

Práctica 3 - Mafft

Pablo Luesia
Juan José Gómez
David Guimerá

Mafft - Introducción

- ▶ Es una herramienta para alineamiento de secuencias.
- ▶ Esta disponible tanto online como para descargar en los principales sistemas operativos.

Mafft - ¿Cómo usarlo?

- ▶ Se puede descargar desde la web oficial para diferentes sistemas operativos (Mac OS X, Linux y Windows).
- ▶ También tiene una versión web (la que vamos a utilizar en la sesión)

Mafft - Versión web

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

Download version

[Mac OS X](#)

[Windows](#)

[Linux](#)

[Source](#)

[Online version](#)

[Alignment](#)

[mafft --add](#)

[Merge](#)

[Phylogeny](#)

[Rough tree](#)

[Merits / limitations](#)

[Algorithms](#)

[Tips](#)

[Benchmarks](#)

[Feedback](#)



Contact email address, kazutaka.katoh@aist.go.jp, is temporarily unavailable from email to this address but have received no response, then please re-send the email.

About

MAFFT is a multiple sequence alignment program for unix-like operating systems. It offers a range of multiple alignment methods (sequences), FFT-NS-2 (fast; for alignment of <~30,000 sequences), etc.

Download and Installation

- [Mac OS X](#)
- [Linux](#)
- [Windows](#) **Updated!**
- [Source](#)
- [Changelog](#)

The latest version is 7.392 (2018/Mar/3).

Input Format

Fasta format. [example1 \(LSU rRNA\)](#), [example2 \(protein\)](#)

The type of input sequences (amino acid or nucleotide) is automatically recognized.

Mafft - Características

- ▶ La base de su algoritmo esta fundamentado en la Transformada Rápida de Fourier (FFT).
- ▶ Ofrece un amplio abanico de algoritmos, los cuales estas orientados a dos objetivos: precisión de los resultados o eficiencia.

Selección algoritmo Mafft - Scroll versión online

Advanced settings

Strategy:

- ☒ Auto (FFT-NS-1, FFT-NS-2, FFT-NS-i or L-INS-i; depends on data size) [Updated](#)

Progressive methods

- ☐ FFT-NS-1 (Very fast; recommended for >2,000 sequences; progressive method)
- ☐ FFT-NS-2 (Fast; progressive method)
- ☐ G-INS-1 (Slow; progressive method with an accurate guide tree)

Iterative refinement methods

- ☐ FFT-NS-i (Slow; iterative refinement method)
- ☐ E-INS-i (Very slow; recommended for <200 sequences with multiple conserved domains and long gaps) [Help](#)
- ☐ L-INS-i (Very slow; recommended for <200 sequences with one conserved domain and long gaps) [Help](#)
- ☐ G-INS-i (Very slow; recommended for <200 sequences with global homology) [Help](#)
- ☐ Q-INS-i (Extremely slow; secondary structure of RNA is considered; recommended for a global alignment of h nucleotides; the number of iterative cycles is restricted to two, 2016/May) [Help](#)

Align unrelated segments, too? *in Alpha Testing* (2014/Mar)

If the input data is expected to be globally conserved but locally contaminated by unrelated segments, try 'Unalign

Mafft - Algoritmos orientados a precisión

- ▶ **L-INS-i**: método de los mas precisos. Recomendado cuando se pretenden alinear menos de 200 secuencias. Es iterativo.
- ▶ **G-INS-i**: método apropiado para secuencias de longitudes similares y recomendado para cuando se pretenden alinear menos de 200 secuencias.
- ▶ **E-INS-i**: método recomendado para secuencias que contienen largas regiones difilmente alineables (muchos gaps). De nuevo, recomendado para cuando se quiere alinear menos de 200 secuencias.

Mafft - Algoritmos orientados a eficiencia

- ▶ **FFT-NS-1:** método muy rápido recomendado para menos de 2000 secuencias.
- ▶ **FFT-NS-2:** método progresivo algo mas lento que el anterior.

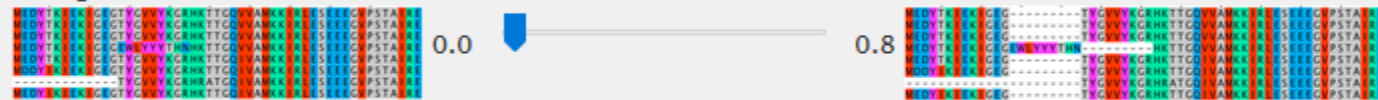
Mafft - Otras opciones (1)

- ▶ Alinear segmentos no relacionados: Permite realizar alineamientos globales de secuencias similares ignorando secuencias con grandes diferencias. Contempla subsecuencias posiblemente contaminadas que provocarían grandes secuencias de gaps en el alineamiento.
 - ▶ Esta opción solo esta disponible con el algoritmo **G-INS-i** y no esta recomendada para cuando se intentan alinear mas de 100 secuencias.

Align unrelated segments, too? in Alpha Testing (2014/Mar)

If the input data is expected to be globally conserved but locally contaminated by unrelated segments, try 'Unalignlevel>0' and possibly 'Leave gappy regions'.

