FACULTAD DE INGENIERÍA Y CIENCIAS EXACTAS
DEPARTAMENTO DE BIOTECNOLOGÍA Y TECNOLOGÍA ALIMENTARIA
UNIVERSIDAD ARGENTINA DE LA EMPRESA

Bioinformática

ANÁLISIS COMPUTACIONAL DE SECUENCIAS

Dr. Lucas L. Maldonado (PhD)

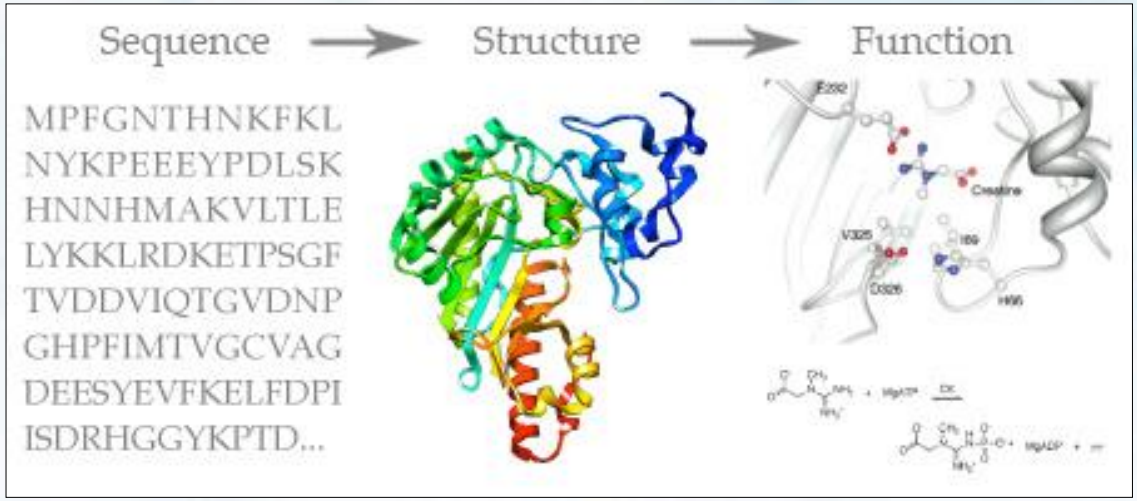
Lic. Biotechnologist and Molecular Biologist

Bioinformatics and genomics specialist

CONICET
Fac. de Medicina - UBA
Fac. de Ciencias Exactas y Naturales – UBA

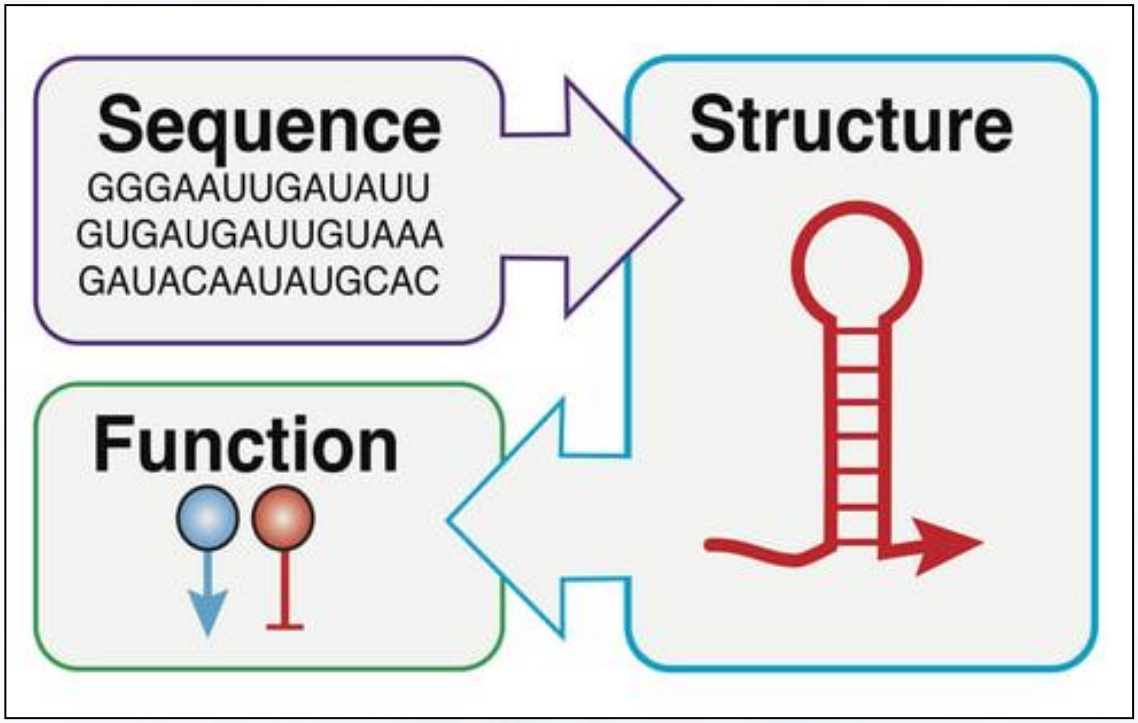
lucamaldonado@uade.edu.ar
lmaldonado@fmed.uba.ar
lucas.l.maldonado@gmail.com.ar

Comparación de secuencias



Las secuencia del ADN determina la secuencia de una proteína.

La secuencia de una proteína determina su estructura 3D.



La estructura 3D de una proteína determina su función biológica.

Por tanto, es muy probable que secuencias similares den lugar a proteínas con estructura y función parecidas.

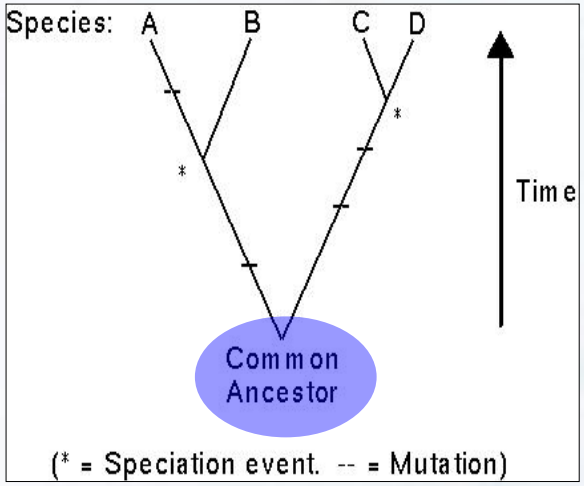
Secuencia → Estructura → Función

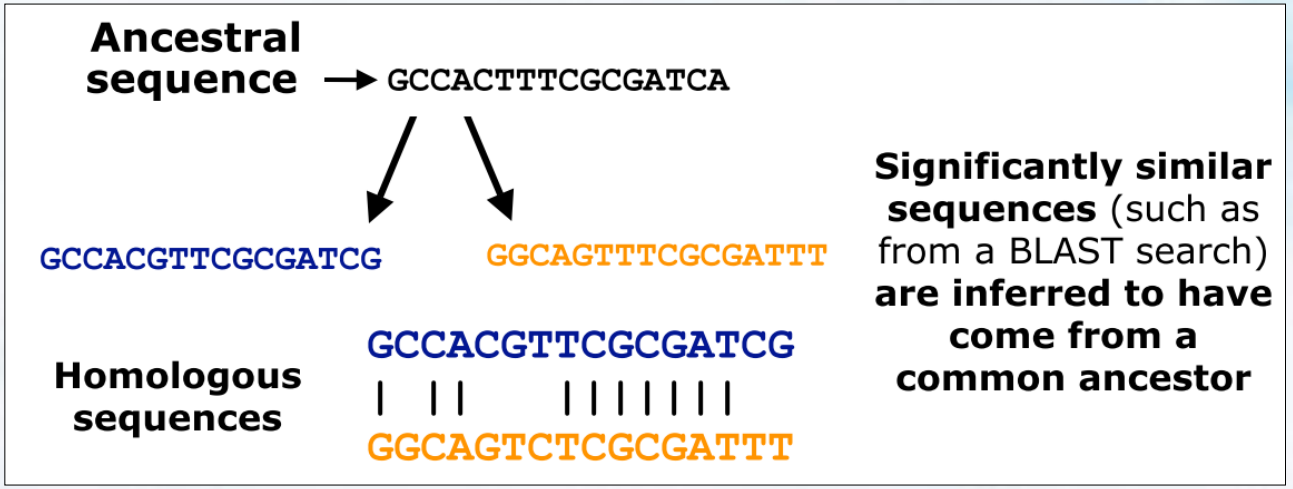
Comparación de secuencias

Similarity implies homology

The probability of two independent randomly evolving sequences converging over any but very small lengths is infinitesimally small.

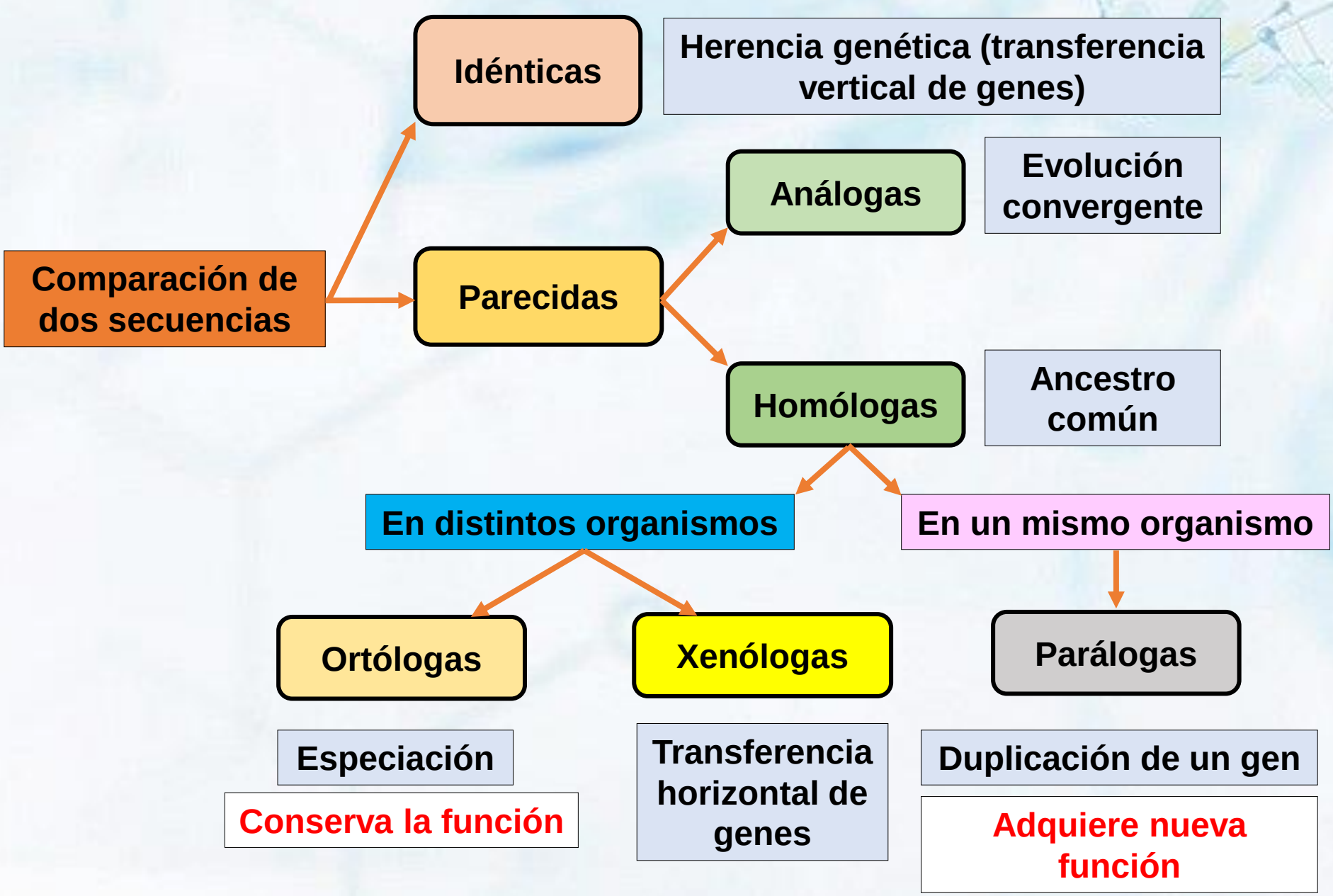
Sequences more similar than expected from random are therefore inferred to have evolved from a common ancestor.





La similitud implica homología …

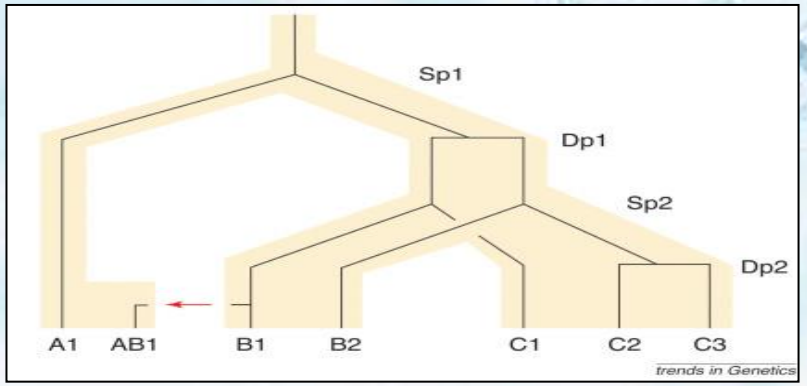
Posibles causas del parecido entre dos secuencias



Diversos tipos de homología

Homólogas:

secuencias que proceden de una misma secuencia ancestral y que, por tanto, presentan cierto grado de similitud.



Parálogas:

secuencias de un mismo organismo, que han aparecido tras un proceso de duplicación génica. Pueden adquirir distinta función.

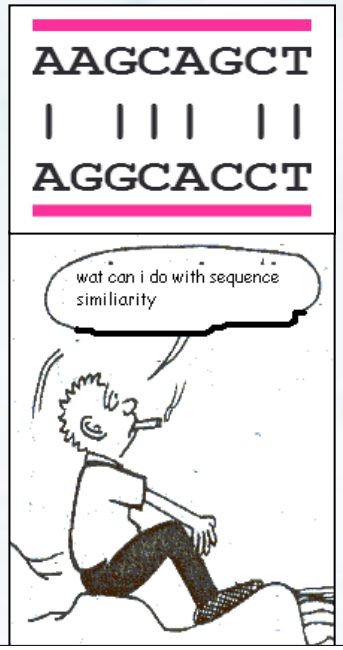
Ortólogas:

secuencias de organismos distintos, que han aparecido durante el proceso de especiación. Conservan la misma función.

Xenólogas:

secuencias de organismos distintos, que han aparecido tras un proceso de transferencia horizontal de genes (virus, simbiosis, etc.)

Comparación de secuencias



Evolution	Homology	%Identity	Alignment methods
Recent relationship or less divergence	Sequence alignments can be used to infer Homology	100 90 80 70 60 50 40 30 20 10 0	Automatic Pairwise Alignment Methods
			Consensus Methods
Increasing divergence	Twilight Zone		Profile Methods
Distant relationship	Midnight Zone		Structure Prediction

Fig 4.1 Percentage identity is an important indicator of the level of evolutionary divergence and functional/structural similarity between compared sequences. Different alignment methods have different areas of optimum application. Pairwise alignment algorithms, for example, perform well at high levels of identity, but below ~50%, the use of consensus information (from multiple alignments) may be necessary. Below ~30%, profile methods are generally used, because they allow insertions, deletions and substitutions to be modelled. Finally, at the lowest levels of identity, where alignments are no longer statistically significant, structure prediction algorithms tend to be used.

¿Que puedo comparar?

¿Contra que voy a comparar?

¿Qué técnica voy a usar?

