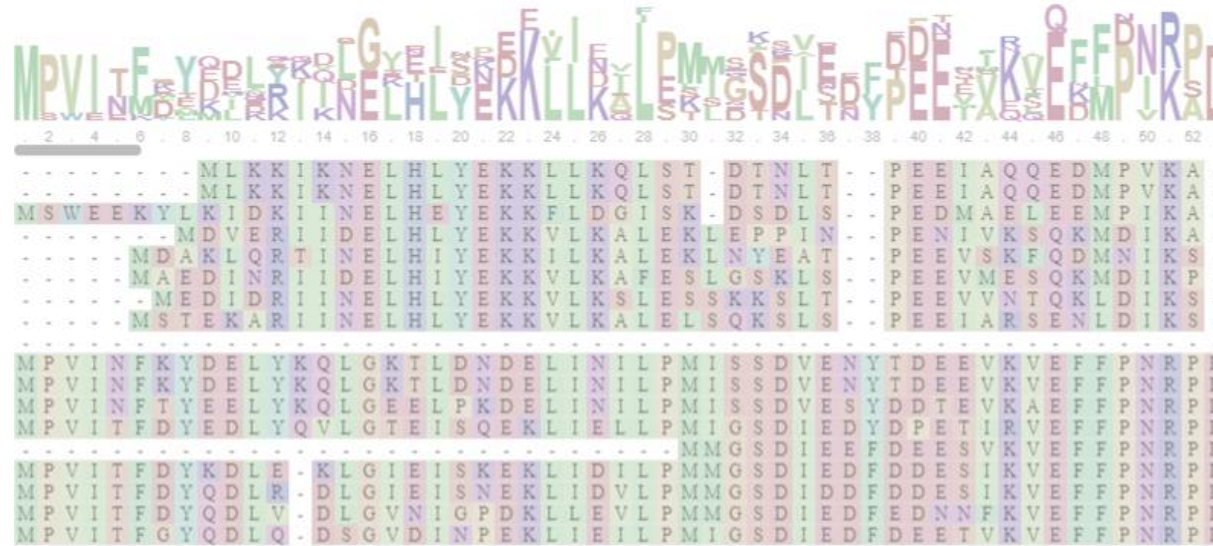


Alineamiento múltiple de secuencias

Multiple sequence alignment (MSA)



BCH441 - Bioinformatics
Boris Steipe

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
<input checked="" type="checkbox"/>	RecName: Full=Organic hydroperoxide resistance transcriptional regulator [Bacillus subtilis subsp. subtilis str. 168]	132	132	89%	5e-39	46.32%	Q34777.1
<input checked="" type="checkbox"/>	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
<input type="checkbox"/>	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!!**

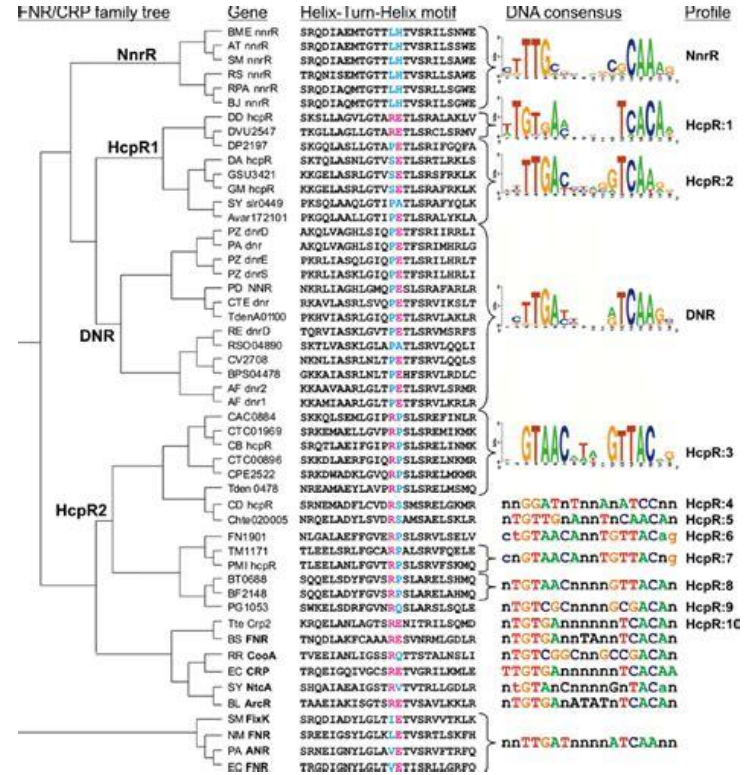
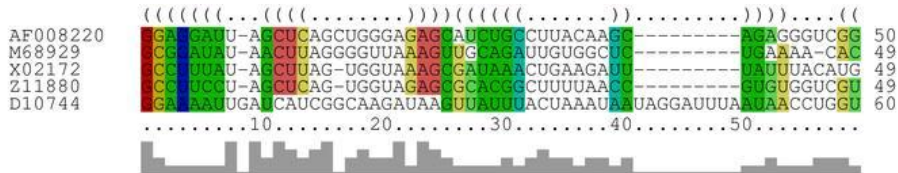
Residuos importantes (e.g. cisteínas) o dominios conservados, el largo de la secuencia afecta en valor de *E*-value (mientras más corta la secuencia, es más alto el valor)

