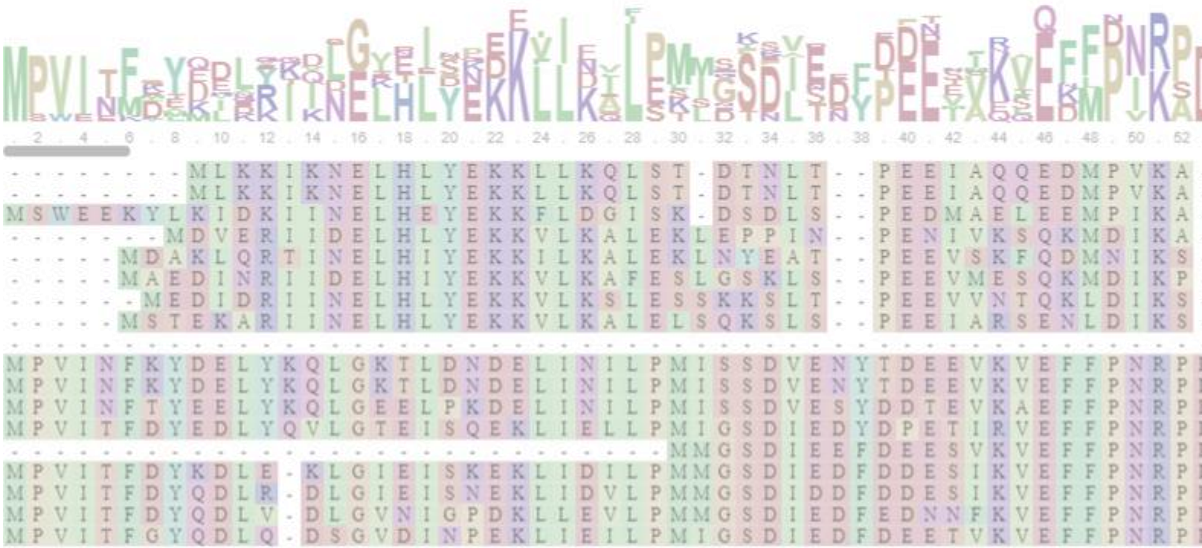
Alineamiento múltiple de secuencias

Multiple sequence alignment (MSA)



BCH441 - Bioinformatics
Boris Steipe

http://steipe.biochemistry.utoronto.ca/abc/index.php/Bioinformatics_Main_Page

Alineamiento de Secuencias

¿Qué método estadístico será usado para evaluar la significancia de un alineamiento?

P-value:

Z-value o ***E*-value**
Expect value

→ BLAST

E-value de un score ***s*** puede ser interpretado como el número de falsos positivos que poseen este puntaje o uno más alto dentro de una base de datos del mismo tamaño.

Mientras más pequeño es el *E*-value (mayor el exponente negativo), más significancia posee el alineamiento

	Description	Max Score	Total Score	Query Cover	**E value**	Per. Ident	Accession
☑	RecName: Full=Organic hydroperoxide resistance transcriptional regulator [Bacillus subtilis subsp. subtilis str. 168]	132	132	89%	5e-39	46.32%	O34777.1
☑	RecName: Full=HTH-type transcriptional regulator SarZ; AltName: Full=Staphylococcal accessory regulator Z [Staphylococcus haemolyticus JCS	96.3	96.3	76%	6e-25	36.21%	Q4L8Q3.1
☐	RecName: Full=HTH-type transcriptional regulator SarZ; AltName: Full=Staphylococcal accessory regulator Z [Staphylococcus saprophyticus sub	96.3	96.3	69%	7e-25	42.86%	Q49ZX3.1

Alineamiento de Secuencias

E-value es un buen parámetro, pero esto no elimina la necesidad de aplicar **sentido común biológico**!!

