



MAFFT Y WEBLOGO

Carlos Marañés - 717788
Dario Ferrer - 721132
Joaquín Gonzalez - 715821



¿QUÉ ES MAFFT?

(for Multiple Alignment using Fast Fourier Transform)

Programa usado para crear alineamientos de secuencias de aminoácidos o nucleótidos.

Primera versión: Algoritmo de alineamiento progresivo, donde las secuencias son procesadas utilizando la transformada rápida de Fourier (FFT).

Actualmente implementa muchos más algoritmos.

USO

- Windows, Mac OS y Linux.
- Se disponen de los fuentes para poder ser compilados
- VERSIÓN WEB -> descriptiva y de fácil uso. Se usará esta en la práctica.

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](#)

 Follow

Contact email address, kazutaka.katoh@a
then please re-send the email to katoh@i

About

MAFFT is a multiple sequence alignment p
of <~200 sequences), FFT-NS-2 (fast; for

Download and Installation

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

The latest version is 7.409 (2018/Aug

DESCARGA DEL RESULTADO

Computing

Job id: .19030620231657XjANCd1BsLZQPUMHiFT5wlsfnormal

Data size: 17 sequences $\times$ (642-266) aa

Command:

```
mafft --thread 10 --threadtb 5 --threadit 0 --reorder --auto input > output
```

Load level of this queue: 🚦

Load level of the cluster: 🚦

Calculation status will be checked 0.6 seconds later.

[Check now](#)

[Progress \(in a new window\)](#)

Cancel

Notify when finished

Set or change email address

DESCARGA DEL RESULTADO

[Clustal format](#) | [Fasta format](#) | [MAFFT result](#) | [View](#) | [Tree](#) | [Refine dataset](#) | [Return to home](#)

View

Reformat to GCG, PHYLIP, MSF, NEXUS, uppercase/lowercase, *etc.* with Readseq

GUIDANCE2 computes the residue-wise confidence scores and extracts well-aligned residues.

Refine dataset

Phylogenetic tree

MAFFT-L-INS-i Result

CLUSTAL format alignment by MAFFT (v7.423)

```
1LYLA      F-----NDELNRNRE-KLAALRQQGVAFPNDFRRDHTSDQLHEEFDAK-----D
c|p2       MP-----
1B8AA      M-----YRTHYS-----SEITEELNG-----
1AS2B      E-----DTAKDNYG-KLPLIQSRDSDRTG--QKRVKFVDLDEAKDSD-----
c|p1       M-----RTEYCGQL-----
1ADJA      -----
j|p2       M-----
j|p2       L-----RDNMK-----
c|p1       MPVITLPDGSQRHYD-----HAVSPMDVALDIG-----
c|p1       M-----RT-----
j|p2       L-----
1ATIA      -----
j|p2       M-----
1SESA      M--VD-LKRLRQEPE-----
```

MAFFT - ALGORITMOS

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](#) [Updated](#) (2015/Jun)
- ☐ 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 highly divergent ncRNAs with <200 sequences $\times$ <1,000 nucleotides; the number of iterative cycles is restricted to two, 2016/May) [Help](#)

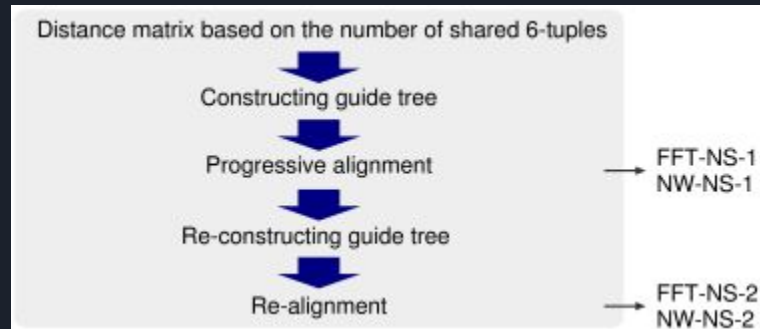


MAFFT - ALGORITMOS

- Compromiso entre velocidad y precisión.
- (1) Métodos progresivos, (2) iterativo con refinamiento e (3) iterativo con refinamiento y consistencia de puntuaciones.
- Velocidad: $(1) > (2) > (3)$
- Precisión $(1) < (2) < (3)$