Aplicaciones del alineamiento de secuencias

Database searches are useful for finding homologues

- Database searches don’t provide precise comparisons
- More precise tools are needed to analyze the sequences in detail including
  - Dot plots for graphic analysis
  - Local or global alignments for residue/residue analysis

Aplicaciones del alineamiento de secuencias

- Alineamientos globales de secuencias de nucleótidos:

- Comparar genomas o cromosomas completos

Análisis genéticos y/o genómicos

Buscar variantes estructurales entre dos genomas:

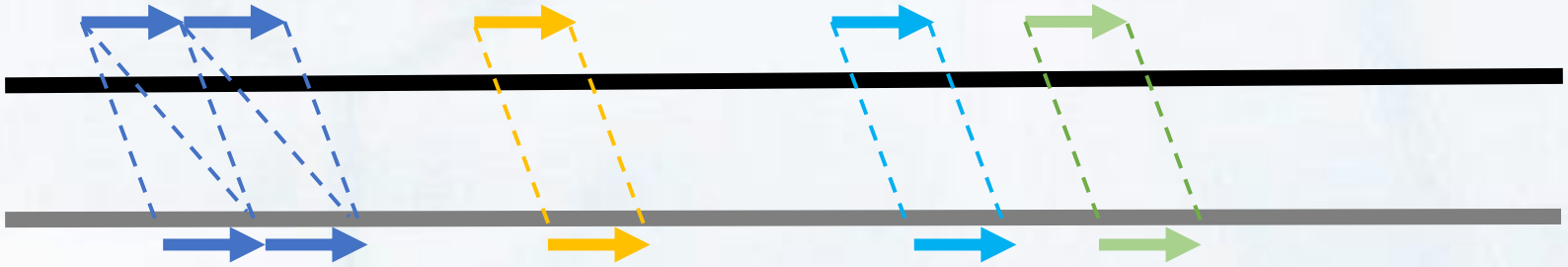
- Traslocaciones
- Transversiones
- Inversiones
- Repeticiones
- Cambios en el numero de copias de genes u otros elementos
- Análisis de variantes genéticas (SNPs e INDELS) – (Aunque menos exacto)
- Buscar zonas conservadas → Si se conservan son importantes o cumplen alguna función
- Determinar la presencia o ausencia de regiones determinadas o genes

Aplicaciones del alineamiento de secuencias

- Alineamientos globales de secuencias de nucleótidos:
- Comparar genomas o cromosomas completos

Anotación estructural de genomas

Sabiendo la localización de los genes en un genomas de referencia (Subject) y teniendo la secuencia genómica de la secuencia query puede inferirse la localización genómica de los genes siempre y cuando los genomas de las especies que se están comparando estén altamente relacionadas, de otro modo puede conducir a errores de anotación.



Aplicaciones del alineamiento de secuencias

- Alineamientos globales de secuencias de nucleótidos:
- Comparar genes completos

¿Por qué me interesa comparar genes?

Aplicaciones del alineamiento de secuencias

- Alineamientos globales de secuencias de nucleótidos:
- Comparar genes completos

¿Por qué me interesa comparar genes?

Gen vs Gen

- Determinar genes homologos
- Cambios o variantes entre 2 genes
- Determinar conservación → Función
- Determinar Dominios compartidos
- Localizar Motivos
- Determinar los productos de genes

Gen vs Cromosoma

- Determinar la localización de los genes
- Estudiar en contexto genómico de un gen
- Localizar intrones y exones
- Estudiar regiones promotoras basados en dominios o regiones características del genomas

Aplicaciones del alineamiento de secuencias

- Alineamientos globales de secuencias de amino ácidos:
- Comparar proteínas completas

¿Por qué me interesa comparar proteínas?

Aplicaciones del alineamiento de secuencias

- Alineamientos globales de secuencias de amino ácidos:
- Comparar proteínas completas

¿Por qué me interesa comparar proteínas?

Proteina vs Proteina

- Determinar genes homólogos
- Sustituciones de aminoácidos o deleciones
- Determinar conservación → Función
- Determinar Dominios compartidos
- Localizar Motivos
- Determinar los productos de genes
- Estudiar las estructuras de la proteínas

Gen vs Proteina

- Determinar el marco de lectura del gen
- mRNA vs proteínas permite identificar regiones 5'UTR y 3'UTR

Aplicaciones del alineamiento de secuencias

Alineamientos Locales de secuencias (BLAST):

Los métodos de alineamientos locales tienen 2 aplicaciones principales:

1. Búsqueda en bases de datos
2. Búsqueda de Dominios y motivos en genes o proteínas

Existen otras aplicaciones de alineamientos locales (BLAST) que requieren de algoritmos o pipelines de análisis mas complejos:

Búsqueda de ortólogos: Reciprocal Best Hits (RBH) are a common proxy for orthology in comparative genomics. Essentially, a RBH is found when the proteins encoded by two genes, each in a different genome, find each other as the best scoring match in the other genome

Scaffolding: Reconstruction de scaffolds o cromosomas basados en un genoma de referencia.

Anotación génica estructural: Inferencia de la localization génica por homología

Anotación Funcional de genes: Inferencia de la función génica por homología.

Aplicaciones del alineamiento de secuencias

Un aspecto importante de la **cacterización de secuencia biológicas son los motivos y los dominios,** ya que sirven **para caracterizar funciones de proteínas desconocidas.**

Secuencias conservadas son secuencias de amino ácidos o nucleótidos similares o idénticos que pueden encontrarse en distintos organismos.

Las secuencias conservadas se denominan **motivos,** que es un elemento conservado en la secuencia de aminoácidos o nucleótidos, que habitualmente se asocia con una función concreta.

Los **motivos** se generan a partir de alineamientos múltiples de secuencias con elementos funcionales o estructurales conocidos, por lo que son útiles para predecir la existencia de esos mismos elementos en otras proteínas de función y estructura

¿Pero que son los Dominios?

Un ***dominio*** es un término más genérico que designa una región de una proteína con interés biológico funcional o estructural.

También se llama ***dominio*** a una región de la estructura tridimensional de una proteína con una función concreta, que incluye regiones no necesariamente contiguas en la secuencia de aminoácidos.

Según Función

Según estructura

Regiones conservadas en una familia de proteínas

Un dominio proteico puede ser funcional si es una unidad modular de la proteína que lleva a cabo una función bioquímica determinada, y estructural si se refiere a un componente estable de la estructura.

Los Dominios son los encargados de:

La presencia secuencias específicas de ciertos aminoácidos pueden contribuir a:

la formación de sitios activos o sitios catalíticos.

mantener la estructura de la proteína

Contener motivos característicos

La secuencia de amino ácidos determina la estructura secundaria de las proteínas de manera que contribuyen a:

la formación de sitios activos o sitios catalíticos

mantener la estructura de la proteína

Estabilizar la estructura nativa funcional de las proteínas

Dinamica molecular de las proteina

A nivel funcional

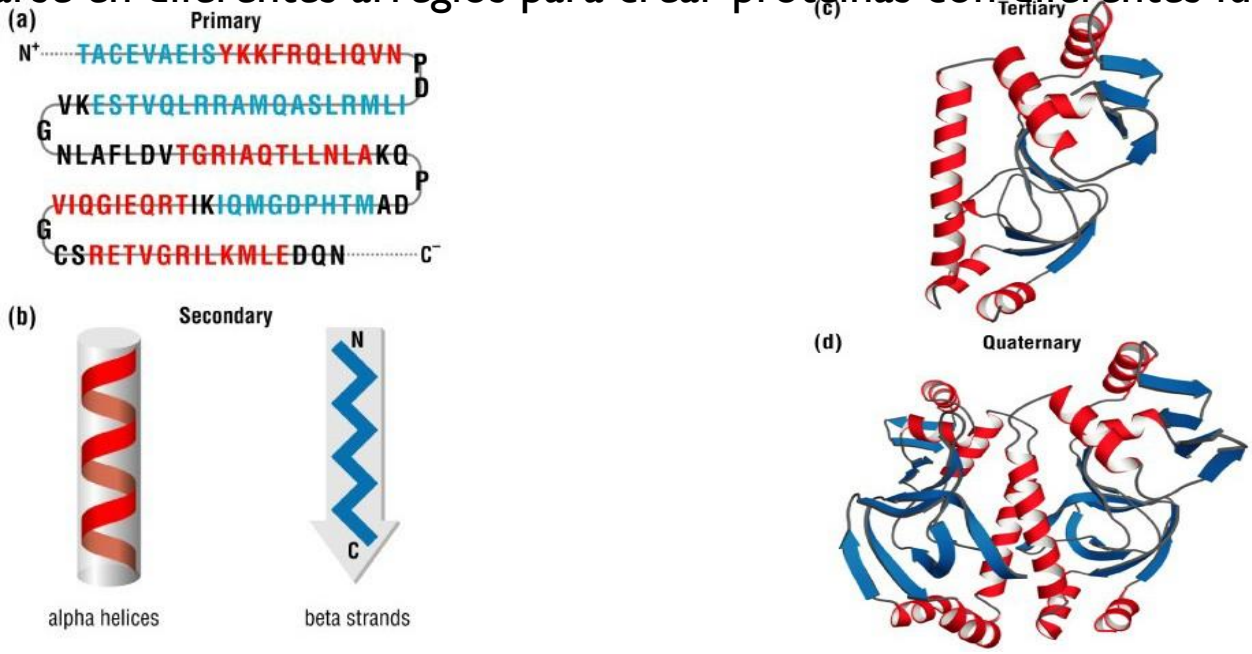
A nivel estructural

Regiones conservadas en una familia de proteínas

Un dominio En general, los dominios varían en longitud desde aproximadamente 50 aminoácidos hasta 250 aminoácidos de longitud. puede ser funcional si es una unidad modular de la proteína que lleva a cabo una función bioquímica determinada, y estructural si se refiere a un componente estable de la estructura.

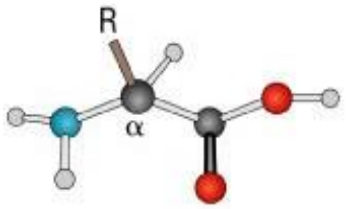
Los cuatro niveles de la estructura proteica

- Un dominio de proteína es una parte conservada de una secuencia de proteína dada y una estructura terciaria que puede evolucionar, funcionar y existir independientemente del resto de la cadena de proteína.
- Cada dominio forma una estructura tridimensional compacta y, a menudo, puede ser estable y plegado de forma independiente.
- Muchas proteínas constan de varios dominios estructurales.
- Un dominio puede aparecer en una variedad de proteínas diferentes.
- La evolución molecular usa dominios como bloques de construcción y estos pueden recombinarse en diferentes arreglos para crear proteínas con diferentes funciones.

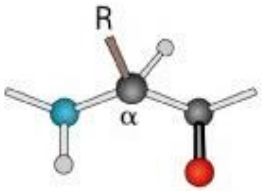


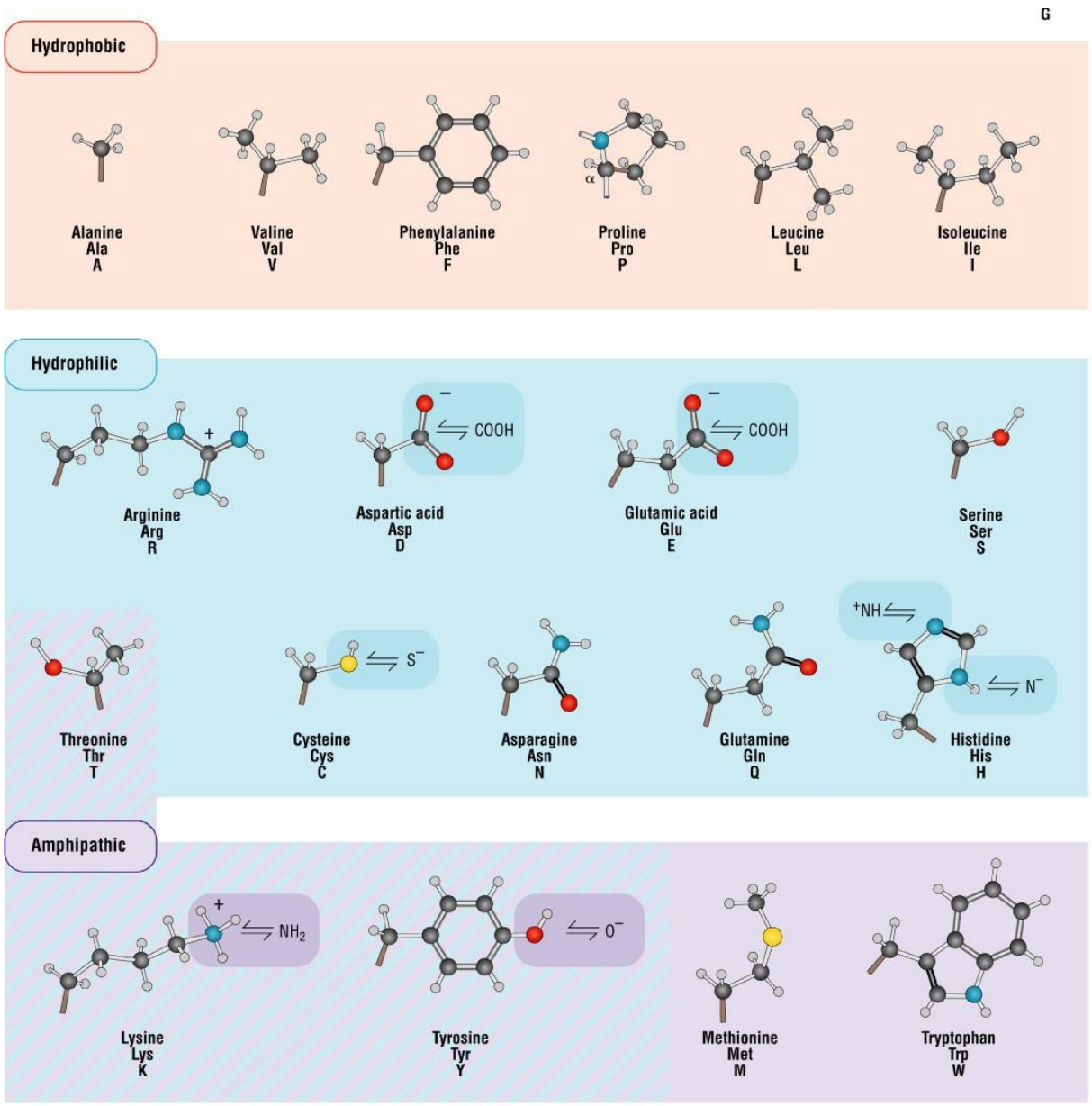
Los amino ácidos: estructura y carácter químico

Estructura general de un aminoácido

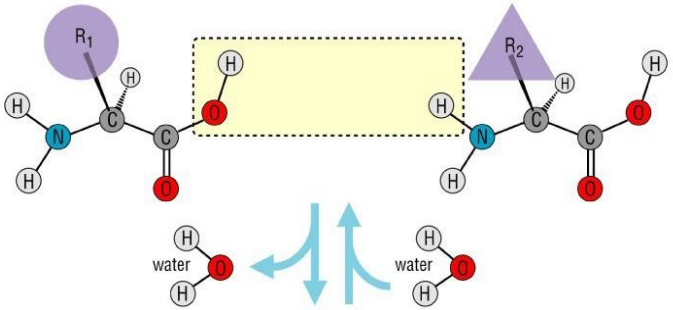


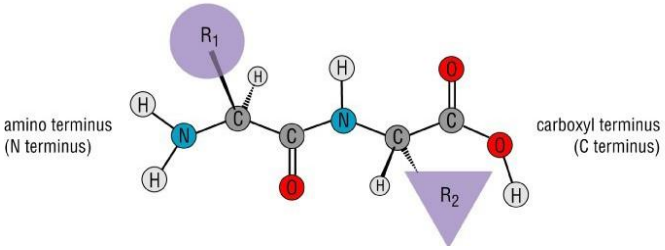
Formando enlace peptídico

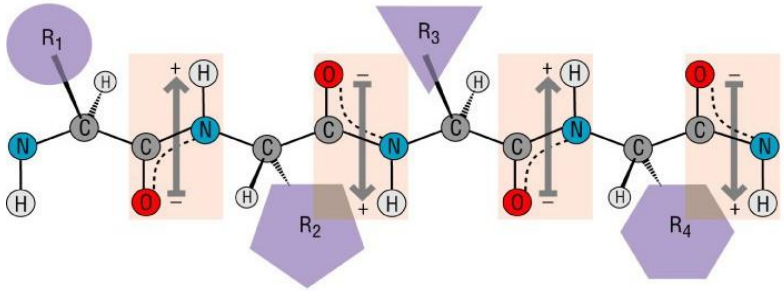




La propiedades del enlace peptídico afectan la estabilidad y flexibilidad de las proteínas