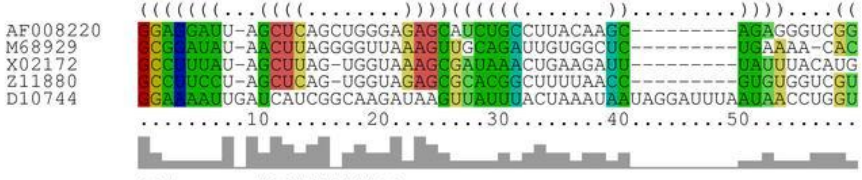
Residuos importantes (e.g. cisteínas) o dominios conservados, el largo de la secuencia afecta en valor de *E*-value (mientras más corta la secuencia, es más alto el valor)

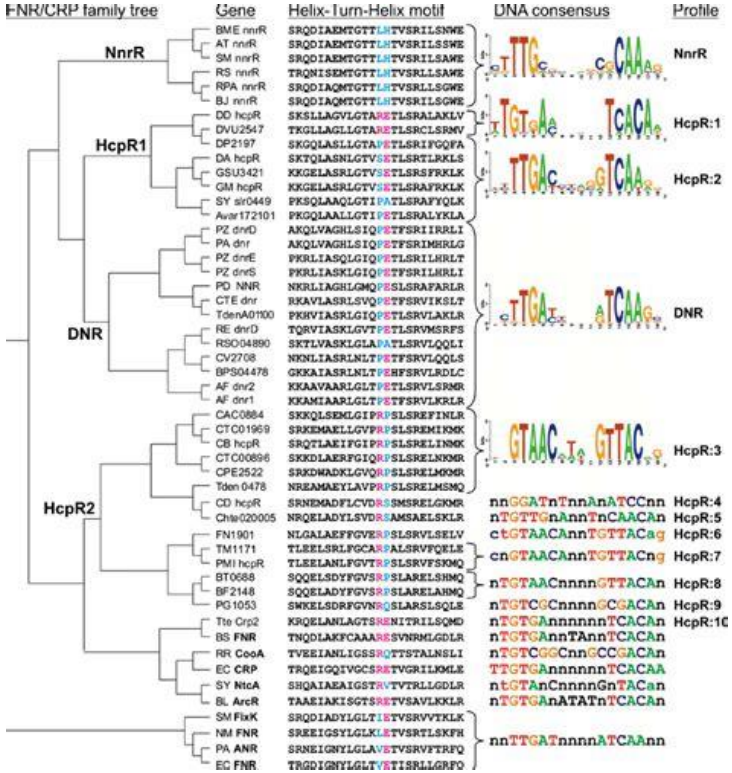
Alineamiento Múltiple de Secuencias (MSA)

Multiple sequence alignments (MSA)

¿Para que sirven los MSA?

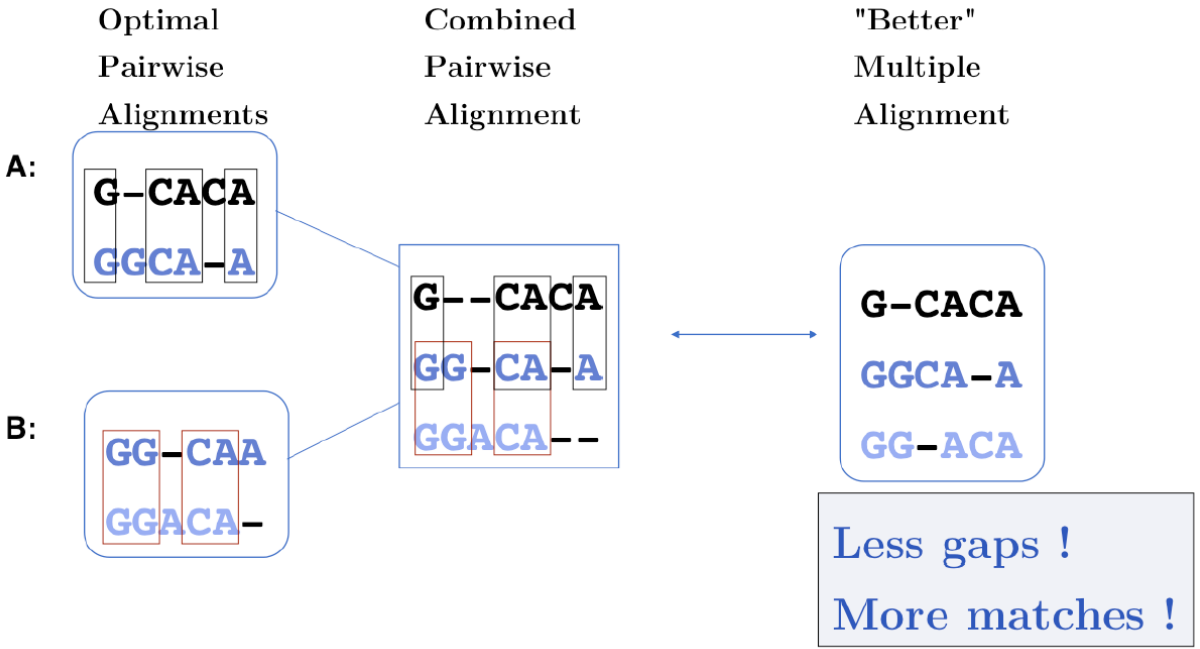
1. Relaciones filogenéticas
2. Patrones conservados (e.g. regiones variables; residuos conservados; límites de dominios)
3. Modelamiento estructural
4. Predicción de estructura secundaria
5. Reconstrucción filogenética
6. Búsquedas más sensibles de homología (e.g. HMM search)