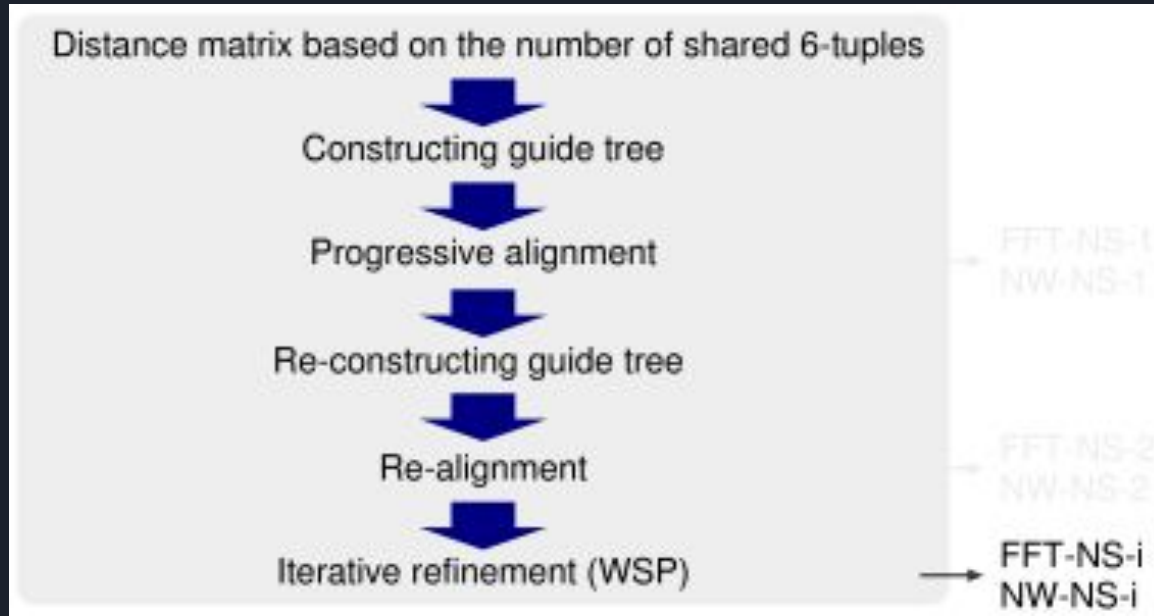
MAFFT - ALGORITMOS PROGRESIVOS

- FFT-NS-1: Opción más simple y uno de los métodos más rápidos actualmente disponibles.
- FFT-NS-2: Más lento que el anterior pero más preciso, ya que la matriz de distancias utilizada se vuelve a procesar.
- G-INS-1: Más lento aún pero con un árbol guía más preciso.