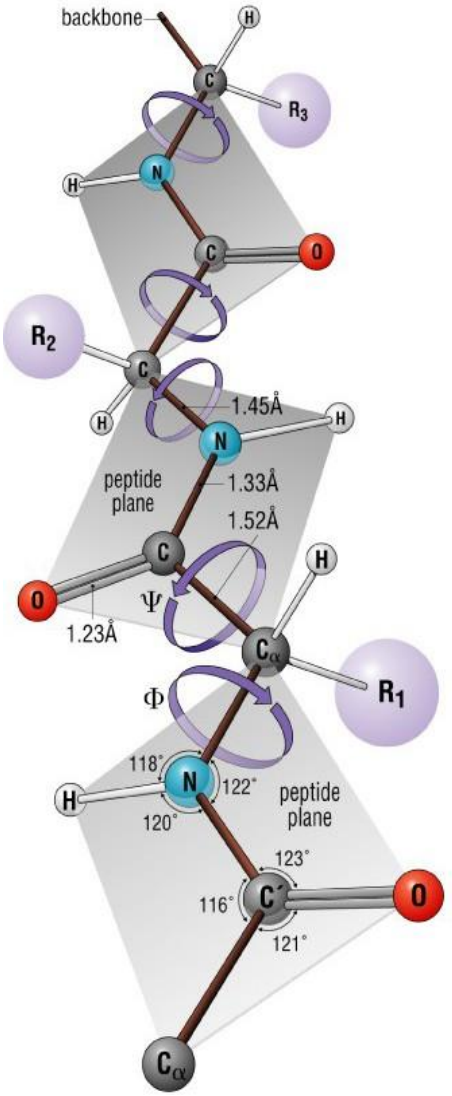






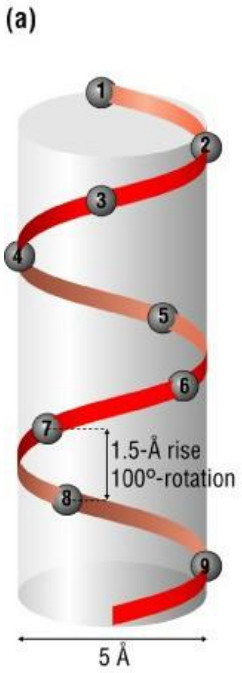
Los enlaces tipo amida son muy estables

Caracter de doble enlace parcial afecta la rotación de la cadena polipeptídica

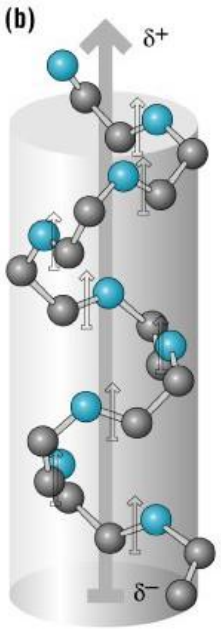


Resonancia de los enlaces tiene dos efectos: incremento de la estabilidad y momento dipolar

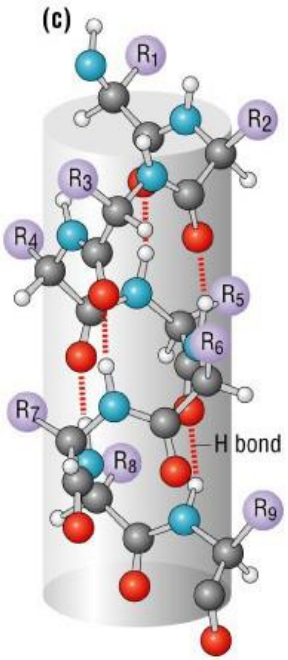
Estructura secundaria: Alfa hélices



Distancias

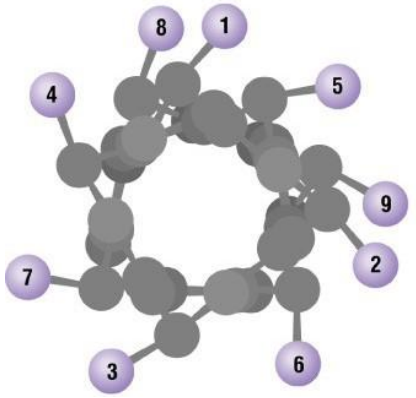


Dipolo



Puentes de hidrógeno

Vista superior



Las cadenas laterales determinan el carácter hidrofílico, hidrofóbico o anfipático de una alfa hélice

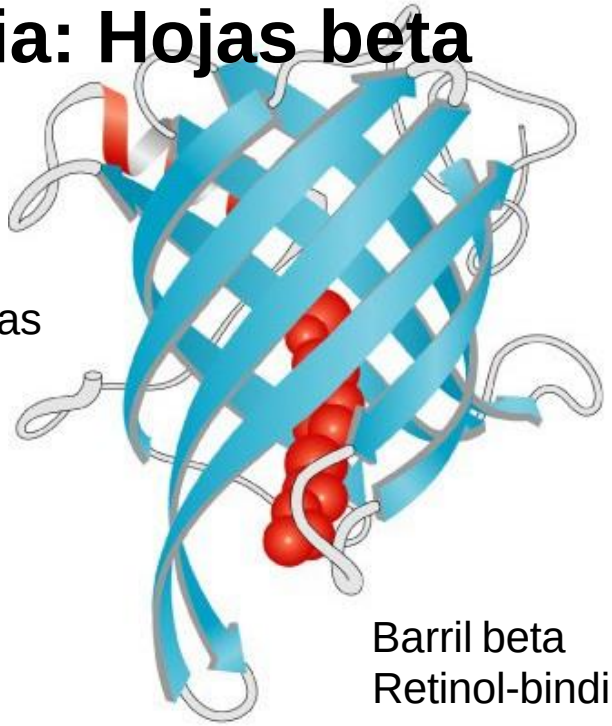
Variantes poco frecuentes de alfa hélices

Average Conformational Parameters of Helical Elements

Conformation	Phi	Psi	Omega	Residues per turn	Translation per residue
Alpha helix	–57	–47	180	3.6	1.5
3-10 helix	–49	–26	180	3.0	2.0
Pi-helix	57	–70	180	4.4	1.15
Polyproline I	–83	+158	0	3.33	1.9
Polyproline II	–78	+149	180	3.0	3.12
Polyproline III	–80	+150	180	3.0	3.1

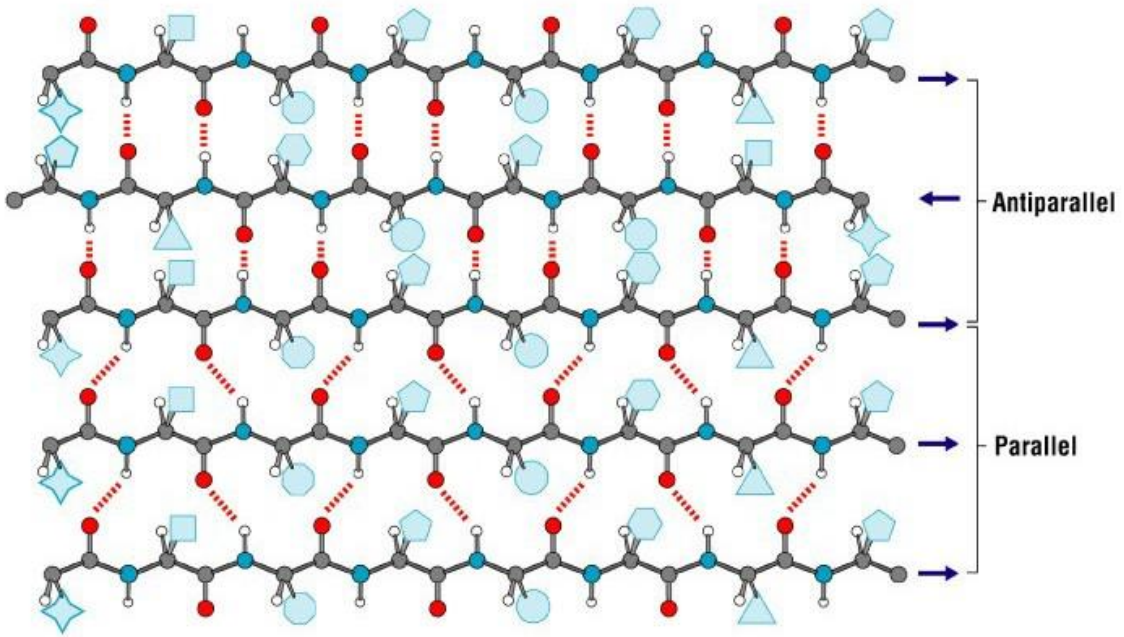
Estructura secundaria: Hojas beta

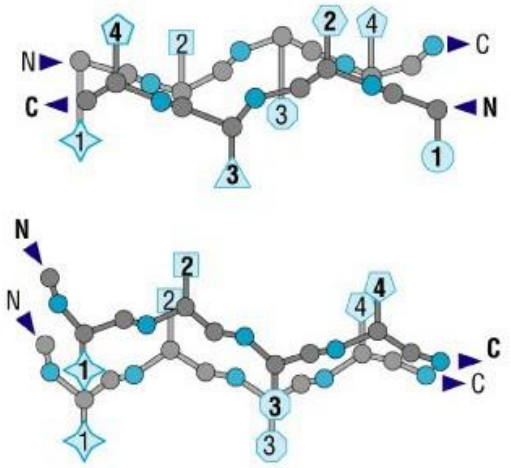
Hojas beta anfipáticas



Barril beta
Retinol-binding protein

Estructura hoja beta





¿Cual es la diferencia entre motivos y Dominios?

Dominios	Los dominios son unidades estructuralmente grandes.
	Son segmentos de proteína plegados independientemente que se pueden diferenciar tanto estructural como funcionalmente.
	es una estructura estable con una función específica y definida en la proteína y puede existir independientemente de la proteína
	Permiten predecir funciones biológicas y ususalmente se encuentran en proteínas con funciones homologas
Motivos	El motivo es una disposición de estructuras secundarias de la molécula de proteína.
	No suele ser estable por sí mismo, a diferencia de un dominio.
	Puede o no formar parte de un dominio.
	No permiten predecir funciones biológicas y pueden encontrarse en proteínas con funciones diferentes
	Pueden Distinguirse motivos de ADN o de proteínas: pequeños segmentos con alguna característica particular

¿CÓMO PODEMOS ENCONTRAR DOMINIOS EN UNA PROTEÍNA?

Los motivos y dominios son construidos a partir de un MSA (Alineamiento múltiple de Secuencias) en las cuales, las secuencias están relacionadas entre sí.

Esto sirve para hallar las regiones conservadas.

Una vez halladas las regiones que se consideran motivos y dominios se prosigue a almacenar la información de consenso en una base de datos para que así sirvan como base en la identificación de las funciones de una proteína desconocida que presente los mismos patrones.

Esto puede ser almacenado de dos formas: Expresiones regulares o mediante un modelo estadístico.

OBJETIVO DEL ALINEAMIENTO DE DOS SECUENCIAS

El interés de comparar dos secuencias o buscar una en una base de datos reside en el hecho que *en las secuencias biológicas (AN, Proteínas) una elevada similitud entre dos secuencias suele conllevar similitudes en su estructura o función.*

El objetivo de la comparación a pares es pues encontrar secuencias con (sub)patrones comunes pero de las que no se conoce previamente si están relacionadas biológicamente

OBJETIVOS GENERALES DE LOS MSA

- Un MSA genera información resumida sobre un conjunto de secuencias relacionadas

- Esta sirve para distintos fines
  - Para descubrir **patrones comunes** en las secuencias identificados con estructura/función
  - Para poder decidir sobre las relaciones (evolutivas) entre ellas.
  - Para la identificación de miembros potenciales.
  - Para revelar disimilaridades.

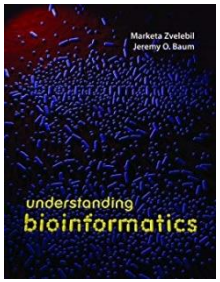


Figure 4.10
Pairwise and multiple alignments of part of the catalytic domains of five PI3-kinases and a cAMP-dependent protein kinase.
(A) Pairwise alignment of PI3-kinase p110α and the protein kinase does not align the important active-site residues and the DFG motif (in green).

(B) Multiple alignment of the protein kinase with a set of five PI3-kinases (which have considerable overall homology to each other) has the effect of forcing the best-conserved regions to be matched. Here the DFG motif and the important N and D (green) residues are aligned correctly in all the sequences. In addition it is apparent that a G (green) is also totally conserved (identical) and that three more residues are conserved in their physicochemical properties (blue).

El alineamiento de dos secuencias se suele utilizar para llevar a cabo búsquedas en BD.

El alineamiento múltiple de secuencias (AMS) permite identificar los residuos conservados que resultan cruciales para mantener la estructura y/o función.

```
(A)  p110α          TFILGIGDRHNSNIMVKDDG-QLFHIDFGHFLDHKKKKFGYKRERVPFVLT--QDFLIVI 142
     cAMP-kinase    QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPE 179

(B)  p110β          SYVLGIG----------DRHSDNINVKKTGQLFHIDFGHILGNFKSKFGIKRERVPFILT 136
     p110δ          TYVLGIG----------DRHSDNIMIRESGQLFHIDFGHFLGNFKTKFGINRERVPFILT 136
     p110α          TFILGIG----------DRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLT 135
     p110γ          TFVLGIG----------DRHNDNIMITETGNLFHIDFGHILGNYKSFLGINKERVPFVLT 135
     p110_dicti     TYVLGIG----------DRHNDNLMVTKGGRLFHIDFGHFLGNYKKKFGFKRERAPFVFT 135
     cAMP-kinase    QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCG--TPEYLA 177
```

La fortaleza de los AMS

Alineamiento de dos secuencias

Los aa importantes para el mantenimiento de la estructura y/o función de la proteína se encuentran bajo presión evolutiva: o se conservan o, si cambian, lo hacen por aa parecidos. Un alineamiento múltiple de secuencias (AMS) de proteínas homólogas permite identificar estos aa.


Si dos secuencias están muy próximas en la evolución, habrán cambiado muy poco y será difícil detectar qué aa son los realmente importantes

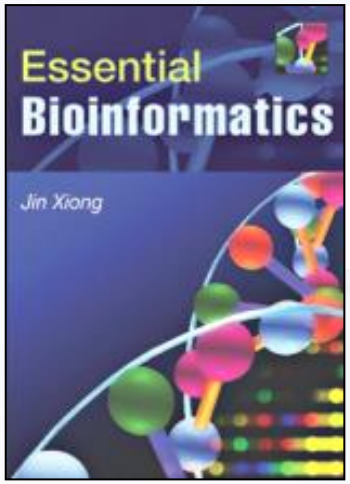

Si dos secuencias están muy lejanas será difícil hacer un alineamiento capaz de detectar los residuos importantes.