Unalignlevel:



↑ Default

This feature is available only when G-INS-1 or G-INS-i is selected in the **Strategy** section above.

- ☒ Try to align gappy regions anyway
- ☐ Leave gappy regions (Not recommended for >~1,000 sequences)

Mafft - Otras opciones (2)

- ▶ **Scoring:** se pueden modificar los parámetros del *scoring*, como la penalización de añadir gaps (1.53 por defecto) o la matriz de puntuación para secuencias de aminoácidos y nucleótidos.
- ▶ **Árbol guía:** se puede cambiar el algoritmo para generar el árbol guía.

Parameters:

Scoring matrix for amino acid sequences: BLOSUM62 ▾

Scoring matrix for nucleotide sequences: 200PAM / $\kappa=2$ ▾

↑ Switch it to '1PAM / $\kappa=2$ ' when aligning closely related DNA sequences.

Gap opening penalty: 1.53 (1.0 - 5.0)

Offset value: 0.0 (0.0 - 1.0)

Score of N in nucleotide data: [Example](#)

↓ Long stretches of **Ns** tend to be gapped (excluded from the alignment).

☒ (nzero) **N** has no effect on the alignment score.

☐ (nwildcard) **N** is treated like a wildcard. **Experimental option** (2016/Apr/26)

↑ Try this if **Ns** should be aligned with usual letters.

Guide tree:

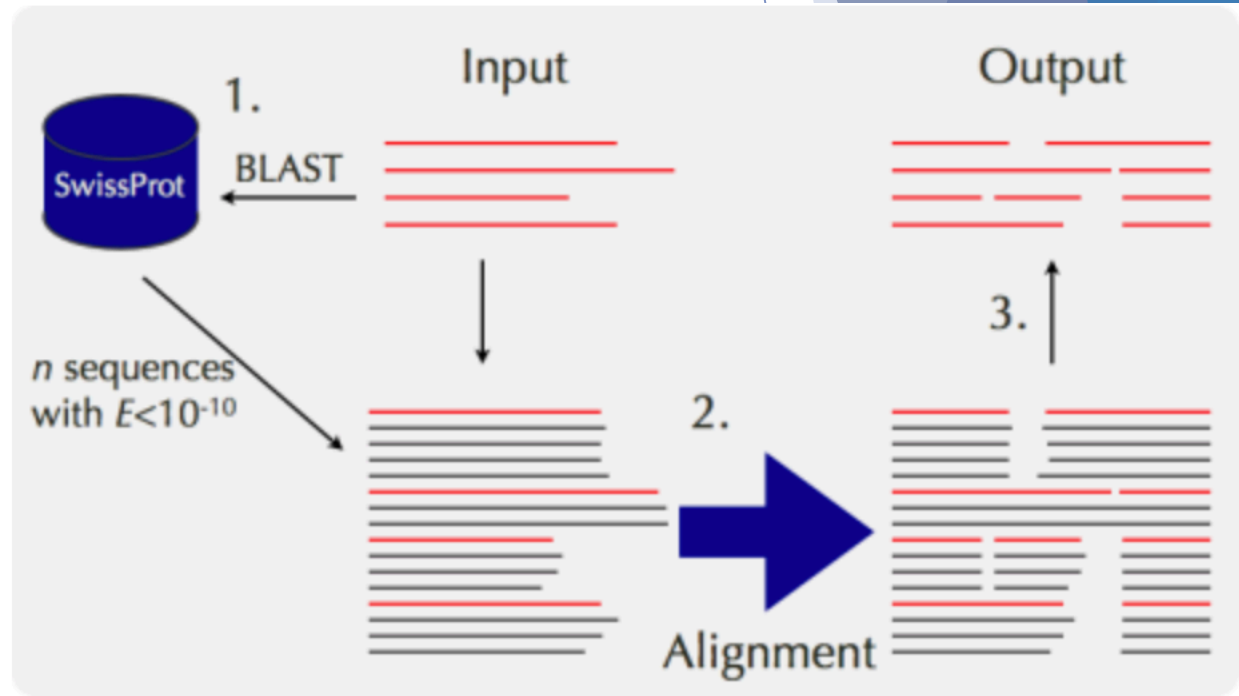
☒ Default ☐ UPGMA

☒ Output guide tree

To display the tree, follow the "Refine dataset" link in the result page.

Mafft - Otras opciones (3)

- ▶ Se puede mejorar la precisión de un alineamiento utilizando las secuencias homologas a las de entrada. Mafft sigue los siguientes pasos:
 - ▶ -Busca un numero especificado por el usuario de secuencias homologas con una distancia máxima introducida por el usuario
 - ▶ -Alinea las secuencias de entrada y sus homologas usando el algoritmo L-INS-i
 - ▶ -Elimina las secuencias homologas.



Mafft - Limitaciones

- ▶ En general, es poco escalable para largas secuencias de DNA. Incluso los algoritmos mas eficientes (por ejemplo FFS-NS-2) intentan usar una rigurosa programación dinámica que toma mucho tiempo de CPU. En esos casos, es mas recomendable usar FAST y BLAST.
- ▶ Se asume que el orden de los bloques alineables es conservado para todas las secuencias de entrada.