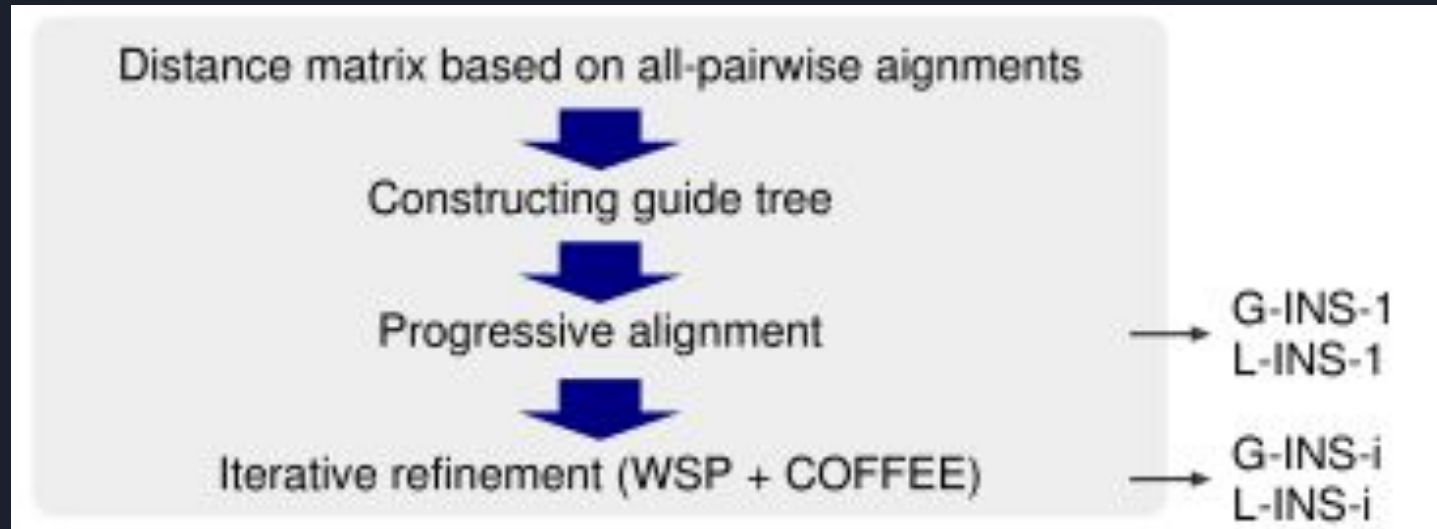


MAFFT - ALGORITMOS ITERATIVOS CON REFINAMIENTO

- FFT-NS-i: Se aplica una capa de iteración sobre el algoritmo progresivo FFT-NS-2 para ganar precisión.



MAFFT - ALGORITMOS ITERATIVOS CON REFINAMIENTO Y PUNTUACIONES CONSISTENTES



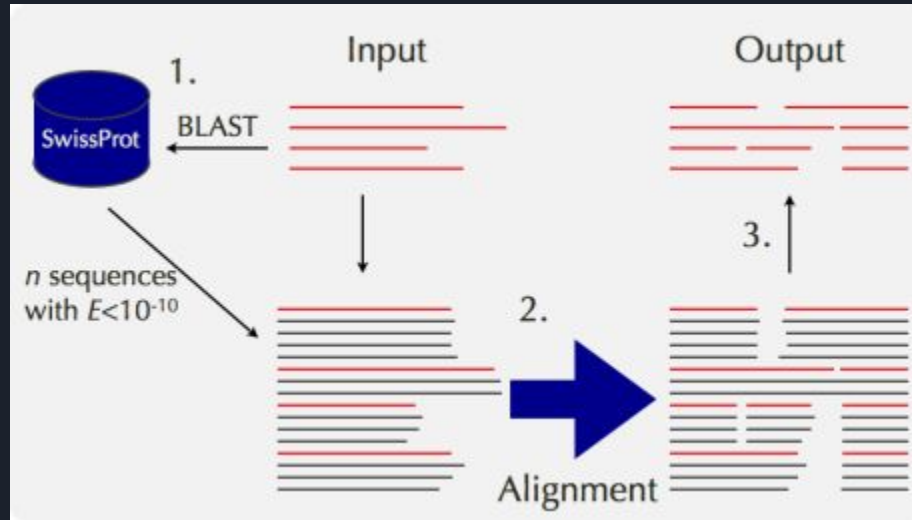


MAFFT - ALGORITMOS ITERATIVOS CON REFINAMIENTO Y PUNTUACIONES CONSISTENTES

- Todos estos algoritmos son muy lentos y no se recomienda utilizarlo con más de 200 secuencias.
- E-INS-i: Recomendado para menos de 200 secuencias. Se recomienda cuando la naturaleza de las secuencias a alinear no es clara.
- L-INS-i: Se asume que las secuencias de entrada tengan únicamente un dominio alineable (procedencia clara).
- G-INS-i: Asume que una región puede ser alineada e intenta alinearlas globalmente utilizando el algoritmo de Needleman-Wunsch.
- Q-INS-i: Extremadamente lento. Utiliza una función objetivo de consistencia incorporando información estructural.