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 \neq Mejor alineamiento multiple

Optimal
Pairwise
Alignments

Combined
Pairwise
Alignment

"Better"
Multiple
Alignment

Algoritmos de MSA se
basan en procesos
heurísticos!!

Pairwise alignment: $O(L^2)$

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

Less gaps !
More matches !

Aún es tema de investigación en bioinformática!

A:

G-CACA
GGCA-A

B:

GG-CAA
GGACA-

G--CACA
GG-CA-A
GGACA--

G-CACA
GGCA-A
GG-ACA

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 similitud de secuencias
- Buscar motivos o patrones conservados
- Recapitular filogenia
- Concentrarse en regiones 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)

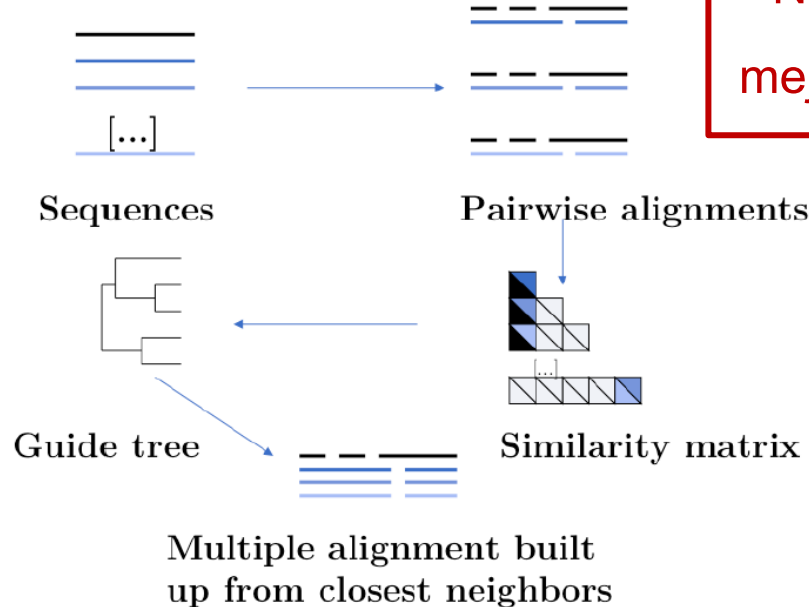
Alineamiento progresivo:

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:

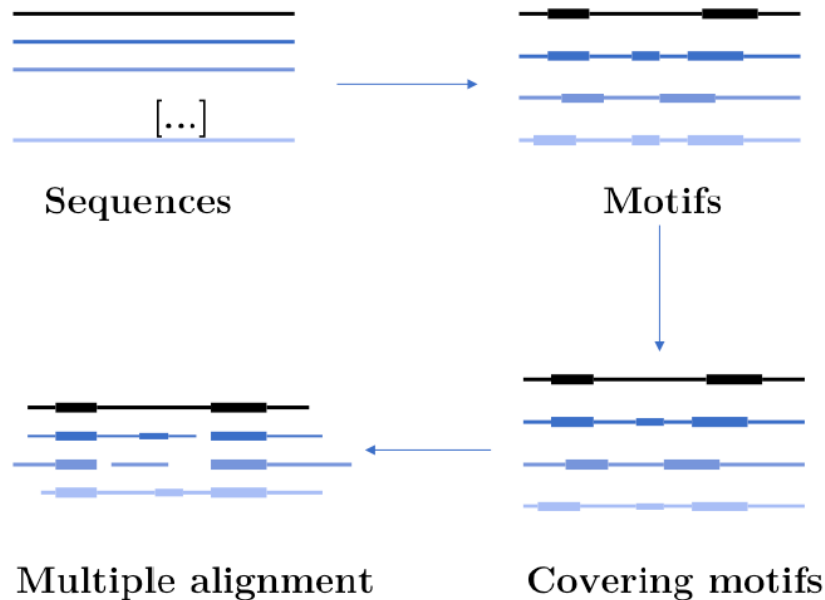
MUSCA:

(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.



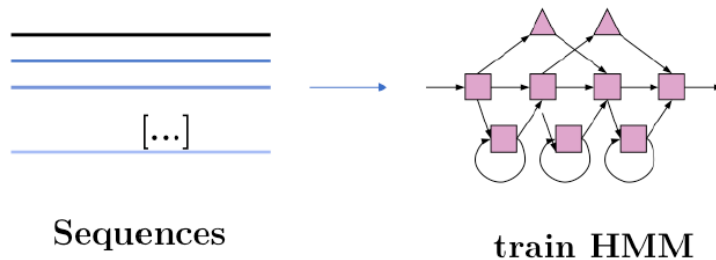
Alineamiento Múltiple de Secuencias (MSA)

Alineamiento probabilístico:

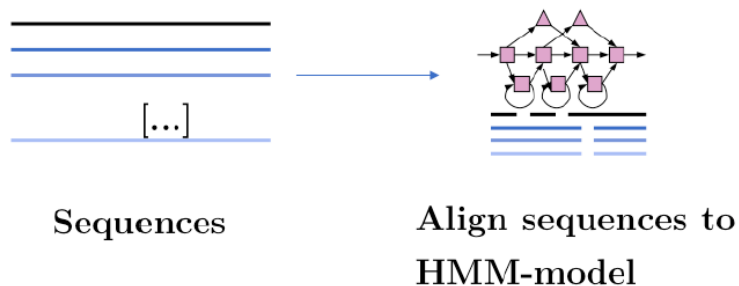
SAM:

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"

N3 (Basi) plot(12M) (99 letters)