Este problema se puede resolver alineando el mayor número posible de secuencias homólogas.

Definición de AMS

Un alineamiento múltiple de secuencias es el emparejamiento de 3 o mas secuencias.

- se obtiene insertando en cada secuencia un cierto número de gaps (quizás 0)
- Las secuencias resultantes tienen (deben tener) la misma longitud
- Cada columna tendrá como mínimo un carácter diferente de '-' ("gaps")



CHAPTER FIVE

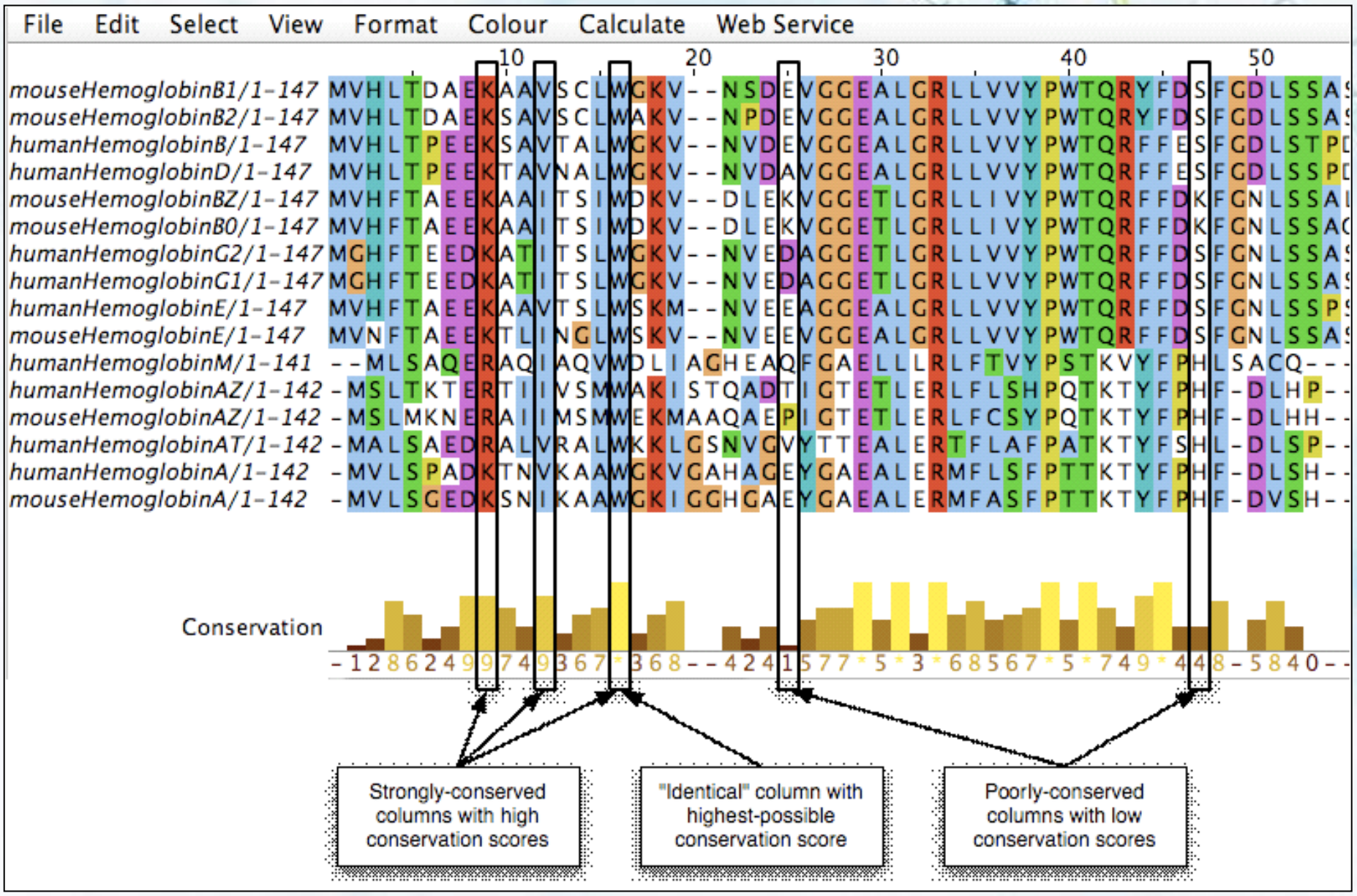
Multiple Sequence Alignment

What is an optimal multiple sequence alignment?