OPTIMIZACIONES

- Para únicamente secuencias de proteínas se puede mejorar la precisión utilizando secuencias homólogas. Para ello, se buscan en el servidor SwissProt, se recogen 50 por defecto y se utilizan para el alineamiento y finalmente se eliminan de la salida.



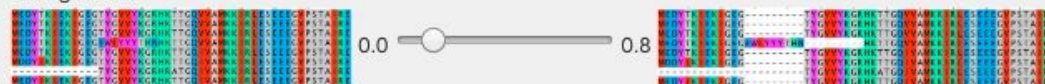
MAFFT- OPCIONES DE ANÁLISIS

Mafft permite realizar el alineamiento de secuencias ignorando aquellos fragmentos con muchas diferencias. En caso de que las secuencias estén contaminadas con segmentos no relacionados, hay que aumentar el valor del nivel de desalineación y marcar la opción que no intenta alinear los huecos.

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- PARÁMETROS DE SCORING

MAFFT permite cambiar los parámetros de scoring. Se puede cambiar el modelo de la matriz de scoring para secuencias de aminoácidos y para secuencias de nucleótidos. También permite cambiar el valor de penalización de los huecos.

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.

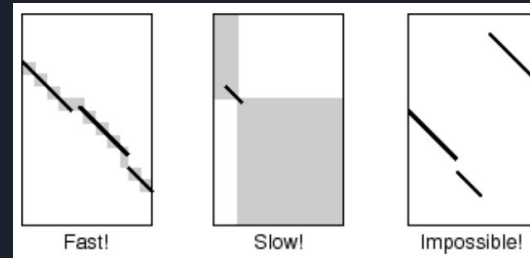


MAFFT - MÉRITOS

- **PRECISIÓN** → L-INS-i es uno de los métodos de alineamiento de secuencia más preciso en la actualidad, recomendado para conjuntos de entre 10 a 100 secuencias.
- **ESCALABILIDAD** → FFT-NS-2 y otros métodos progresivos disponibles en MAFFT permiten alinear un gran número de secuencias de gran tamaño.
- **SISTEMA DE PUNTUACIÓN** → Está diseñado de tal manera que permite la existencia de huecos de gran tamaño. Es por esto que se puede utilizar para alineamientos de secuencias de ARN con gaps de tamaño variable.

MAFFT - LIMITACIONES

- **PRECISIÓN** → No cuenta con extensión de librería. Ésta permite hacer un preprocesado de todas las secuencias a analizar, generando una librería con información acerca de posibles alineamientos entre pares, la cual es posteriormente utilizada para realizar el alineamiento global.
- **ESCALABILIDAD** → Si las secuencias son muy largas y tienen poca relación entre sí, es mejor utilizar otras herramientas como FASTA o BLAST, ya que los algoritmos utilizados en MAFFT tratan de conseguir un alineamiento muy preciso, requiriendo un gran tiempo de CPU.
- Se asume que el orden de los bloques alineables es conservado para todas las secuencias de entrada.



¿QUÉ ES WebLogo?