<http://www.biorxiv.org/q/BASI/Basi.qp>

Q - Large

Mr_01559242	230	CG	-	-	-	P	-	232
Mr_01559243		CG	-	-	-	-	-	81
Mr_01559244	79	CG	-	-	-	-	-	80
Mr_01559245		CG	-	-	-	-	-	96
Mr_01559246		CG	-	-	-	-	-	81
Mr_01559247	49	CG	-	-	-	-	-	81
Mr_01559248	102	CG	-	-	-	-	-	113
Mr_01559249	105	CG	-	-	-	-	-	187
Mr_01559250		CG	-	-	-	-	-	81
Mr_01559251	88	CG	-	-	-	-	-	81
Mr_01559252	91	CG	-	-	-	-	-	81
Mr_01559253	14	CG	-	-	-	-	-	81
Mr_01559254	01	CG	-	-	-	-	-	81
Mr_01559255	01	CG	-	-	-	-	-	81
Mr_01559256	01	CG	-	-	-	-	-	81
Mr_01559257	01	CG	-	-	-	-	-	81
Mr_01559258	01	CG	-	-	-	-	-	81
Mr_01559259	01	CG	-	-	-	-	-	81
Mr_01559260	01	CG	-	-	-	-	-	81
Mr_01559261	01	CG	-	-	-	-	-	81
Mr_01559262	01	CG	-	-	-	-	-	81
Mr_01559263	01	CG	-	-	-	-	-	81
Mr_01559264	01	CG	-	-	-	-	-	81
Mr_01559265	01	CG	-	-	-	-	-	81
Mr_01559266	01	CG	-	-	-	-	-	81
Mr_01559267	01	CG	-	-	-	-	-	81
Mr_01559268	01	CG	-	-	-	-	-	81
Mr_01559269	01	CG	-	-	-	-	-	81
Mr_01559270	01	CG	-	-	-	-	-	81
Mr_01559271	01	CG	-	-	-	-	-	81
Mr_01559272	01	CG	-	-	-	-	-	81
Mr_01559273	01	CG	-	-	-	-	-	81
Mr_01559274	01	CG	-	-	-	-	-	81
Mr_01559275	01	CG	-	-	-	-	-	81
Mr_01559276	01	CG	-	-	-	-	-	81
Mr_01559277	01	CG	-	-	-	-	-	81
Mr_01559278	01	CG	-	-	-	-	-	81
Mr_01559279	01	CG	-	-	-	-	-	81
Mr_01559280	01	CG	-	-	-	-	-	81
Mr_01559281	01	CG	-	-	-	-	-	81
Mr_01559282	01	CG	-	-	-	-	-	81
Mr_01559283	01	CG	-	-	-	-	-	81
Mr_01559284	01	CG	-	-	-	-	-	81
Mr_01559285	01	CG	-	-	-	-	-	81
Mr_01559286	01	CG	-	-	-	-	-	81
Mr_01559287	01	CG	-	-	-	-	-	81
Mr_01559288	01	CG	-	-	-	-	-	81
Mr_01559289	01	CG	-	-	-	-	-	81
Mr_01559290	01	CG	-	-	-	-	-	81
Mr_01559291	01	CG	-	-	-	-	-	81
Mr_01559292	01	CG	-	-	-	-	-	81
Mr_01559293	01	CG	-	-	-	-	-	81
Mr_01559294	01	CG	-	-	-	-	-	81
Mr_01559295	01	CG	-	-	-	-	-	81
Mr_01559296	01	CG	-	-	-	-	-	81
Mr_01559297	01	CG	-	-	-	-	-	81
Mr_01559298	01	CG	-	-	-	-	-	81
Mr_01559299	01	CG	-	-	-	-	-	81
Mr_01559300	01	CG	-	-	-	-	-	81
Mr_01559301	01	CG	-	-	-	-	-	81
Mr_01559302	01	CG	-	-	-	-	-	81
Mr_01559303	01	CG	-	-	-	-	-	81
Mr_01559304	01	CG	-	-	-	-	-	81
Mr_01559305	01	CG	-	-	-	-	-	81
Mr_01559306	01	CG	-	-	-	-	-	81
Mr_01559307	01	CG	-	-	-	-	-	81
Mr_01559308	01	CG	-	-	-	-	-	81
Mr_01559309	01	CG	-	-	-	-	-	81
Mr_01559310	01	CG	-	-	-	-	-	81
Mr_01559311	01	CG	-	-	-	-	-	81
Mr_01559312	01	CG	-	-	-	-	-	81
Mr_01559313	01	CG	-	-	-	-	-	81
Mr_01559314	01	CG	-	-	-	-	-	81
Mr_01559315	01	CG	-	-	-	-	-	81
Mr_01559316	01	CG	-	-	-	-	-	81
Mr_01559317	01	CG	-	-	-	-	-	81
Mr_01559318	01	CG	-	-	-	-	-	81
Mr_01559319	01	CG	-	-	-	-	-	81
Mr_01559320	01	CG	-	-	-	-	-	81
Mr_01559321	01	CG	-	-	-	-	-	81
Mr_01559322	01	CG	-	-	-	-	-	81
Mr_01559323	01	CG	-	-	-	-	-	81
Mr_01559324	01	CG	-	-	-	-	-	81
Mr_01559325	01	CG	-	-	-	-	-	81
Mr_01559326	01	CG	-	-	-	-	-	81
Mr_01559327	01	CG	-	-	-	-	-	81
Mr_01559328	01	CG	-	-	-	-	-	81
Mr_01559329	01	CG	-	-	-	-	-	81
Mr_01559330	01	CG	-	-	-	-	-	81
Mr_01559331	01	CG	-	-	-	-	-	81
Mr_01559332	01	CG	-	-	-	-	-	81
Mr_01559333	01	CG	-	-	-	-	-	81
Mr_01559334	01	CG	-	-	-	-	-	81