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSVSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCTISDFYPGA--VTVAWKADS--
AALGCTVKDYFPEP--VTVSWNSG---
VSLTCTVKGFYPSD--IAVEWESNG--
```

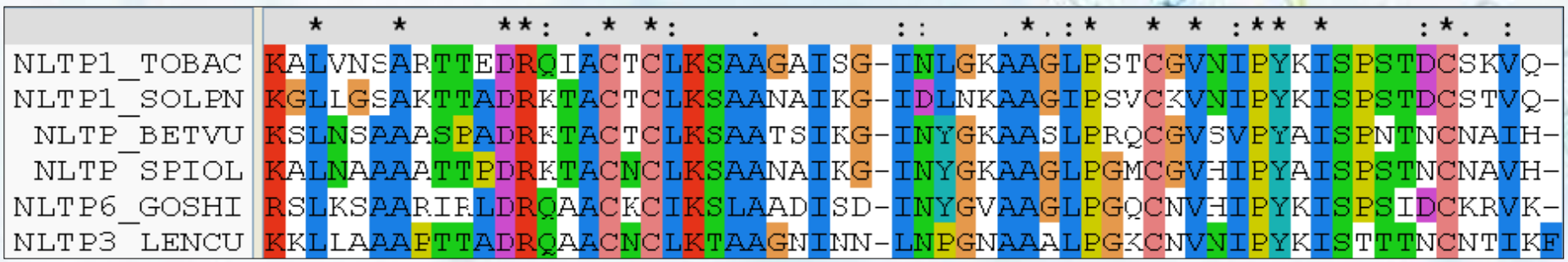
Idea: try to have a maximum of similar residues in a given column of the alignment

Alineamiento óptimo de múltiples secuencias



Distintos grados de conservación

Para ser útil, un MSA debe incluir un amplio rango de similitudes



Multiple sequence alignment

'One amino acid sequence plays coy; a pair of homologous sequences whisper; many aligned sequences shout out loud.' In nature, even a single sequence contains all the information necessary to dictate the fold of the protein.

To be informative a multiple alignment should contain a distribution of closely- and distantly-related sequences. If all the sequences are very closely related, the information they contain is largely redundant, and few inferences can be drawn. If all the sequences are very distantly related, it will be difficult to construct an accurate alignment (unless all the structures are available), and in such cases the quality of the results, and the inferences they might suggest, are questionable. Ideally, one has a complete range of similarities, including distant relatives linked through chains of close relationships.

El porcentaje de identidad entre secuencias

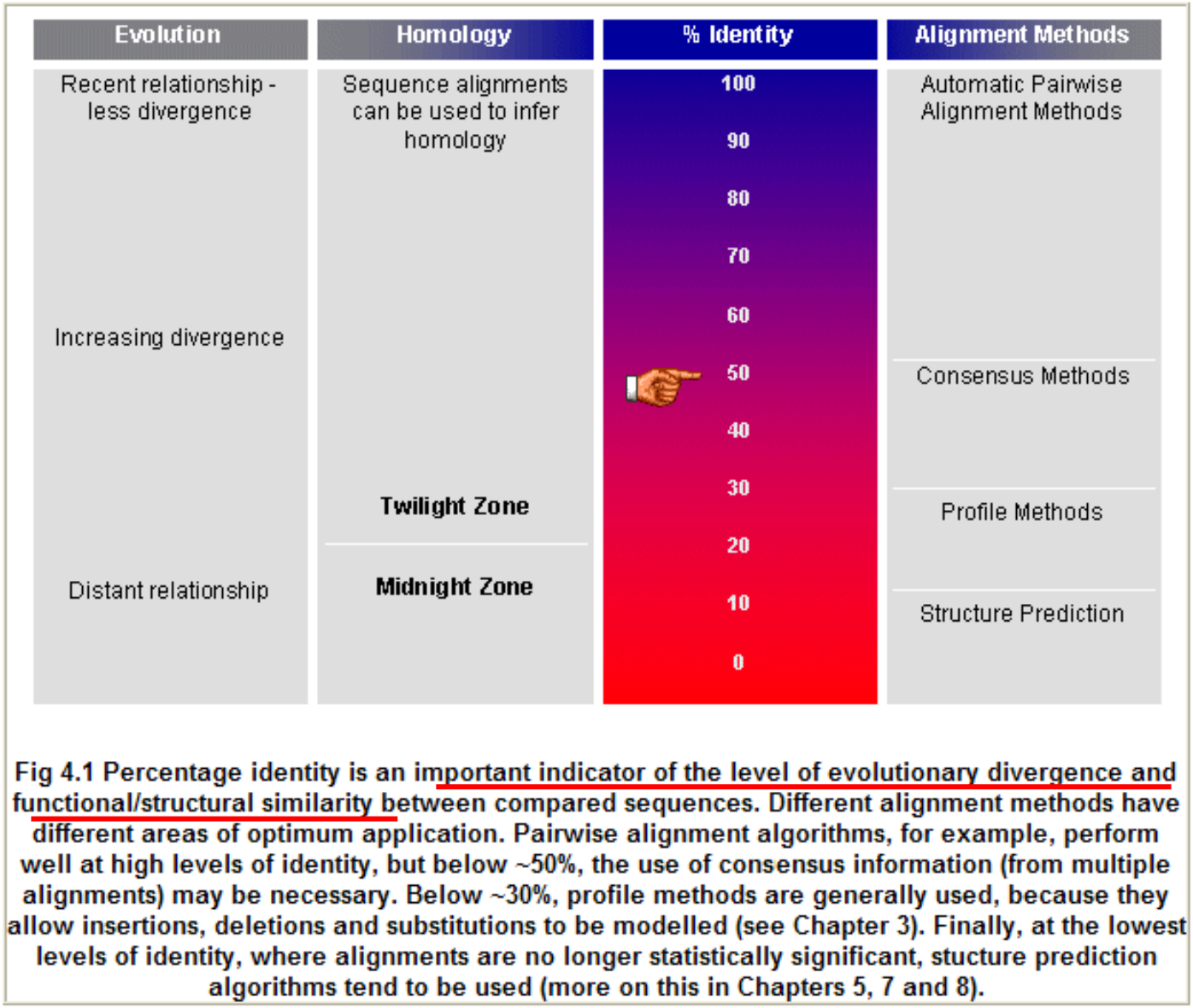
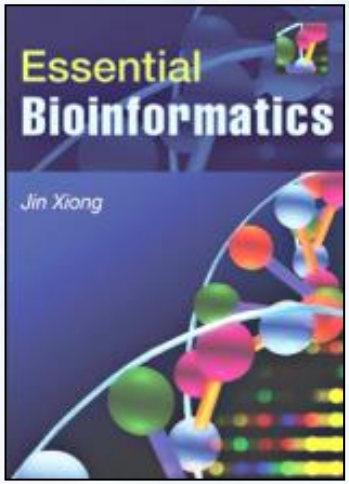


Fig 4.1 Percentage identity is an important indicator of the level of evolutionary divergence and functional/structural similarity between compared sequences. Different alignment methods have different areas of optimum application. Pairwise alignment algorithms, for example, perform well at high levels of identity, but below ~50%, the use of consensus information (from multiple alignments) may be necessary. Below ~30%, profile methods are generally used, because they allow insertions, deletions and substitutions to be modelled (see Chapter 3). Finally, at the lowest levels of identity, where alignments are no longer statistically significant, stucture prediction algorithms tend to be used (more on this in Chapters 5, 7 and 8).



CHAPTER FIVE

Multiple Sequence Alignment

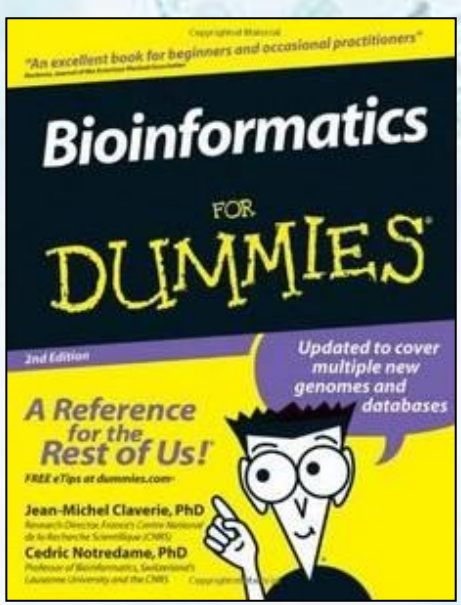
A natural extension of pairwise alignment is multiple sequence alignment, which is to align multiple related sequences to achieve optimal matching of the sequences. Related sequences are identified through the database similarity searching described in Chapter 4. As the process generates multiple matching sequence pairs, it is often necessary to convert the numerous pairwise alignments into a single alignment, which arranges sequences in such a way that evolutionarily equivalent positions across all sequences are matched.

There is a unique advantage of multiple sequence alignment because it reveals more biological information than many pairwise alignments can. For example, it allows the identification of conserved sequence patterns and motifs in the whole sequence family, which are not obvious to detect by comparing only two sequences. Many conserved and functionally critical amino acid residues can be identified in a protein multiple alignment. Multiple sequence alignment is also an essential prerequisite to carrying out phylogenetic analysis of sequence families and prediction of protein secondary and tertiary structures. Multiple sequence alignment also has applications in designing degenerate polymerase chain reaction (PCR) primers based on multiple related sequences.

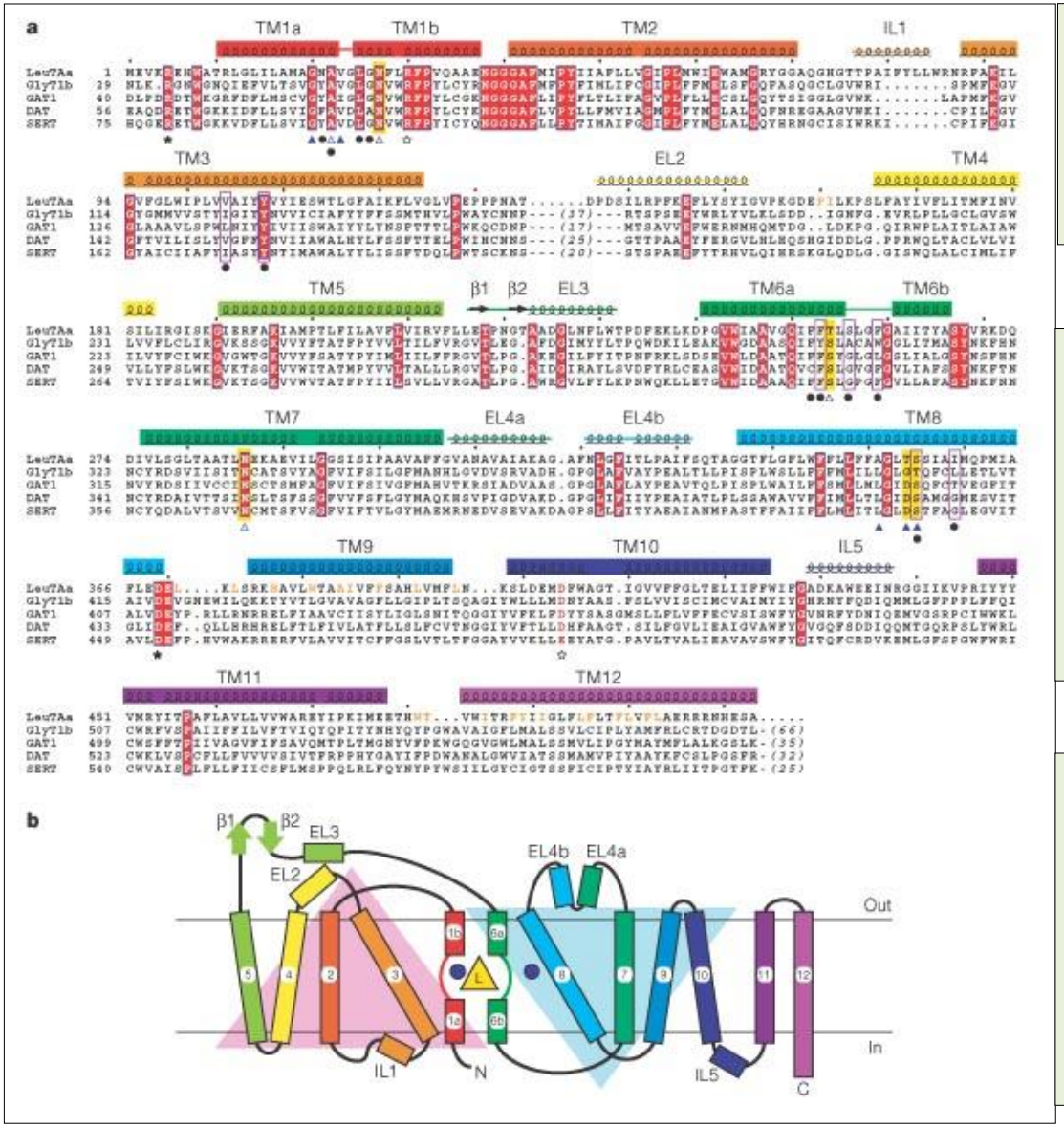
Criterios para la construcción de un AMS

Table 9-1 Main Criteria for Building a Multiple Sequence Alignment

Criterion	*Meaning*
Structural similarity	Amino acids that play the same role in each structure are in the same column. Structure-superposition programs are the only ones that use this criterion.
Evolutionary similarity	Amino acids or nucleo tides related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column. No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it.
Functional similarity	Amino acids or nucleotides with the same function are in the same column. No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it — or you can edit your alignment manually.
Sequence similarity	Amino acids in the same column are those that yield an alignment with maximum similarity. Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, their structural, evolutionary, and functional similarities are equivalent to sequence similarity.



Regiones conservadas y no conservadas

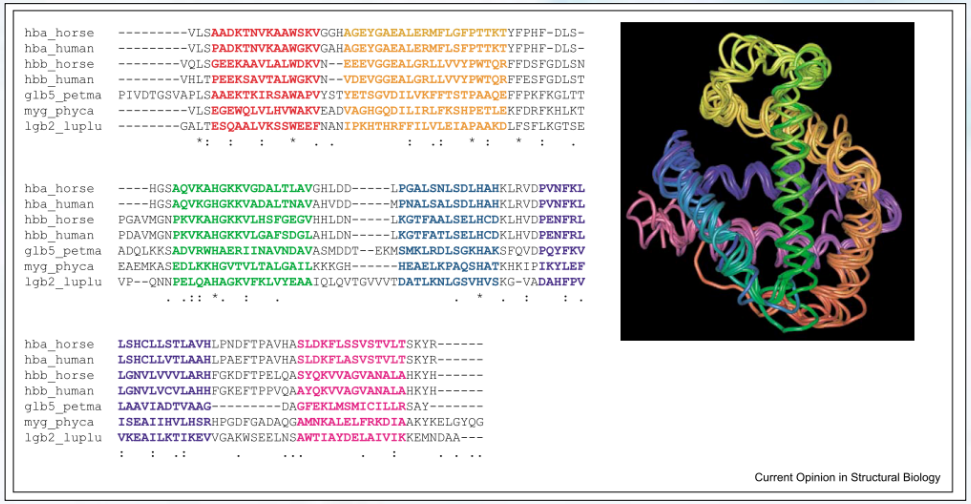


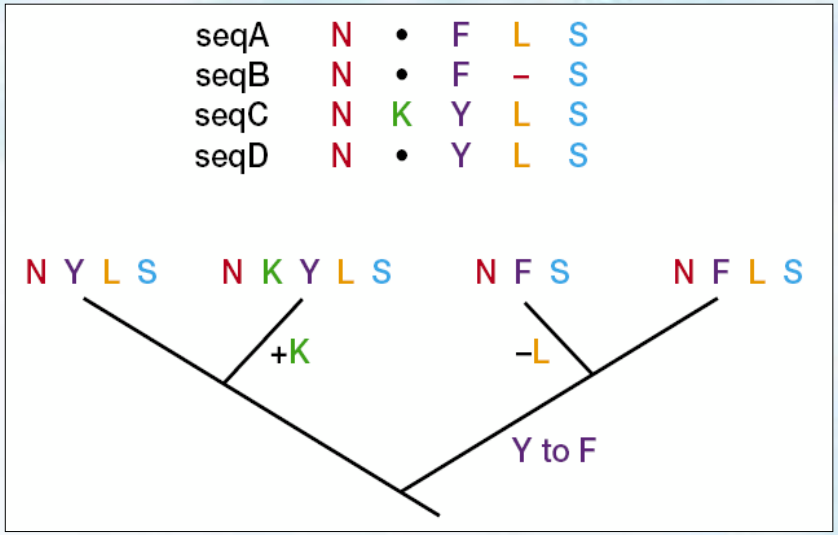
Al alinear muchas secuencias, las columnas que contienen aa idénticos o similares destacarán claramente.

Estos aa conservados corresponden a aquellas regiones de la proteína que son importantes para el mantenimiento de la estructura y/o función.

Las regiones de la proteína que toleran indels suelen corresponder a regiones expuestas (bucles sin elementos de estructura secundaria.

MSA: predicción de estructuras y filogenia





Superposición de secuencias y estructuras: Los bloques conservados están relacionados a la estructura de la proteína y son importantes para la función de la misma

MSA y relaciones filogenéticas: La conservación de dominios puede definir las familias de proteínas y la conservación de residuos específicos pueden definir la topología

Aplicaciones de los MSA

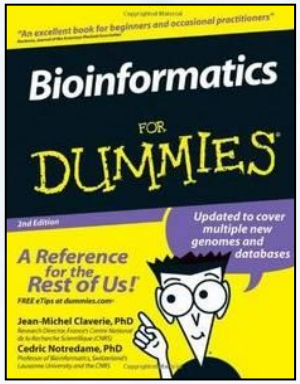


Table 9-2 Main Applications of Multiple Sequence Alignments

Application	*Procedure*
Extrapolation	A good multiple alignment can help convince you that an uncharacterized sequence is really a member of a protein family. Alignments that include Swiss-Prot sequences are the most informative. Use the ExPASyBLAST server (at `www.expasy.ch/tools/blast/`) to gather and align them.
Phylogenetic analysis	If you carefully choose the sequences you include in your multiple alignment, you can reconstruct the history of these proteins. Use the Pasteur Phylip server at `bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html`.
Pattern identification	By discovering very conserved positions, you can identify a region that is characteristic of a function (in proteins or in nucleic-acid sequences). Use the logo server for that purpose: `www-lmmb.ncifcrf.gov/~toms/sequencelogo.html`.
Domain identification	It is possible to turn a multiple sequence alignment into a profile that describes a protein family or a protein domain (PSSM). You can use this profile to scan databases for new members of the family. Use NCBI-BLAST to produce and analyze PSSMs: `www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml#pssm`.

Aplicaciones de los AMS

Aplicaciones de los MSA

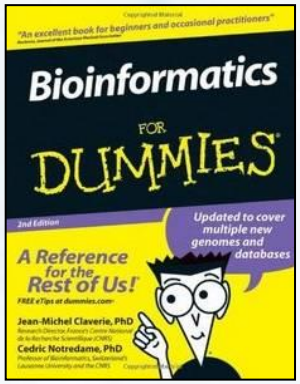
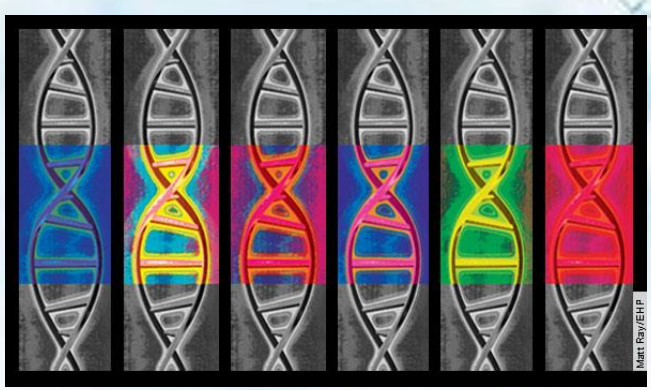
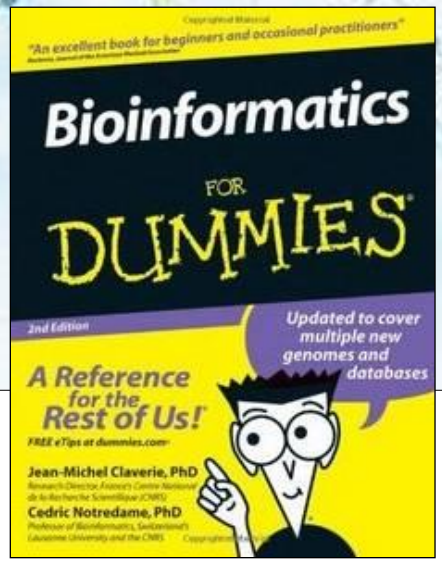


Table 9-2 Main Applications of Multiple Sequence Alignments

Application	*Procedure*
DNA regulatory elements	You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potentially similar binding sites. Use the Gibbs sampler to identify these sites: `bayesweb.wadsworth.org/gibbs/gibbs.html`
Structure prediction	A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for both proteins and RNA. Sometimes it can also help in the building of a 3-D model.
nsSNP analysis	Various gene alleles often have different amino-acid sequences. Multiple alignments can help you predict whether a Non-Synonymous Single-Nucleotide Polymorphism is likely to be harmful. See the SIFT site for more details: `blocks.fhcrc.org/sift/SIFT.html`.
PCR analysis	A good multiple alignment can help you identify the less-degenerated portions of a protein family, in order to fish out new members by PCR (polymerase chain reaction). If this is what you want to do, you can use the following site: `blocks.fhcrc.org/codehop.html`.

Selección de secuencias para un MSA





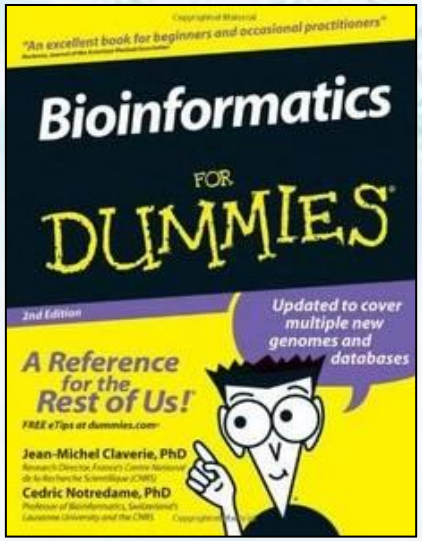
Choosing the Right Sequences

Anyone who ever worked in a lab knows that molecular biology is very much like cooking: It's all about selecting the right ingredients and putting them into the pot at the right time and in the right order.

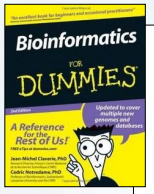
Building a multiple alignment obeys the same rule. Before you build your alignment, you must carefully select the sequences you want to align. These sequences are members of the same protein family, and they all share a common ancestor. The family is usually too large to be entirely included in your multiple alignment, and picking the right sequences is an art. If you want to be good at this game, you need to know what you want to show with your alignment — and you need to know how the multiple alignment programs work.

Selección de secuencias para un MSA

Table 9-3 A Few Guidelines for Selecting Sequences	