Alineamiento Múltiple de Secuencias (MSA)

Óptimo alineamiento de pares ≠ Mejor alineamiento multiple



Algoritmos de MSA se basan en procesos heurísticos!!

Pairwise alignment: $O(L^2)$

n-wise alignment: $O(L^n)$

Aún es tema de investigación en bioinformática!

Alineamiento Múltiple de Secuencias (MSA)

Algoritmos de MSA : Maximizar el puntaje total (Score = **s**)

- Muchos métodos propuestos
- Más común: suma de **s** de todos los alineamientos de 2 secuencias (*pairwise alignments*)

¿Cómo se define el objetivo de un MSA biológicamente significativo?

Alineamiento Múltiple de Secuencias (MSA)

Algunos objetivos biológicos a considerar en un MSA:

- Minimizar el número de indels (gaps)
- Maximizar la similud de secuencias
- Buscar motivos o patrones conservados
- Recapitular filogenia
- Concentrarse en regions locales de alineamiento no indels

Cada uno de estos objetivos sugiere una estrategia diferente de alineamiento!!

Alineamiento Múltiple de Secuencias (MSA)

Objective	Algorithm	Type of alignment algorithm
Maximize similarity, Minimize gaps	Bounded optimal solution search	**Exact**
Align according to phylogeny	Align most similar first, then merge together	**Progressive**
Retain conserved regions	Conserved regions guide alignment	**Consistency based**
Maximize similarity to model	Create a model, align each sequence to that	**Probabilistic**
"Learn" about important regions and extend the alignment from secure seeds	Improve alignment from draft alignments	**Iterated**

Programas que incluyen pasos correspondientes a más de un tipo de algoritmo dan “mejores resultados”

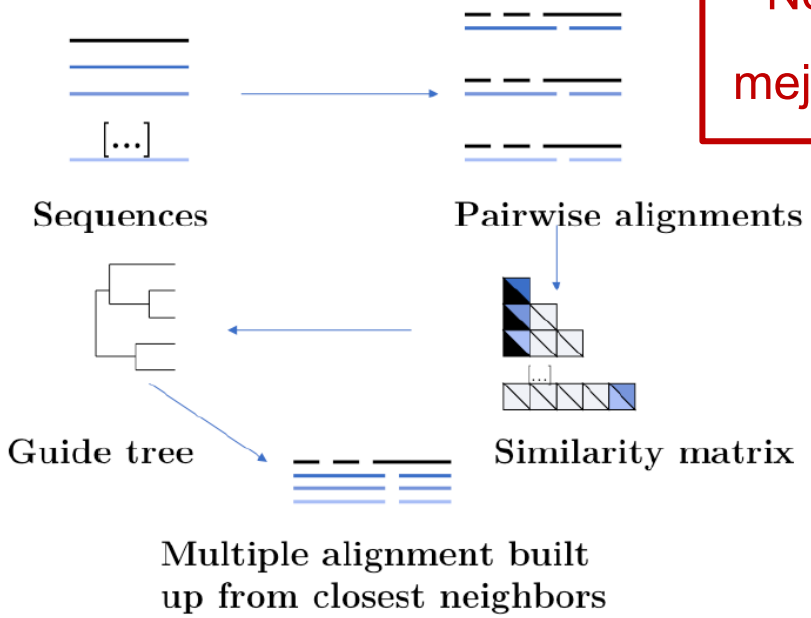
Alineamiento Múltiple de Secuencias (MSA)

Alineamiento progresivo:

Clustal W, X:
Considered by many to be *The Standard* method.