Mr_01559335	01	CG	-	-	-	-	-	81
Mr_01559336	01	CG	-	-	-	-	-	81
Mr_01559337	01	CG	-	-	-	-	-	81
Mr_01559338	01	CG	-	-	-	-	-	81
Mr_01559339	01	CG	-	-	-	-	-	81
Mr_01559340	01	CG	-	-	-	-	-	81
Mr_01559341	01	CG	-	-	-	-	-	81
Mr_01559342	01	CG	-	-	-	-	-	81
Mr_01559343	01	CG	-	-	-	-	-	81
Mr_01559344	01	CG	-	-	-	-	-	81
Mr_01559345	01	CG	-	-	-	-	-	81
Mr_01559346	01	CG	-	-	-	-	-	81
Mr_01559347	01	CG	-	-	-	-	-	81
Mr_01559348	01	CG	-	-	-	-	-	81
Mr_01559349	01	CG	-	-	-	-	-	81
Mr_01559350	01	CG	-	-	-	-	-	81
Mr_01559351	01	CG	-	-	-	-	-	81
Mr_01559352	01	CG	-	-	-	-	-	81
Mr_01559353	01	CG	-	-	-	-	-	81
Mr_01559354	01	CG	-	-	-	-	-	81
Mr_01559355	01	CG	-	-	-	-	-	81
Mr_01559356	01	CG	-	-	-	-	-	81
Mr_01559357	01	CG	-	-	-	-	-	81
Mr_01559358	01	CG	-	-	-	-	-	81
Mr_01559359	01	CG	-	-	-	-	-	81
Mr_01559360	01	CG	-	-	-	-	-	81
Mr_01559361	01	CG	-	-	-	-	-	81
Mr_01559362	01	CG	-	-	-	-	-	81
Mr_01559363	01	CG	-	-	-	-	-	81
Mr_01559364	01	CG	-	-	-	-	-	81
Mr_01559365	01	CG	-	-	-	-	-	81
Mr_01559366	01	CG	-	-	-	-	-	81
Mr_01559367	01	CG	-	-	-	-	-	81
Mr_01559368	01	CG	-	-	-	-	-	81
Mr_01559369	01	CG	-	-	-	-	-	81
Mr_01559370	01	CG	-	-	-	-	-	81
Mr_01559371	01	CG	-	-	-	-	-	81
Mr_01559372	01	CG	-	-	-	-	-	81
Mr_01559373	01	CG	-	-	-	-	-	81
Mr_01559374	01	CG	-	-	-	-	-	81
Mr_01559375	01	CG	-	-	-	-	-	81
Mr_01559376	01	CG	-	-	-	-	-	81
Mr_01559377	01	CG	-	-	-	-	-	81
Mr_01559378	01	CG	-	-	-	-	-	81
Mr_01559379	01	CG	-	-	-	-	-	81
Mr_01559380	01	CG	-	-	-	-	-	81
Mr_01559381	01	CG	-	-	-	-	-	81
Mr_01559382	01	CG	-	-	-	-	-	81
Mr_01559383	01	CG	-	-	-	-	-	81
Mr_01559384	01	CG	-	-	-	-	-	81
Mr_01559385	01	CG	-	-	-	-	-	81
Mr_01559386	01	CG	-	-	-	-	-	81
Mr_01559387	01	CG	-	-	-	-	-	81
Mr_01559388	01	CG	-	-	-	-	-	81
Mr_01559389	01	CG	-	-	-	-	-	81
Mr_01559390	01	CG	-	-	-	-	-	81
Mr_01559391	01	CG	-	-	-	-	-	81
Mr_01559392	01	CG	-	-	-	-	-	81
Mr_01559393	01	CG	-	-	-	-	-	81
Mr_01559394	01	CG	-	-	-	-	-	81
Mr_01559395	01	CG	-	-	-	-	-	81
Mr_01559396	01	CG	-	-	-	-	-	81
Mr_01559397	01	CG	-	-	-	-	-	81
Mr_01559398	01	CG	-	-	-	-	-	81
Mr_01559399	01	CG	-	-	-	-	-	81
Mr_01559400	01	CG	-	-	-	-	-	81
Mr_01559401	01	CG	-	-	-	-	-	81
Mr_01559402	01	CG	-	-	-	-	-	81
Mr_01559403	01	CG	-	-	-	-	-	81
Mr_01559404	01	CG	-	-	-	-	-	81
Mr_01559405	01	CG	-	-	-	-	-	81
Mr_01559406	01	CG	-	-	-	-	-	81
Mr_01559407	01	CG	-	-	-	-	-	81
Mr_01559408	01	CG	-	-	-	-	-	81
Mr_01559409	01	CG	-	-	-	-	-	81
Mr_01559410	01	CG	-	-	-	-	-	81
Mr_01559411	01	CG	-	-	-	-	-	81
Mr_01559412	01	CG	-	-	-	-	-	81
Mr_01559413	01	CG	-	-	-	-	-	81
Mr_01559414	01	CG	-	-	-	-	-	81
Mr_01559415	01	CG	-	-	-	-	-	81
Mr_01559416	01	CG	-	-	-	-	-	81
Mr_01559417	01	CG	-	-	-	-	-	81
Mr_01559418	01	CG	-	-	-	-	-	81
Mr_01559419	01	CG	-	-	-	-	-	81
Mr_01559420	01	CG	-	-	-	-	-	81
Mr_01559421	01	CG	-	-	-	-	-	81
Mr_01559422	01	CG	-	-	-	-	-	81
Mr_01559423	01	CG	-	-	-	-	-	81
Mr_01559424	01	CG	-	-	-	-	-	81
Mr_01559425	01	CG	-	-	-	-	-	81
Mr_01559426	01	CG	-	-	-	-	-	81
Mr_01559427	01	CG	-	-	-	-	-	81
Mr_01559428	01	CG	-					