Problem	***Diagnostic***
Proteins or DNA	Use proteins whenever possible. You can turn them back into DNA *after* doing the multiple alignment.
Many sequences	Start with 10–15 sequences; avoid aligning more than 50 sequences.
Very different sequences	Sequences that are less than 30 percent identical to more than half the other sequences in the set often cause troubles.
Identical sequences	They never help. Unless you have a *very* good reason to do so, avoid incorporating into your multiple alignment any sequence that's more than 90 percent identical to another sequence in the set.
Partial sequences	Multiple-sequence-alignment programs prefer sequences that are roughly the same length. Programs often have difficulties comparing items in a mixture of complete sequences and shorter fragments.
Repeated domains	Sequences with repeated domains cause trouble for most multiple-alignment programs — especially if the number of domains is different. When this happens, you may be better off extracting the domains yourself with Dotlet or Lalign (see Chapter 8) and making a multiple alignment of those segments.



¿Qué tipo de secuencias necesito para un AMS?



Making the right compromise between similarity and new information

If you think that very similar sequences give very good alignments, you're right! However, a multiple sequence alignment that's correct isn't enough; it must also be useful.

For instance, an alignment that only contains very similar sequences brings little information. You can use it to extrapolate annotations, but you can't do phylogeny, structure prediction, or any of the other useful applications that we list in Table 9-2. These other tasks require being able to observe mutation patterns in every column — which isn't possible if you have an alignment in which most columns are entirely conserved.

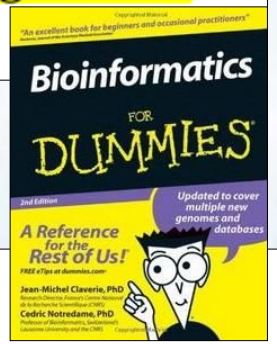
Grabbing the most distantly related sequences you can find doesn't work, either. Multiple-sequence-alignment programs can't use sequences that are too different — even if these sequences are homologous. In fact, two things multiple-sequence-alignment programs *really* don't like are

- Sequences that are very different from every other sequence in the group
- Sequences that need long insertions/deletions to be properly aligned

No es una buena idea seleccionar muchas secuencias

If you can, make sure that *each sequence is between 30 and 70 percent identical with more than half of the sequences in the set.* This way, you're making a reasonable trade-off between new information and alignment quality.

Before adding a sequence to a multiple alignment, you can try to figure out whether it's a good choice by making pairwise comparisons

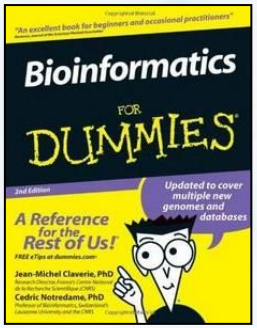


Choosing the right number of sequences

In our opinion, you should start with a relatively small number of sequences — between 10 and 15 sequences would be suitable for most cases. After you get something interesting happening with this small set, you can always increase its size. In any case, it's hard to see any reason for generating a multiple alignment with more than 50 sequences, unless you're interested in building some extensive phylogenetic tree.

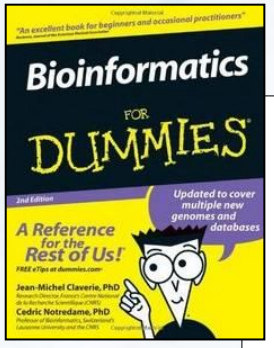
If you start with hundreds of sequences, you immediately hit troubles. There are good reasons why:

Hacer un MSA con muchas secuencias es complicado



- ***Computing* big alignments is difficult.** Public servers do not have infinite resources. Your job may take a very long time to run (if it runs). For you, this makes it difficult to tune parameters and check alternatives.
- ***Building* big alignments is difficult.** Multiple alignment programs are not very good at handling very large sets of sequences.
- ***Displaying* big alignments is difficult.** You can't print them, and they clog your computer when you want to visualize them. If columns are longer than one page, interpretation becomes impossible.
- ***Using* big alignments is difficult.** Tree-building and structure-prediction programs cannot handle them easily.
- **Making *accurate* big alignments is difficult.** You want your multiple sequence alignment to be highly reliable so you can be confident that all the sequences it contains are true members of the family. A major cause for concern is that multiple-sequence-alignment programs make mistakes. The curse is that these mistakes do not add up, they *multiply* — making it easy for a tiny number of bad sequences to ruin an entire alignment. Of course, the more sequences you have, the more likely this is to happen. The best way to avoid such a disaster is to start small — and gradually increase the size of your multiple sequence alignment until it contains all the sequences you're interested in.

Cómo poner nombre a las secuencias de un MSA



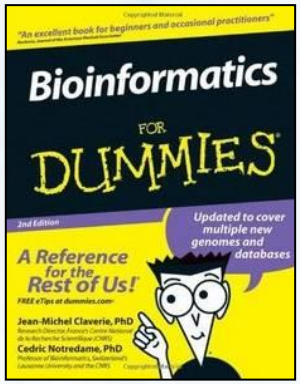
Naming your sequences the right way

Naming may sound like a trivial issue, but it's not. Multiple-sequence-alignment programs have no standard way of handling the sequence names. If you stick to the following four rules, however, you will NEVER get into trouble (if you don't, you're on your own):

- Never use white spaces in your sequence names.
- Do not use special symbols. Stick to plain letters, numbers, and the underscore (_) to replace spaces. Avoid ALL other symbols, especially those that are the most tempting for special sequences (such as @, #, _,^ and so on).
- Never use names longer than 15 characters.
- Never give the same name to two different sequences in your set. Although some programs accept it, others (such as ClustalW) don't.

If you don't obey these naming rules, some multiple-sequence-alignment programs may automatically change the name of your sequences, without the courtesy of telling you.

Consejos sobre cómo hacer un AMS



When you select your sequences, the general rule is that you want them to be as distantly related as possible — without requiring too many gaps in order to be properly aligned. These two criteria are mutually exclusive, so finding the right trade-off requires a bit of strategy. The following steps show you a general approach that you should have in mind when gathering your sequences:

1. **Select a few sequences.**

   See the section "Gathering your sequences with online BLAST servers," later in this chapter.

2. **Compute a multiple alignment by using one of the servers we introduce in this chapter.**

3. **Evaluate the quality of your alignment visually.**

4. **If your alignment looks good, keep the sequences.**

   A good alignment contains nicely conserved blocks separated by regions with insertions and deletions. If you have such an alignment, your sequence set is probably appropriate — and you can try to extend it by adding a few new sequences.

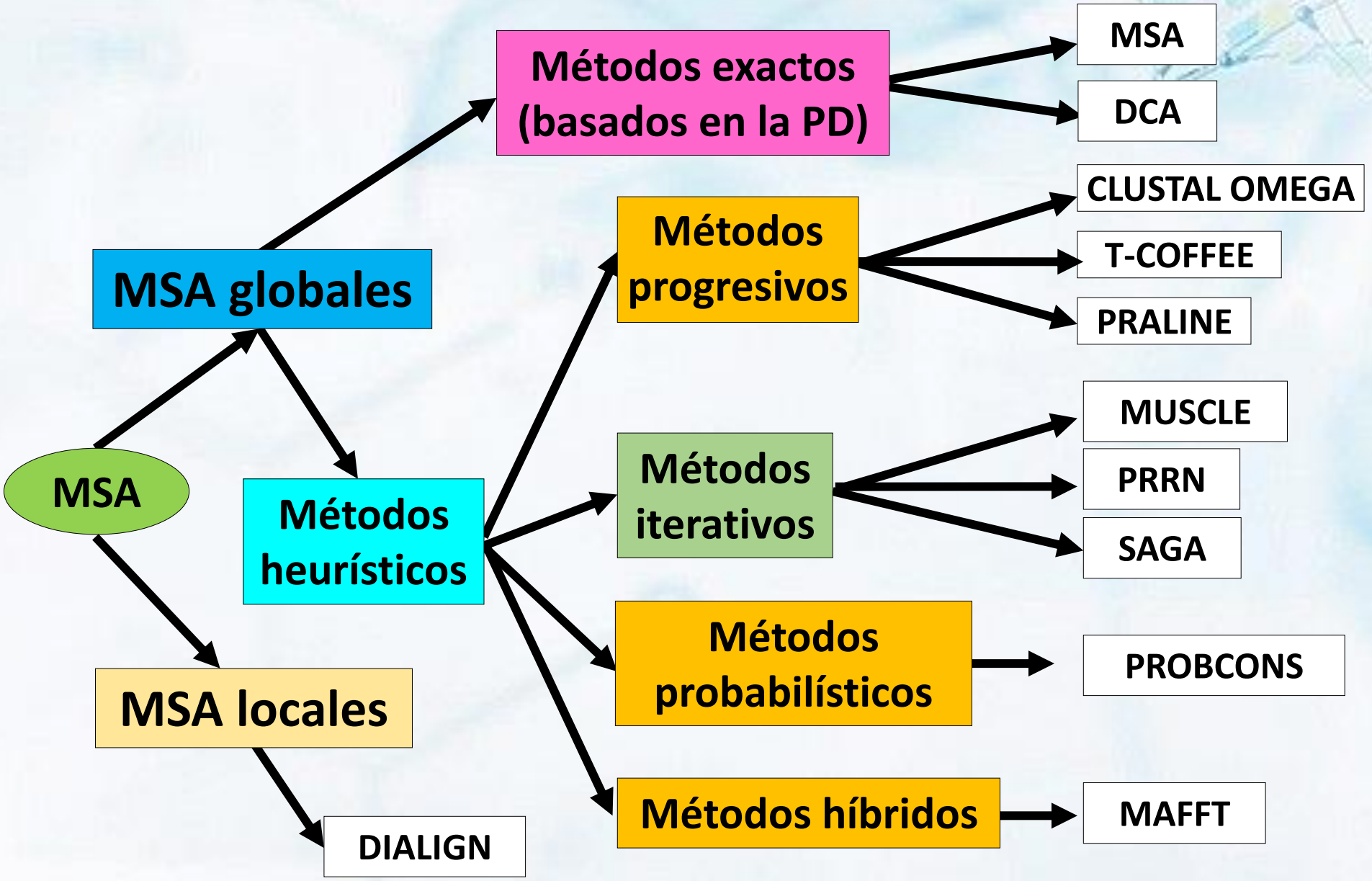
5. **If your alignment is difficult to interpret,**

   a. Examine the sequences more closely — try to remove the trouble-makers that are the most distantly related, or those that cause long insertions/deletions.

   b. Redo the alignment with the smaller set.

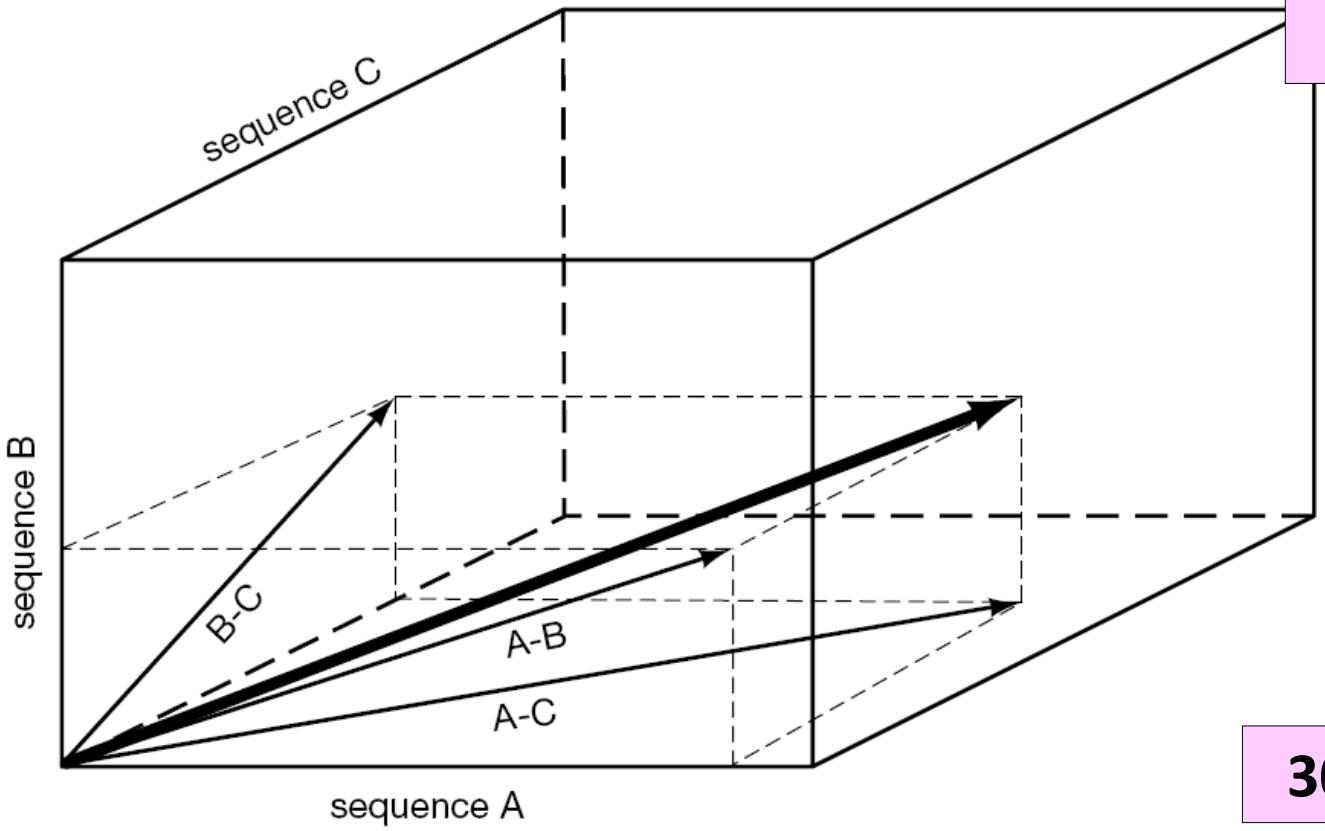
   c. Keep trimming the set until you get a multiple alignment that's easier to interpret.

Métodos para hacer MSA globales

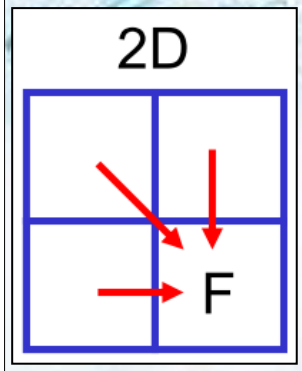


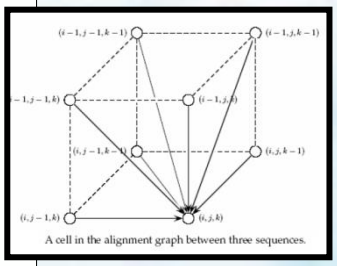
Métodos exactos (basados en la PD)

Métodos exactos (basados en la PD)



$300^2 = 9 \times 10^4$





$300^3 = 2{,}7 \times 10^7$

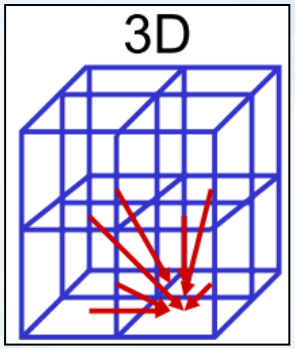


Figure 4.2. Alignment of three sequences by dynamic programming. Arrows on the surfaces of the cube indicate the direction for filling in the scoring matrix for pairs of sequences, A with B, etc., performed as previously described. The alignment of all three sequences requires filling in the lattice of the cube space with optimal alignment scores following the same algorithm. The best score at each interior position requires a consideration of all possible moves within the cube up to that point in the alignment. The trace-back matrix will align positions in all three sequences including gaps.

El espacio de búsqueda con 3 secuencias (PD)

Métodos exactos (basados en la PD)

Para alinear dos secuencias de 300 aa, el algoritmo de programación dinámica (PD) utiliza una matriz bidimensional: El número de operaciones que hay que realizar es 300^2.

Para alinear N secuencias de 300 aa, el algoritmo de PD utiliza una matriz n-dimensional: El número de operaciones que hay que realizar es 300^N.

Este método necesita una gran cantidad de recursos computacionales y mucho tiempo. En la práctica apenas se utiliza: sólo si n = 3 ó, si las secuencias son cortas (entre 200 y 300 residuos), n = 7.

El algoritmo de programación dinámica

MSA mediante el método progresivo

Métodos exactos (basados en la PD)

