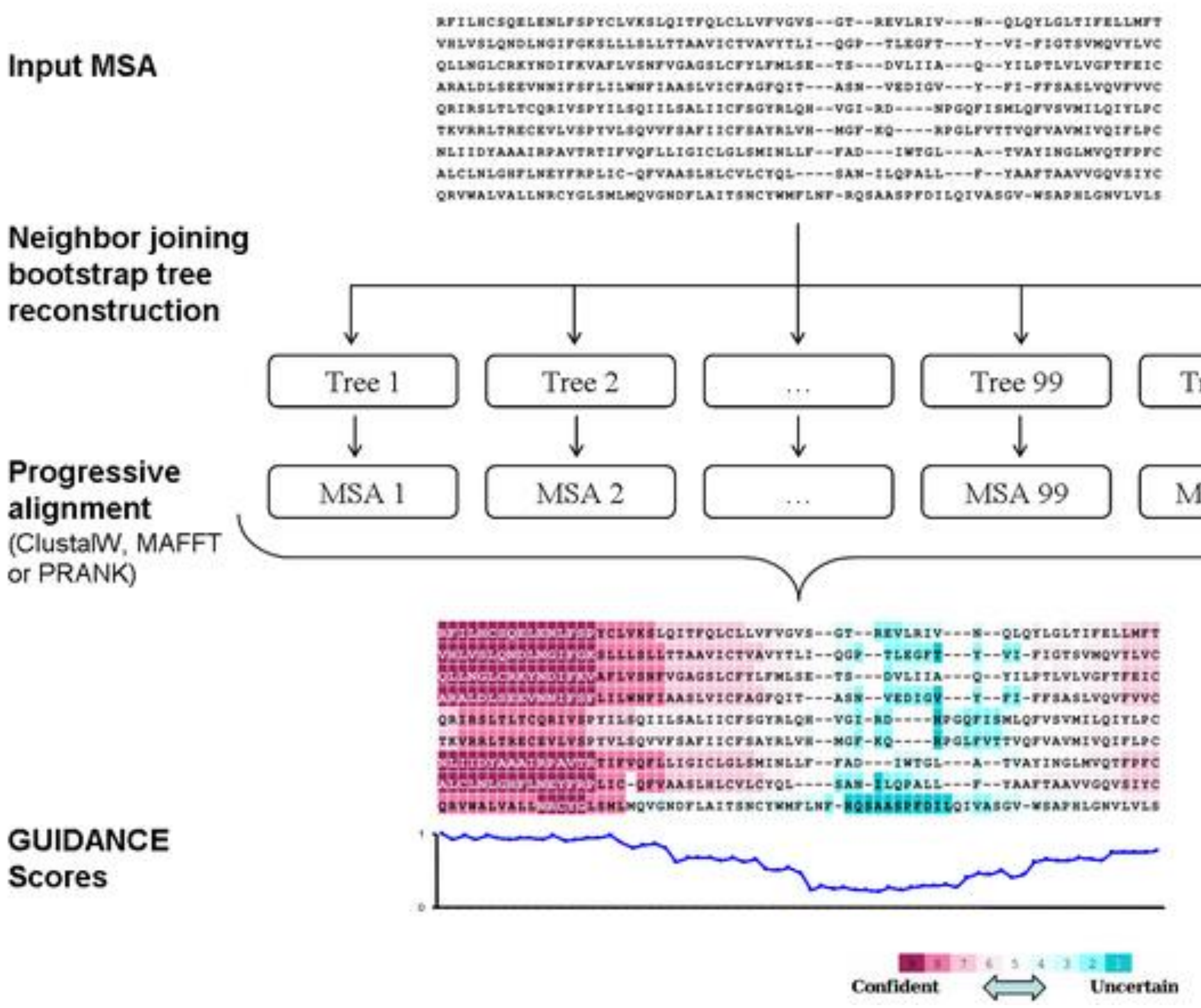
- What is GUIDANCE good for?
- What is GUIDANCE not good for?

- Input
  - Advanced options
    - Number of bootstrap repeats
    - Output order
    - Input MSA
    - Advanced MAFFT\PRANK options

- Methodolgy
  - What are the GUIDANCE scores?
    - Constructing the set of MSAs
    - Calculation of the GUIDANCE scores
  - What are the HoT scores?
  - Running time

- Output
  - MSA Colored according to the confidence score
  - MSA file
  - GUIDANCE column score
  - GUIDANCE residue score
  - GUIDANCE sequence score
  - GUIDANCE residue-pair score
  - Remove unreliable columns below a certain cutoff
  - Remove unreliable sequences below a certain cutoff
  - Mask specific residues below a certain cutoff -- NEW

Guidance2 Server



```
RFILHCSQELENLFSPYCLVKSLQITFQLCLLVFVGVS--GT--REVLRIV---N--QLQYLGLTIFELLMFT
VHLVSLQNDLNGIFGKSLLLSLLTTAAVICTVAVYTLI--QGP--TLEGFT---Y--VI-FIGTSVMQVYLVC
QLLNGLCRKYNDIFKVAFLVSNFVGAGSLCFYLFMLSE--TS---DVLIIA---Q--YILPTLVLVGFTFEIC
ARALDLSEEVNNIFSFLILWNFIAASLVICFAGFQIT---ASN--VEDIGV---Y--FI-FFSASLVQVFVVC
QRIRSLTLTCQRIVSPYILSQIILSALIICFSGYRLQH--VGI-RD----NPGQFISMLQFVSVMILQIYLPC
TKVRRLTRECEVLVSPYVLSQVVFSAFIICFSAYRLVH--MGF-KQ----RPGLFVTTVQFVAVMIVQIFLPC
NLIIDYAAAIRPAVTRTIFVQFLLIGICLGLSMINLLF--FAD---IWTGL---A--TVAYINGLMVQTFPFC
ALCLNLGHFLNEYFRPLIC-QFVAASLHLCVLCYQL----SAN-ILQPALL---F--YAAFTAAVVGQVSIYC
QRVWALVALLNRCYGLSMLMQVGNDFLAITSNCYWMFLNF-RQSAASPFDILQIVASGV-WSAPHLGNVLVLS
```

Bibliografia

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucelic Acids Research*, *32*, 1792–1797. Retrieved from http://nar.oxfordjournals.org/content/32/5/1792.full.pdf+html

Higgins, D. G., & Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, *73*(1), 237–244. doi:10.1016/0378-1119(88)90330-7

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. doi:10.1093/molbev/mst010

Katoh, K., & Standley, D. M. (2014). MAFFT: iterative refinement and additional methods. *Methods in Molecular Biology*, *1079*, 131–146. doi:10.1007/978-1-62703-646-7_8

Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., & Pupko, T. (2010). GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Research*, *38*(Web Server issue), W23-8. doi:10.1093/nar/gkq443