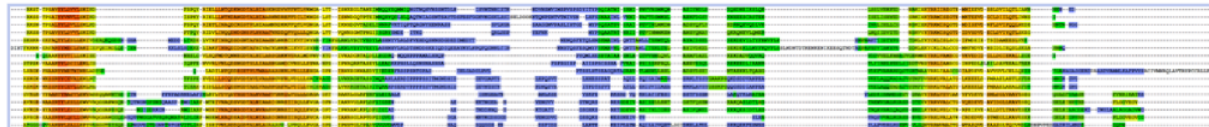
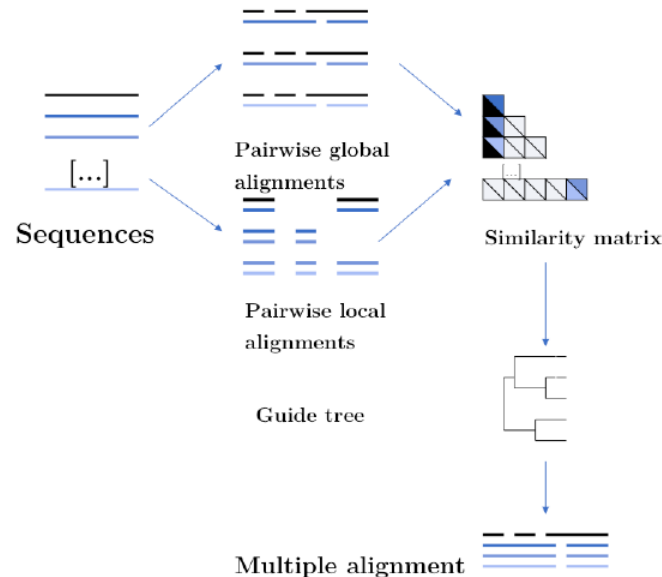
Alineamiento Múltiple de Secuencias (MSA)