1. Build pairwise similarity matrix
2. **Build guide tree.**
3. join neighbors into profiles according to guide tree.
4. Align according to tree.

Key limitation: early errors persist! Best performance is for globally alignable, gap-poor sequence sets. Performance progressively worse for multidomain proteins and distant similarities.



DISCLAIMER:

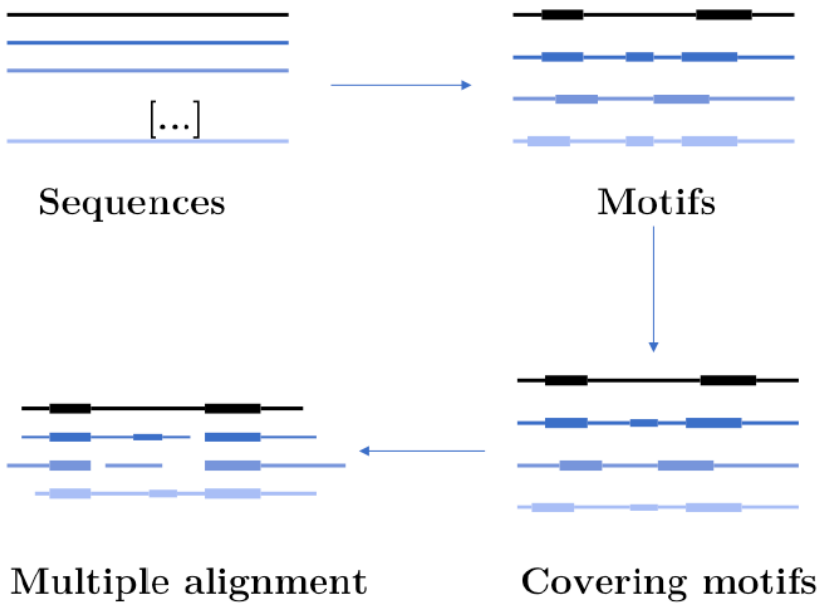
No lo usen, hay mejores algoritmos

Alineamiento Múltiple de Secuencias (MSA)

Alineamiento consistente:

MUSCA:
(I) TEIRESIAS motif discovery.
(II) Use set covering algorithm to select motifs that are common to sequence set.

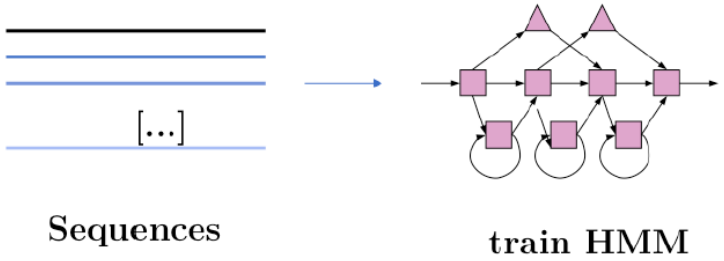
MEME:
Postulate motif, compare residue composition of motif with background, choose motifs to maximize composition difference, output.



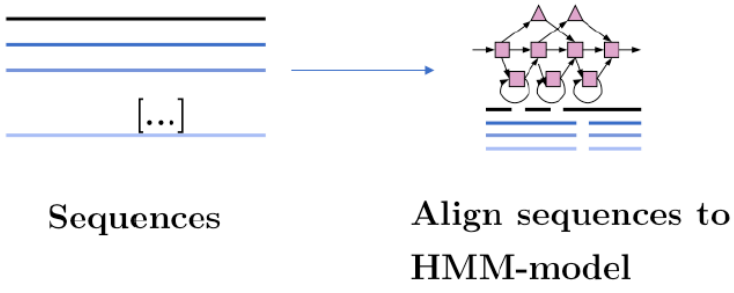
Alineamiento Múltiple de Secuencias (MSA)

Alineamiento probabilístico:

SAM:
Build HMM from input sequences.
Align sequences to HMM.



HMMER usa este tipo de algoritmo



Alineamiento Múltiple de Secuencias (MSA)

Alineamiento basado en perfiles (iterativos):

The MSA derived from aligning sequences to a profile in a **PSI-BLAST** search is also a model based alignment.