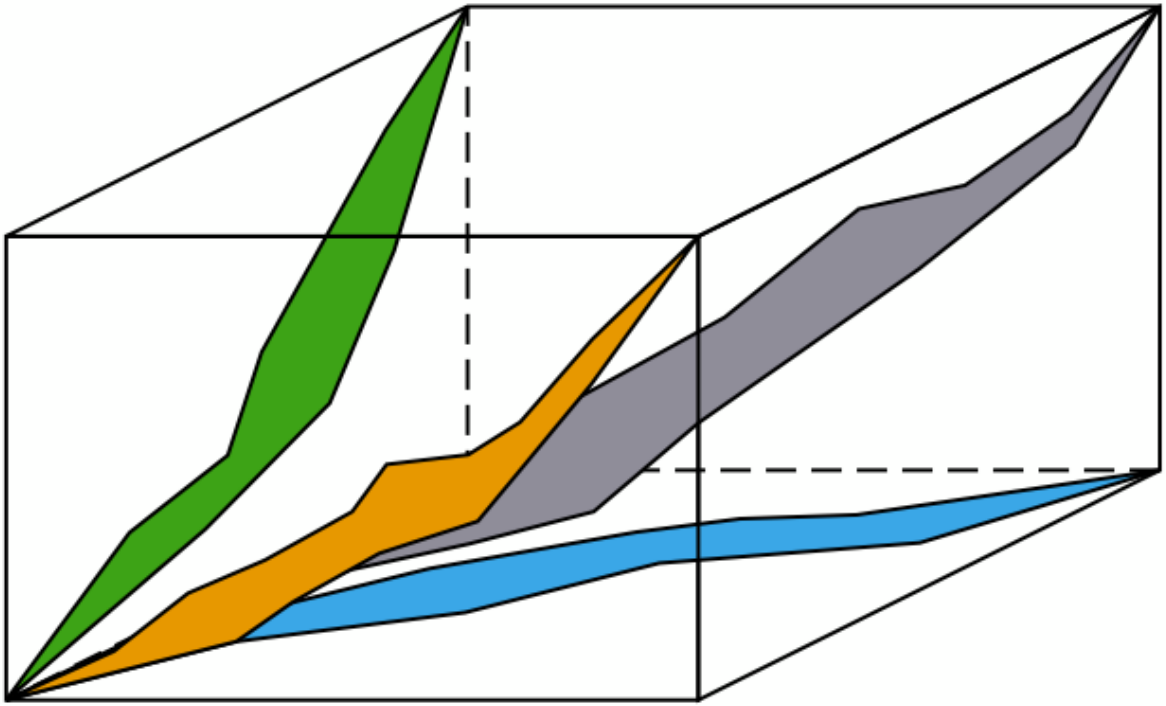


Figure 4.3. Bounds within which an optimal alignment will be found by MSA for three sequences. For MSA to find an optimal alignment among three sequences by the DP algorithm, it is only necessary to calculate optimal alignment scores within the gray volume. This volume is bounded on the one side by the optimal alignments found for each pair of sequences, and on the other by a heuristic multiple alignment of the sequences. The colored areas on each cube surface are two-dimensional projections of the gray volume.

Reducción heurística del espacio de búsqueda

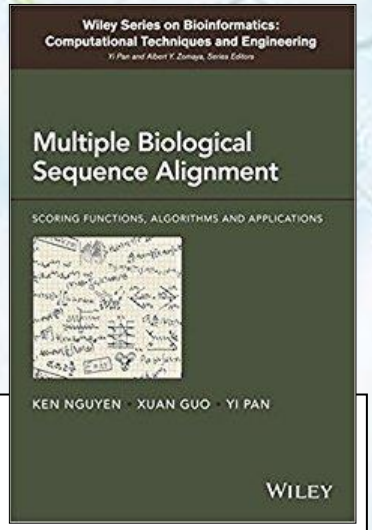
MSA mediante el método progresivo

J Mol Evol (1987) 25:351–360

Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees

Da-Fei Feng and Russell F. Doolittle
Department of Chemistry, University of California–San Diego, La Jolla, California 92093, USA

Journal of Molecular Evolution



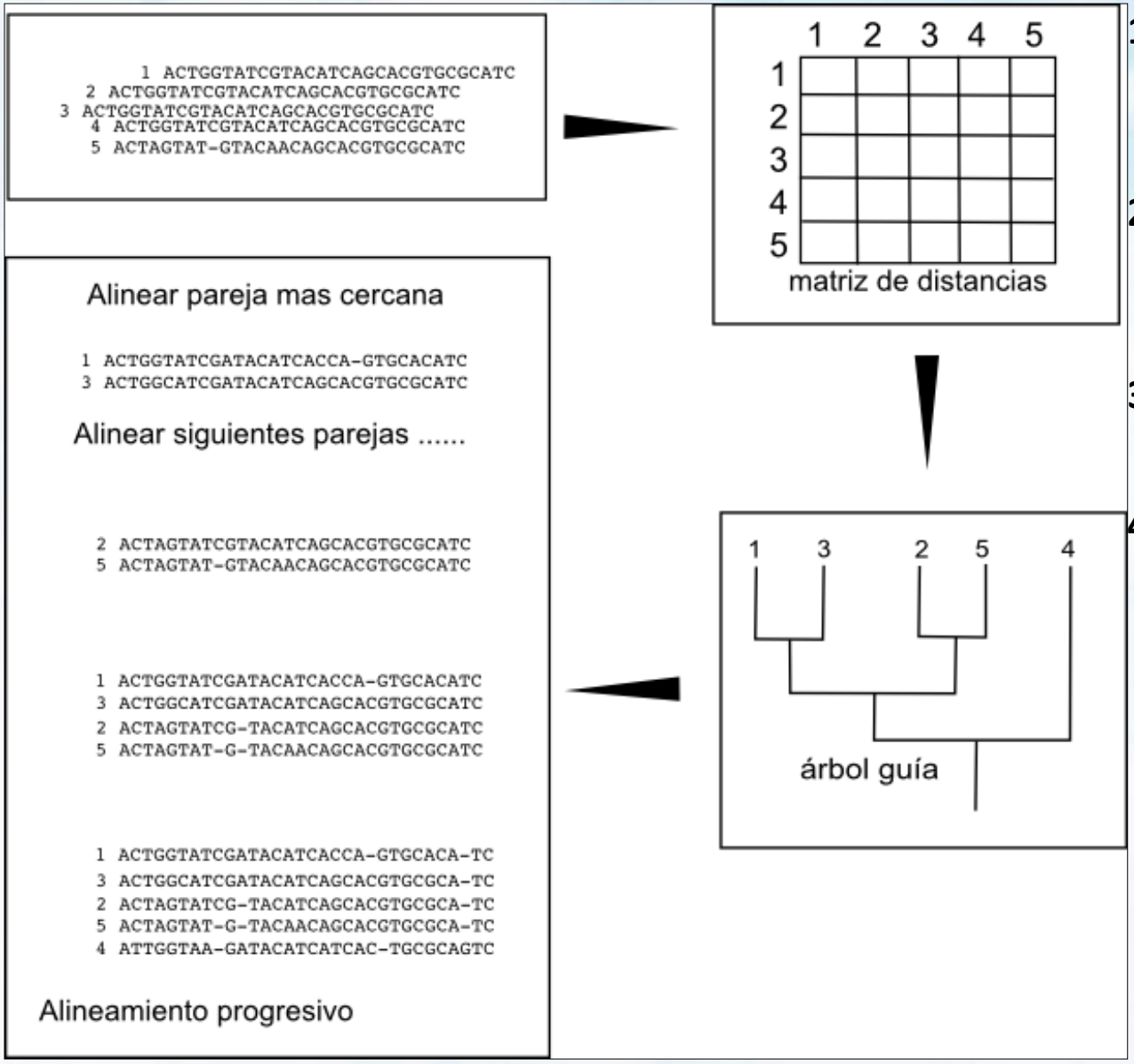
5.2 PROGRESSIVE ALIGNMENT

Progressive MSA is introduced by Feng and Doolittle in 1987 [72]. The main idea is that a pair of sequences with minimum edit distance is most likely to originate from a recently diverged species. Thus, these optimally aligned sequences may review the most reliable hidden information. The algorithm is as follows: (i) calculate pair-wise alignment score and convert them into distances. (ii) Construct a dendrogram, or a guiding tree, from the distances. UPGMA and NJ clustering methods are suitable for this task. (iii) Sequentially align the sequences in their order of addition to the tree. Gap insertions in pair-wise alignments are preserved. A number of alignment programs are based on this technique, such as ClustalW [55, 59], MULTALIN [73], PILEUP, PIMA and KALIGN. The difference between these programs is minimal.

MSA mediante el método progresivo

1. **Compara todas las secuencias de pares utilizando el algoritmo de programación dinámica. Se cuenta el número de residuos y se otorga un score al alineamiento**
2. **Hace un análisis de grupos (*cluster analysis*) para generar una jerarquía de secuencias badas en su similitud. Con esos datos se genera un árbol filogenético (dendograma) que servirá de guía para establecer el orden en la construcción del MSA.**
3. **El MSA comienza alineando las dos secuencias más parecidas. A continuación va añadiendo de forma progresiva las secuencias (o grupos de secuencias) que más se parecen a las que ya están alineadas.**
4. **El MSA depende mucho de los alineamientos iniciales de a pares. Cualquier error cometido en las primeras etapas se va arrastrando durante todo el proceso.**

MSA mediante el método progresivo



1. **Calculo del alineamiento de a pares y asignación de un score**
2. **Conversión de los scores de a pares a una matriz de distancia**
3. **Construcción de un dendograma o árbol guía**
4. **Se seleccionan las secuencias más parecidas y se comienza el alineamiento de las secuencias según el orden del árbol guía**

MSA mediante el método progresivo

Ventajas e inconvenientes del método progresivo

The advantage of progressive alignment algorithms is their capability of aligning a large number of sequences. However, they do not give optimal solution since sequences are iteratively aligned in pair-wise following the order of a guiding tree. Errors made in early stage of alignment will be propagated through the final result.

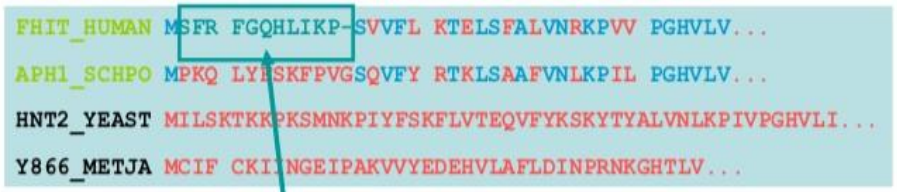
limits of progressive alignment

- initial pairwise alignment
- the very first sequences to be aligned are the most closely related in the tree
  - if they align well, there will be few errors
  - the more distantly related the more errors
- choice of suitable scoring matrices and gap penalties

when to use progressive alignment?

- for more closely related sequences
- large number of sequences

MSA mediante el método progresivo



Alinear las dos secuencias más cercanas

El alineamiento genera un consenso que se utiliza para alinear las secuencias que quedan.

Desde el punto de vista del alineamiento del primer par, el gap puede insertarse en cualquier lugar

Alinear las dos secuencias más cercanas

Una vez insertado el gap no se puede mover porque es parte del consenso.

MSA mediante el método progresivo

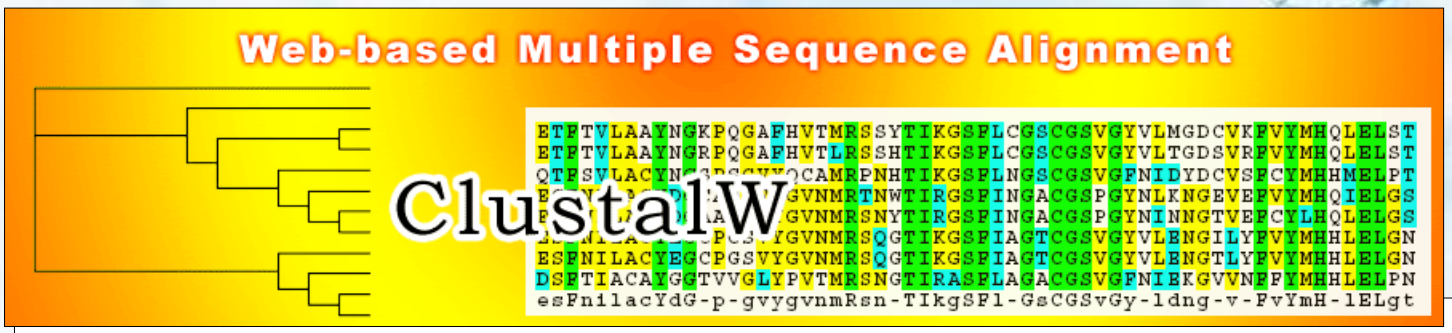
Alinear la secuencia siguiente

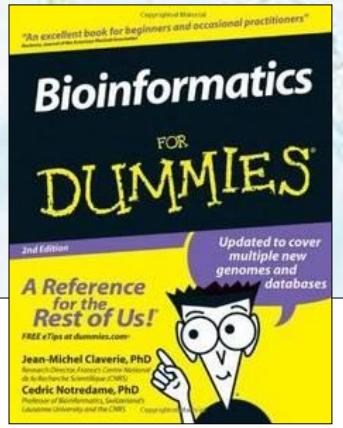
(usa el "consenso" de guía cuando alinea la siguiente secuencia)

Con suerte, el resultado llegue a ser *similar* al resultado que obtenido por un veradero método de alineamiento múltiple.

HEURÍSTICA

Debido al orden de los alineamientos, la posición del gap no puede cambiarse para alinear estas dos Prolinas (lo cual hubiera resultado en un score mayor.





Using ClustalW

ClustalW is by far the most commonly used program for making multiple sequence alignments. If you see a multiple alignment in a scientific publication, you can safely bet that the authors used ClustalW to generate it.

ClustalW uses a progressive method to build its alignments. Instead of aligning all the sequences at the same time, it adds them one by one. If you want to get best results from ClustalW, understanding its underlying principle helps a lot.

Many ClustalW servers are around. They usually run the same version of this program, but their interfaces give you access to different options. At the end of this chapter, we give you a list of servers that run ClustalW. Shopping around to find a ClustalW server that's both fast *and* reliable is worth your time.

ClustalW es el algoritmo progresivo más utilizado

El algoritmo de ClustalW

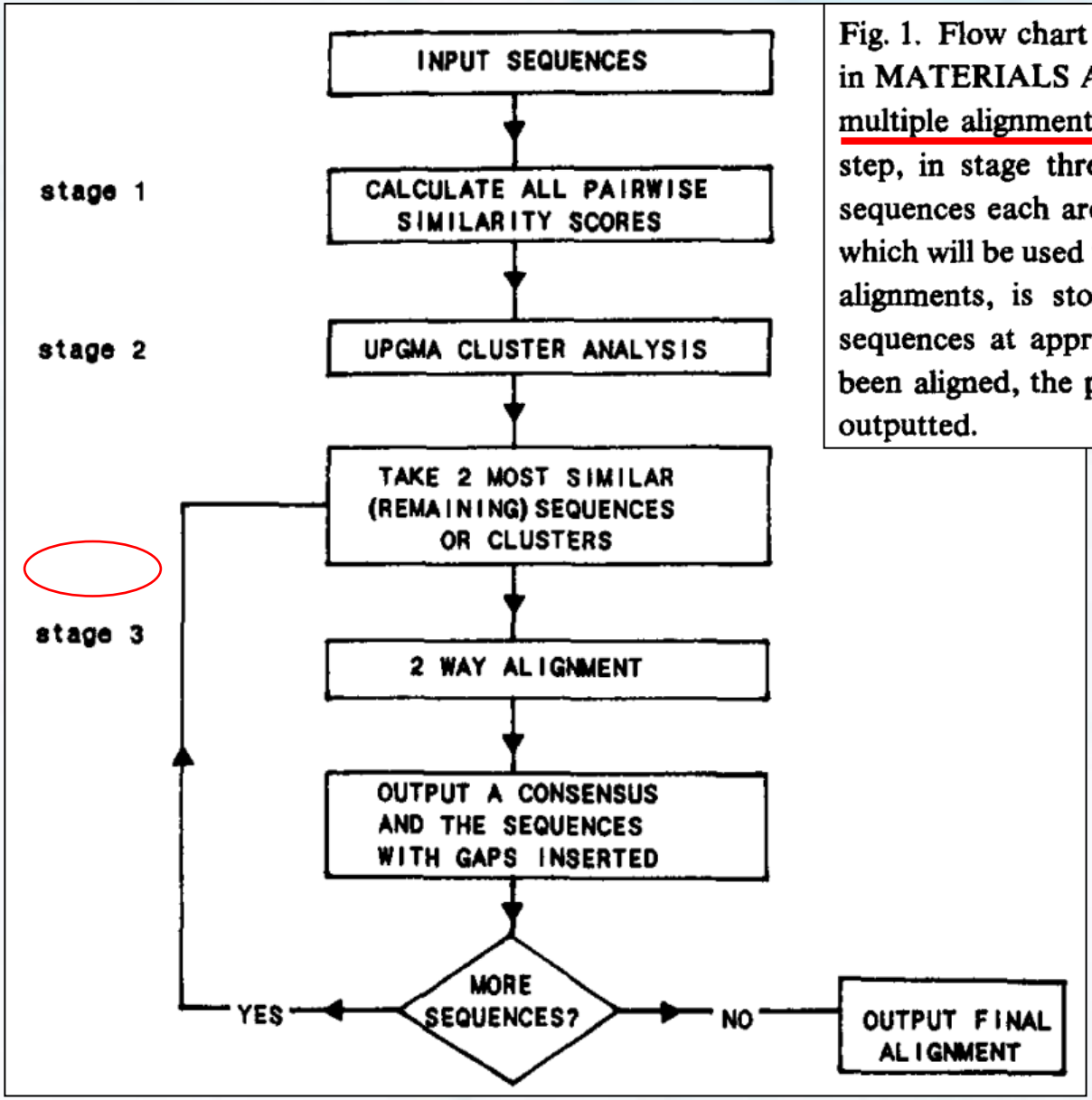


Fig. 1. Flow chart of the multiple alignment strategy described in MATERIALS AND METHODS, section **b**. The core of the multiple alignment process takes place in stage three. At each step, in stage three, two clusters consisting of one or more sequences each are aligned. After each alignment a consensus, which will be used to represent the two aligned clusters in future alignments, is stored and gaps are inserted in the original sequences at appropriate positions. When all sequences have been aligned, the process is complete and the full alignment is outputted.

ClustalW: etapa nº 1

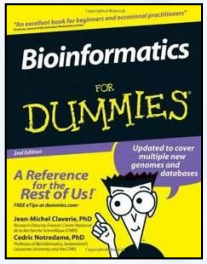
Pairwise alignment: Calculate distance matrix

These scores are calculated as the number of identities in the best alignment divided by the number of residues compared (gap positions are excluded). Both of these scores are initially calculated as per cent identity scores and are converted to distances by dividing by 100 and subtracting from 1.0 to give number of differences per site. We do not correct for multiple substitutions in these initial distances.

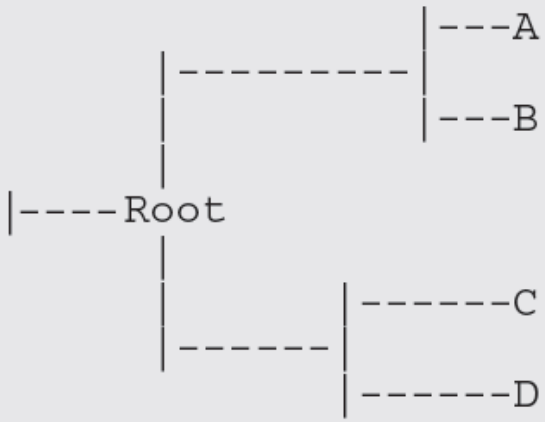
Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6

Se utiliza el algoritmo de programación dinámica para calcular la distancia genética entre cada pareja de secuencias

ClustalW: etapa nº 2



This clustering is named a *dendrogram.* If we were to align four sequences A, B, C, and D, the dendrogram might look like this:



The topology of this dendrogram tells us a simple story: It says that A and B are more similar to each other than they are to C and D. Thus if we align A with B, we are less likely to make a mistake than if we align A with C or D.

To make the progressive alignment, ClustalW follows the dendrogram topology: It starts aligning A with B. After this it aligns C with D. When this is done, Clustal has two small multiple alignments (AB and CD). This is where Clustal pulls out its main trick: It aligns the two alignments as if each of them was a single sequence! It is not as complicated as it seems and there are many ways to do this. For instance, you could replace each alignment with a single consensus sequence. Clustal uses a slightly more sophisticated method, but the idea is essentially the same: It treats multiple alignments like single sequences and aligns them two by two.

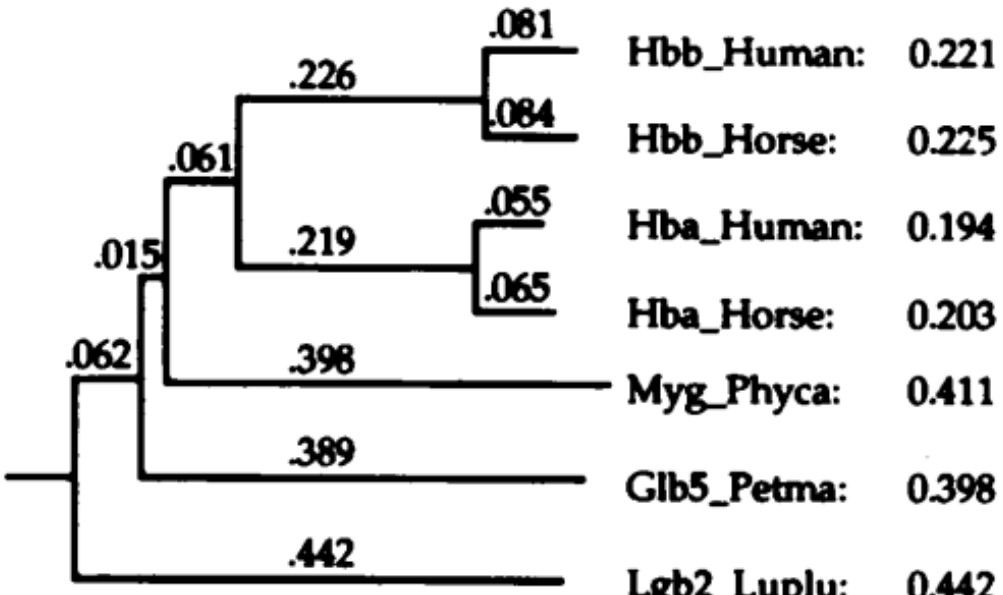
ClustalW agrupa las secuencias y genera un dendrograma

ClustalW: etapa nº 2

Sequence weighting

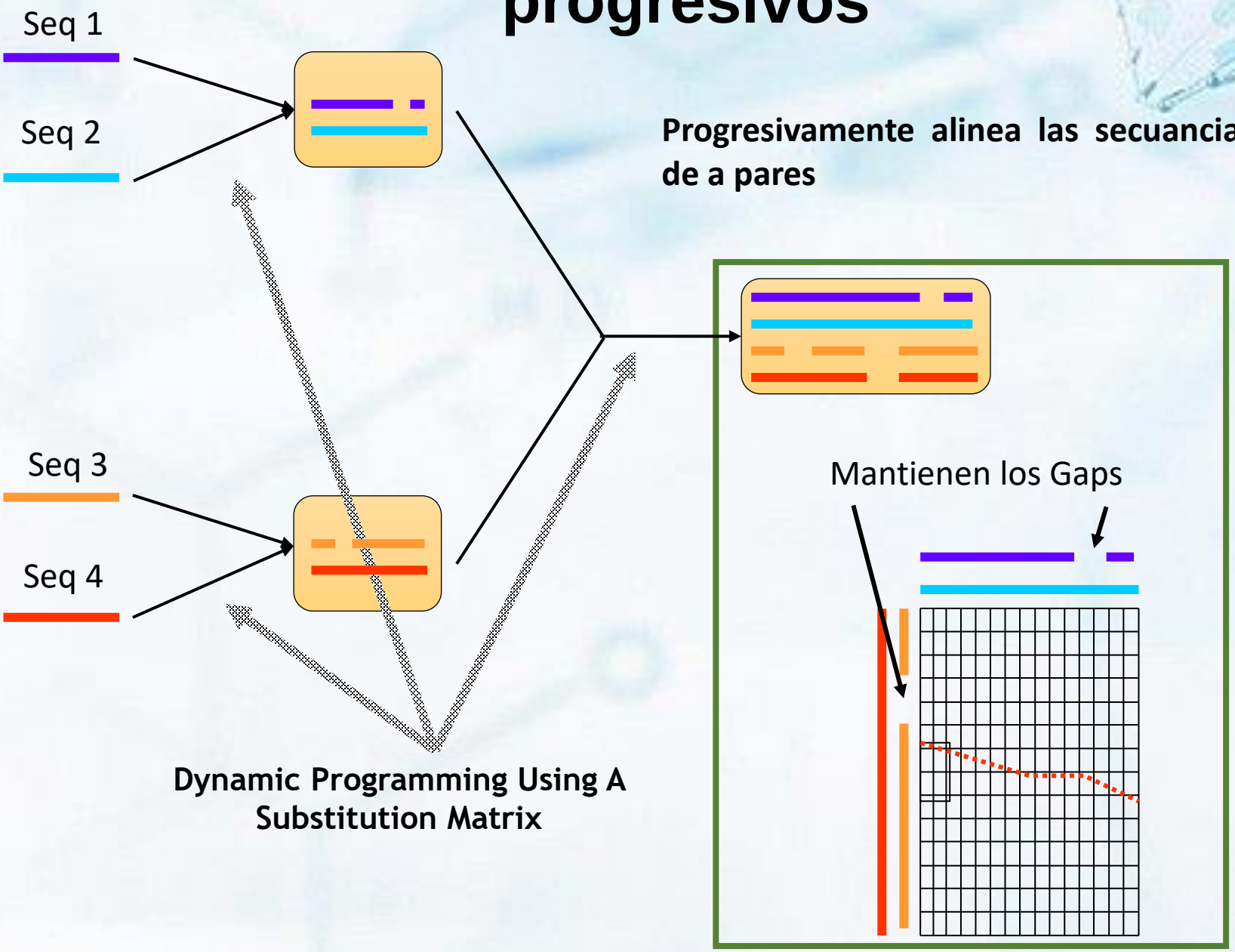
Sequence weights are calculated directly from the guide tree. The weights are normalised such that the biggest one is set to 1.0 and the rest are all less than 1.0. Groups of closely related sequences receive lowered weights because they contain much duplicated information. Highly divergent sequences without any close relatives receive high weights.

Rooted NJ tree (guide tree) and sequence weights

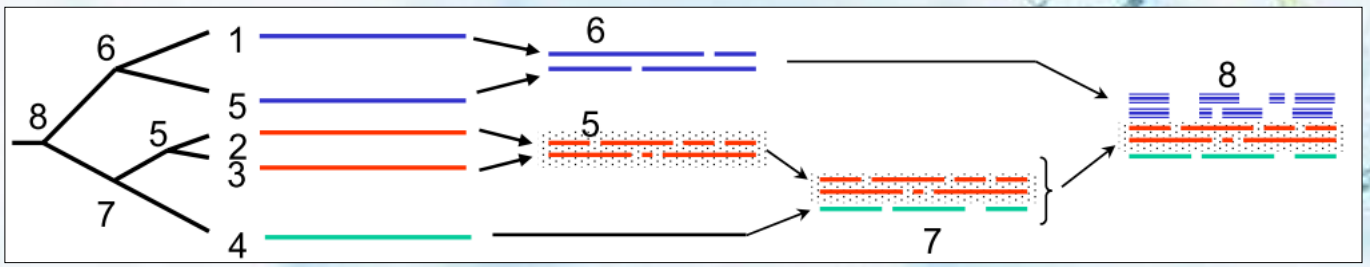


Con esos datos se construye un "árbol guía" que sirve para decidir el orden de los alineamientos (no tiene que ser especialmente preciso).

ClustalW: etapa nº 3: Alineamientos progresivos



ClustalW: etapa nº 3



Progressive alignment: Align following the guide tree