Alineamiento basado en perfiles (iterativos):

TCoffee - a hybrid algorithm significantly improves performance:

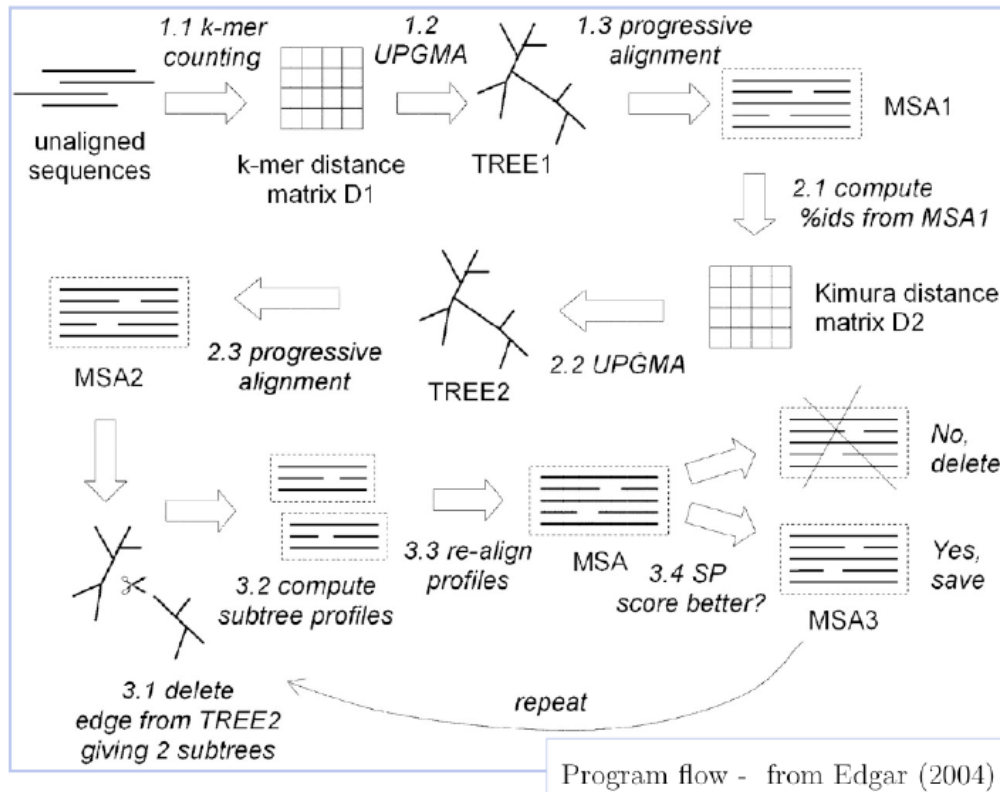
1. Compute **global** pairwise similarity matrix.
2. Compute top 10 non-intersecting **local** alignments.
3. Combine by looking at triplets of sequences.
4. Build guide tree.
5. Align according to tree.

Very good results.



Alineamiento Múltiple de Secuencias (MSA)