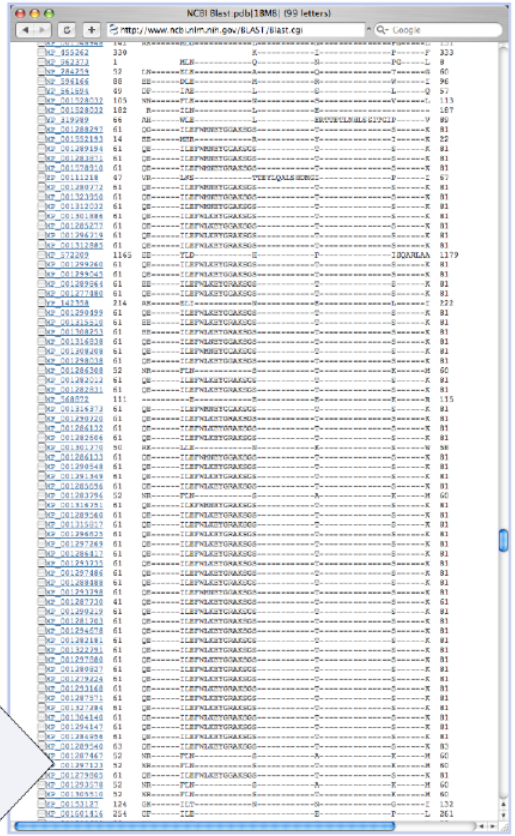
PSI-BLAST

1. Begin with BLAST search
2. Identify significant hits
3. Align to query
4. Compile into position specific scoring matrix (PSSM or "sequence profile")
5. **Repeat search with profile**
6. Add new aligned hits to PSSM
7. Iterate until no new sequences can be added

Results can be displayed as an MSA.

Choose formatting option:

"Flat query-anchored with letters for identities"



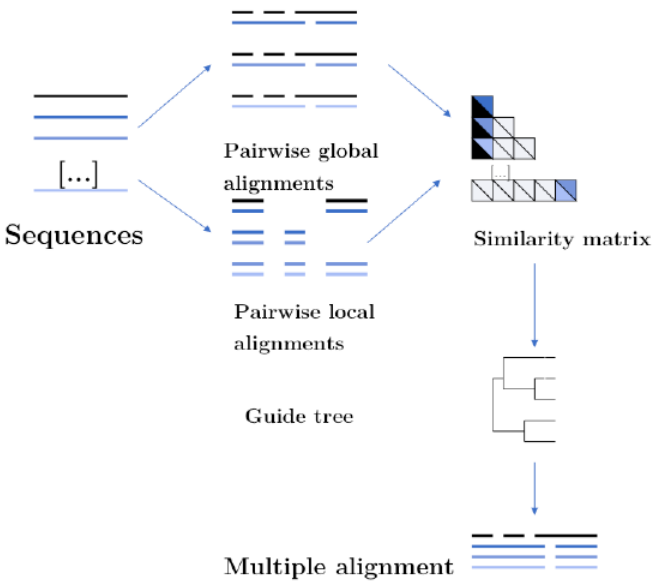
Alineamiento Múltiple de Secuencias (MSA)

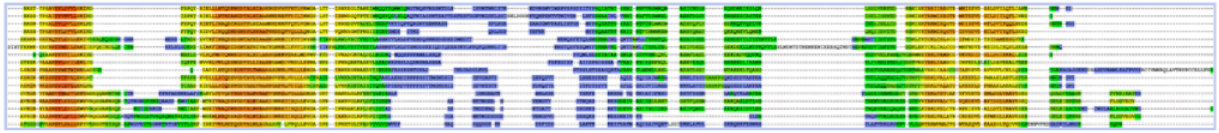
Alineamiento basado en perfiles (iterativos):

TCoffee - a hybrid algorithm significantly improves performance:

1. Compute **global** pairwise similarity matrix.
2. Compute top 10 non-intersecting **local** alignments.
3. Combine by looking at triplets of sequences.
4. Build guide tree.
5. Align according to tree.

Very good results.