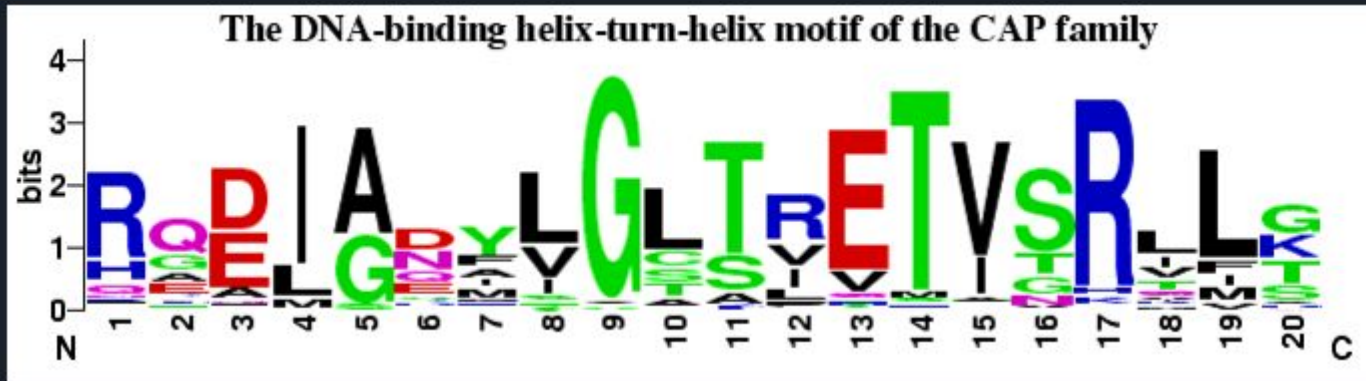
Es una aplicación web diseñada para realizar la generación de secuencias de logos lo más simple y óptima posible.

Creada por Tom Schneider y Mike Stephens



¿QUÉ ES una secuencia de logos?

Es una representación gráfica de los patrones dentro de una secuencia múltiple alineada.





¿POR QUÉ una secuencia de logos?

Provee de una descripción más rica y precisa de similitudes de secuencia que secuencias consensuadas.

Revela de forma más rápida características significativas del alineado.

A CAP-DNA Complex

Helix-Turn-Helix

B CAP recognition site DNA Logo

bits

5' -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 3'

C CAP Helix-Turn-Helix Logo

bits

169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191

Helix Turn Helix

Sidechain-Base Interactions

A CAP-DNA Complex

Helix-Turn-Helix

B CAP recognition site DNA Logo

bits

5' -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 3'

C CAP Helix-Turn-Helix Logo

bits

169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191

Helix Turn Helix

Sidechain-Base Interactions

A CAP-DNA Complex

Helix-Turn-Helix

B CAP recognition site DNA Logo

bits

5' -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 3'

C CAP Helix-Turn-Helix Logo

bits

169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191

Helix Turn Helix

Sidechain-Base Interactions



¿QUÉ ES un Logo?

Formado por pilas de letras.

Cada pila está colocada en una posición de la secuencia.

La altura de cada pila indica la conservación de la secuencia en esa posición.

La altura de las letras en cada pila indica la frecuencia de dicho aminoácido o ácido nucleico encontrado en la posición.

¿CÓMO FUNCIONA?

Multiple Sequence Alignment

Upload Sequence Data:

Seleccionar archivo Ningún archivo seleccionado

Image Format & Size

Image Format:

PNG (bitmap)

Logo Size per Line:

18

X 5

cm

Create Logo

Reset

Advanced Logo Options

Sequence Type:

☐ amino acid ☐ DNA / RNA ☒ Automatic Detection

First Position Number:

1

Logo Range:

-

Small Sample Correction:

☒

Frequency Plot:

☐

Multiline Logo (Symbols per Line):

☐ (32)

Advanced Image Options

Bitmap Resolution:

96 pixels/inch (dpi)

Antialias Bitmaps:

☒

Title:

Y-Axis Height:

(bits)

Show Y-Axis:

☒

Y-Axis Label:

bits

Show X-Axis:

☒

X-Axis Label:

Show Error Bars:

☐

Label Sequence Ends:

☒

Boxed / Boxed Shrink Factor:

☐ / 0.5

Outline Symbols:

☐

Show fine print:

☒

Y-Axis Tic Spacing:

1 (bits)

Colors

Color Scheme:

☒ Default ☐ Black & White ☐ Custom (See Below.)

Symbols			Color			RGB		
⌕ Symbols	⌕ Color	⌕ RGB	⌕ Symbols	⌕ Color	⌕ RGB	⌕ Symbols	⌕ Color	⌕ RGB
KRH	green			purple				
DE	blue			orange				
AVLIPWFM	red			black				
	black			black				
Other								

Create Logo

Reset