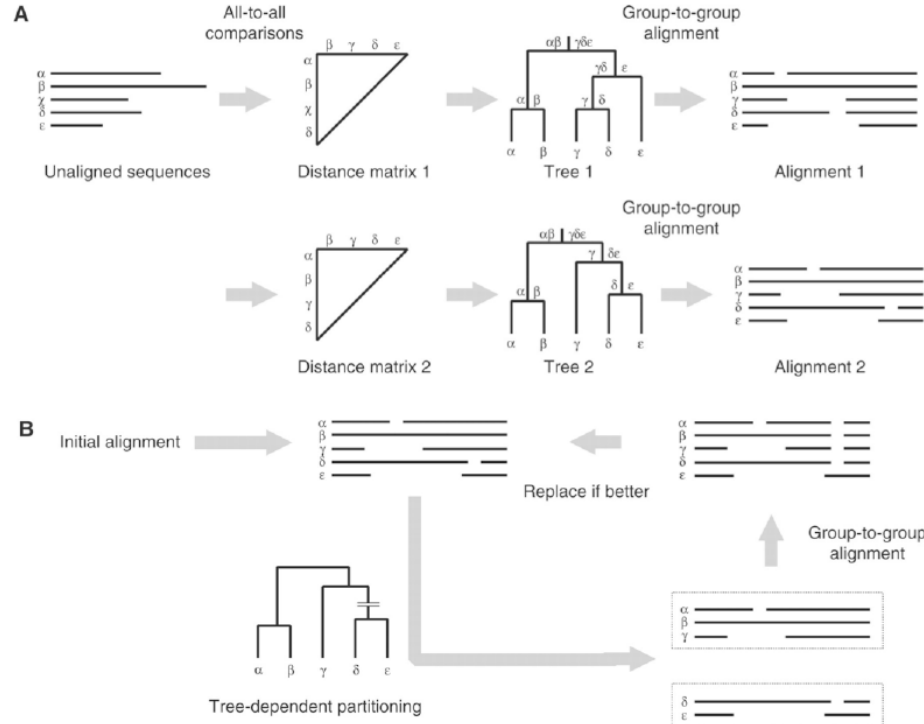
MUSCLE (MSA)



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)



Kazutaka Katoh &
Hiroyuki Toh (2008)

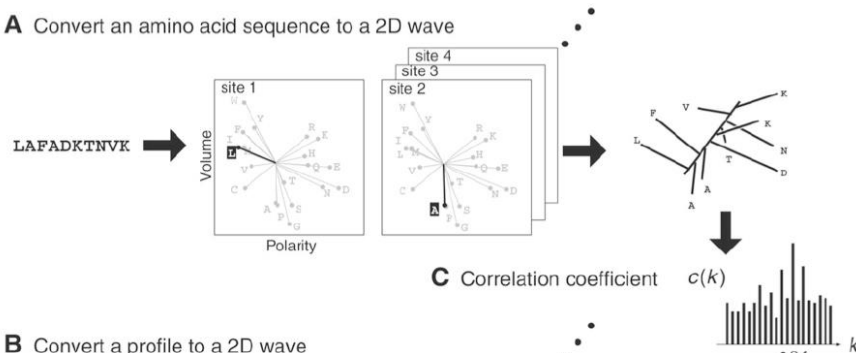
**Recent
developments in the
MAFFT multiple
sequence alignment
program**

*Briefings in
Bioinformatics* 9(4):
286-298.

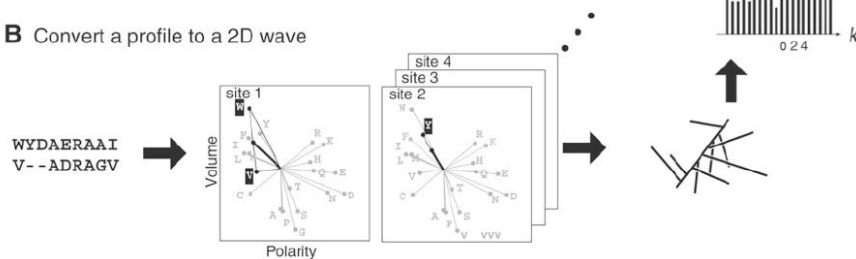
Alineamiento Múltiple de Secuencias (MSA)

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

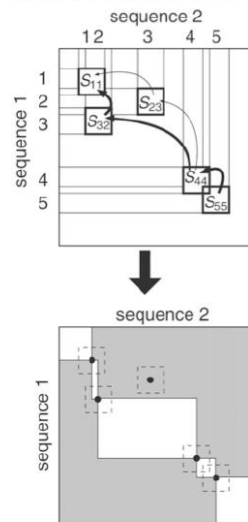
A Convert an amino acid sequence to a 2D wave



B Convert a profile to a 2D wave



D Restrict the area of the DP matrix



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

Alineamiento Múltiple de Secuencias (MSA)

Existen muchos programas para realizar alineamientos múltiples de secuencias

Alineamientos pares de secuencias es problema resuelto

Alineamientos múltiples 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 parólogos 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