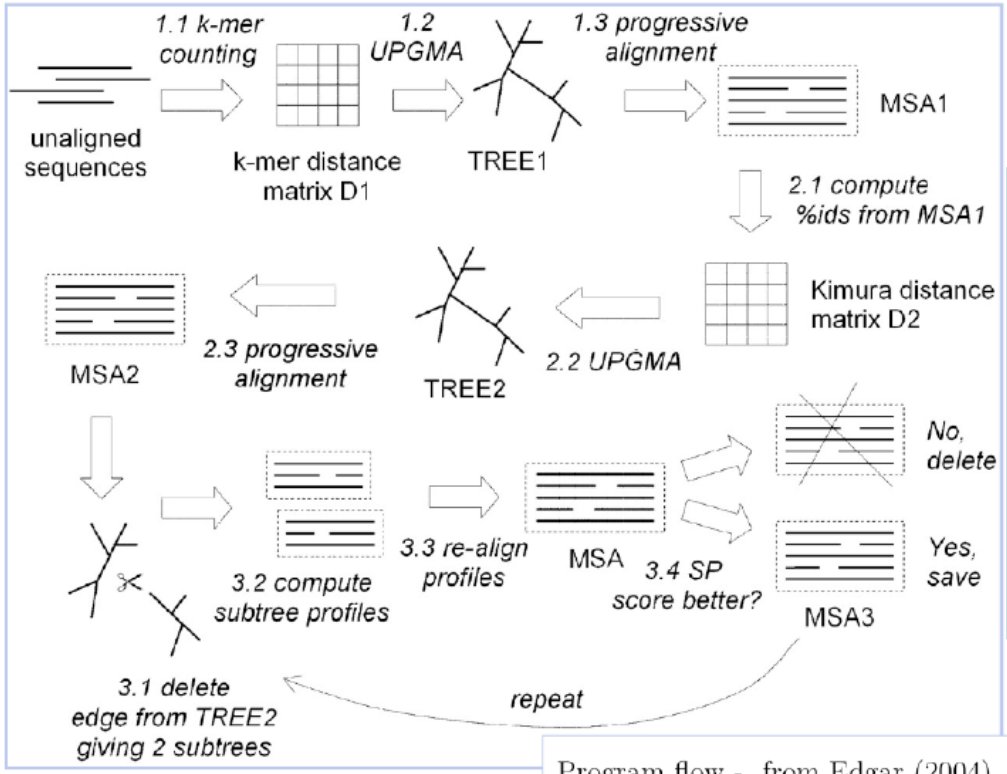
Alineamiento Múltiple de Secuencias (MSA)

MUSCLE (MSA)

Progressive draft alignment

Iterated improvement

Refinement, by aligning to groups of profiles

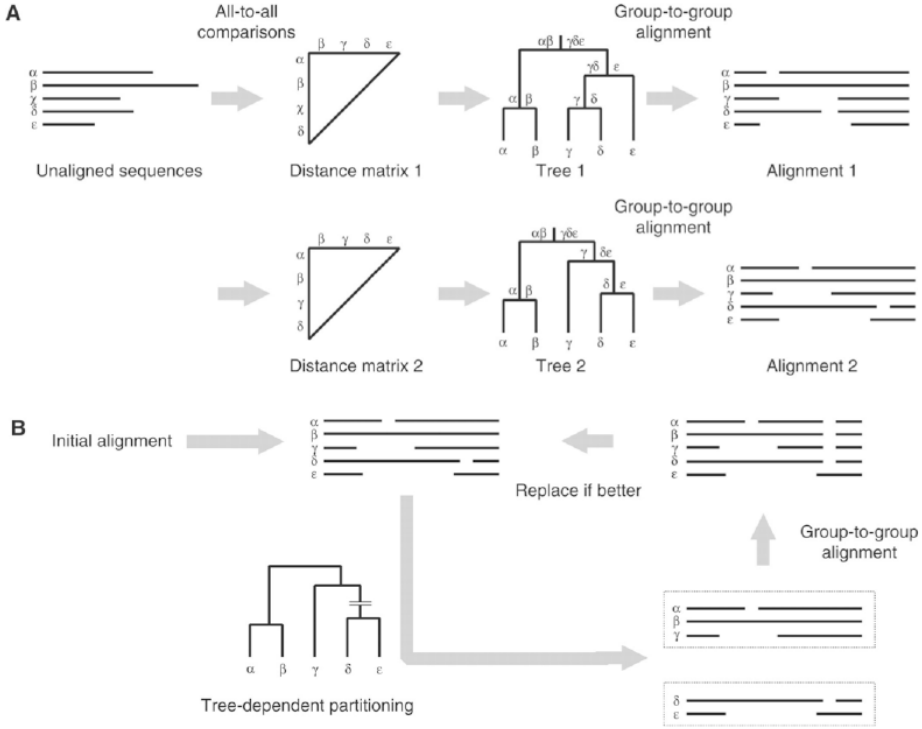


Program flow - from Edgar (2004)

Uno de los más utilizados para alineamientos con un propósito general

Alineamiento Múltiple de Secuencias (MSA)

MAFFT (MSA): progresivo-refinamiento iterativo (WSP o basado en consistencia)

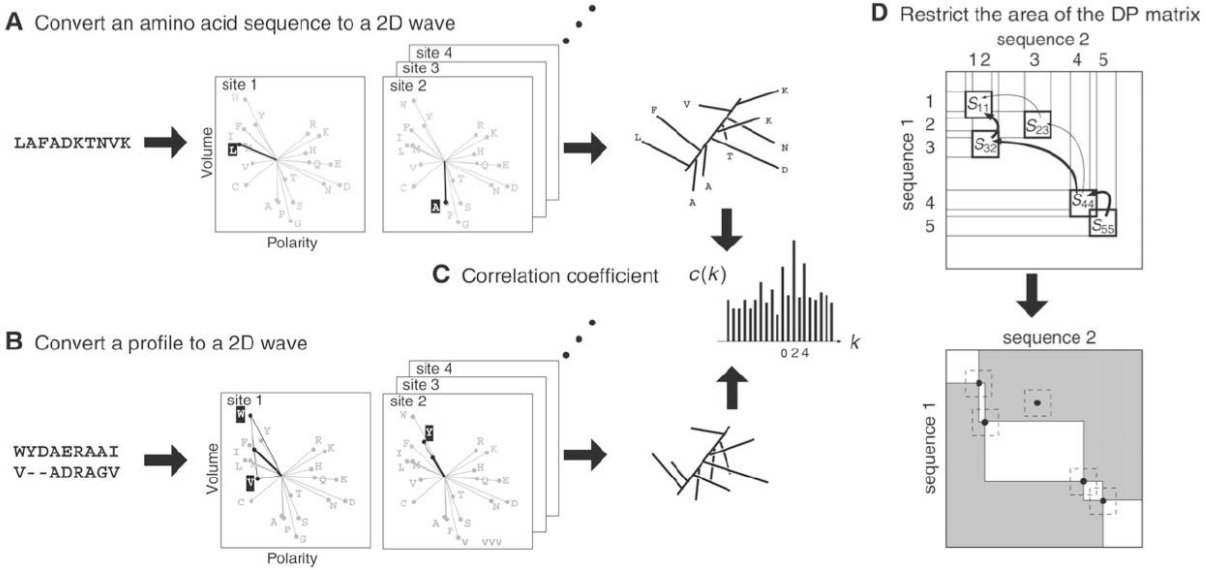


https://mafft.cbrc.jp/alignment/software/tips0.html

Kazutaka Katoh & Hiroyuki Toh (2008) **Recent developments in the MAFFT multiple sequence alignment program** *Briefings in Bioinformatics* 9(4): 286-298.

Alineamiento Múltiple de Secuencias (MSA)

MAFFT (MSA): progresivo-refinamiento iterativo (WSP o basado en consistencia)



Se basa en transformaciones de rápidas de Fourier (FFT) para realizar el agrupamiento de secuencias.

https://mafft.cbrc.jp/alignment/software/tips0.html

Alineamiento Múltiple de Secuencias (MSA)

Existen muchos programas para realizar alineamientos multiples de secuencias

Alineamientos pares de secuencias es problema resuelto

Alineamientos multiples de secuencias NO

Alineamiento Múltiple de Secuencias (MSA)

¿Qué programas debo utilizar?

Depende de lo que quieran hacer con ese alineamiento

- Conservación de patrones: casi cualquier algoritmo lo hara bien
- Definir límites de dominios: indels son muy importantes

Usen más de un algoritmo y comparen resultados

Alineamiento Múltiple de Secuencias (MSA)

Consideraciones a tomar en cuenta!

- Usar solo secuencias homólogas! (considerar si incluir parálogos o homólogos)
- Si el MSA tiene muchos indels revisar secuencias
- Definir que quieren comparar: Dominios específicos de la proteína.
- Recortar extremos que no den información, en general N-terminal y C-terminal no ayudan a identificar patrones conservados y aumentan la cantidad de indels