```
--------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
--------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS--
---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLS--
---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
--------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
          *.      .   *                  *  .   *

PDAVMGNPKVKAHGKKVLGAFSDGLAHLD-----NLKGTFATLSELHCDKLHVDPENFRL
PGAVMGNPKVKAHGKKVLHSFGEGVHHLD-----NLKGTFAALSELHCDKLHVDPENFRL
----HGSAQVKGHGKKVADALTNAVAHVD-----DMPNALSALSDLHAHKLRVDPVNFKL
----HGSAQVKAHGKKVGDALTLAVGHLD-----DLPGALSNLSDLHAHKLRVDPVNFKL
EAEMKASEDLKKHGVTVLTALGAILKKKG-----HHEAELKPLAQSHATKHKIPIKYLEF
ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV
          .. *   .                  *   *     .
```

Se empieza alineando las dos secuencias más parecidas. A este alineamiento se le van añadiendo secuencias o alineamientos por orden decreciente de similitud.

La herramienta y los programas Clustal

http://www.clustal.org/clustal2/

Command-line version — **Graphical version**

Clustal W / Clustal X

Multiple alignment of nucleic acid and protein sequences

Home | Webservers | Download | Documentation | Contact | News

Webservers

You don't necessarily have to go through the hassle to install Clustal on your computer. Instead, you can run Clustal online on several servers on the web:

- EBI web server
- Swiss Institute of Bioinformatics

Download Clustal W/X

Clustal 2 comes in two flavors: the command-line version Clustal W and the graphical version Clustal X. Precompiled executables for Linux, Mac OS X and Windows (incl. XP and Vista) of the most recent version (currently 2.1) along with the source code are available for download here. You can also browse for older versions (Clustal W 1.81, Clustal V etc).
The current version of Clustal 2 is also mirrored on the EBI ftp site .

Clustal 2.1 is released under the GNU Lesser GPL.

Documentation

Help

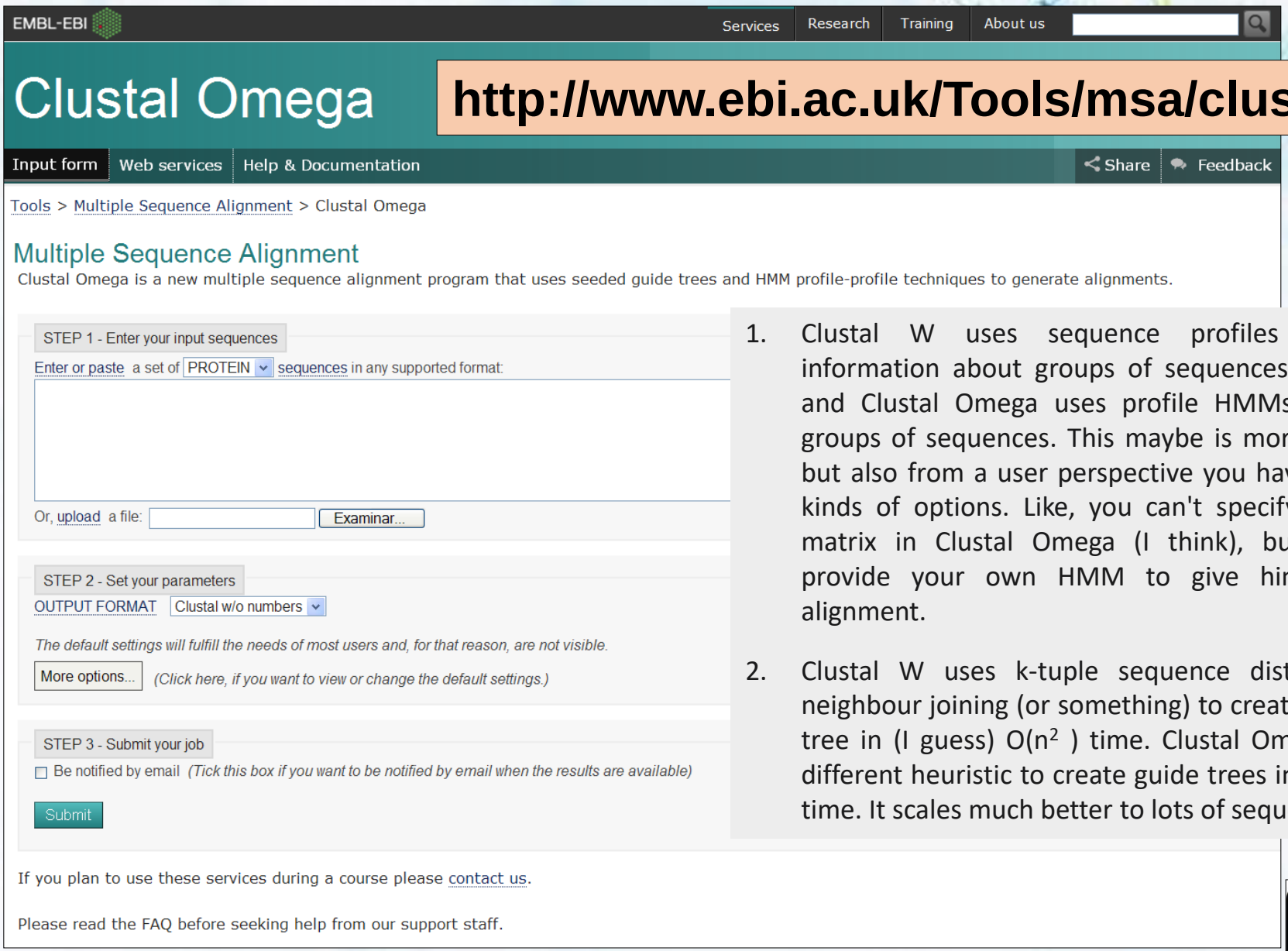
Please have a look at Clustal X's builtin help menu or if you
of date) can be found here: Clustal W and Clustal X. There
structures for profile alignment is "Using Clustal X for multi

El servidor del EBI ha retirado ClustalW2 y en su lugar ofrece Clustal Omega

References

- **Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23, 2947-2948.**
- Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD.

Página principal de Clustal Omega (EBI)



1. Clustal W uses sequence profiles to store information about groups of sequences (probably) and Clustal Omega uses profile HMMs to model groups of sequences. This maybe is more accurate, but also from a user perspective you have different kinds of options. Like, you can't specify a scoring matrix in Clustal Omega (I think), but you can provide your own HMM to give hints to the alignment.

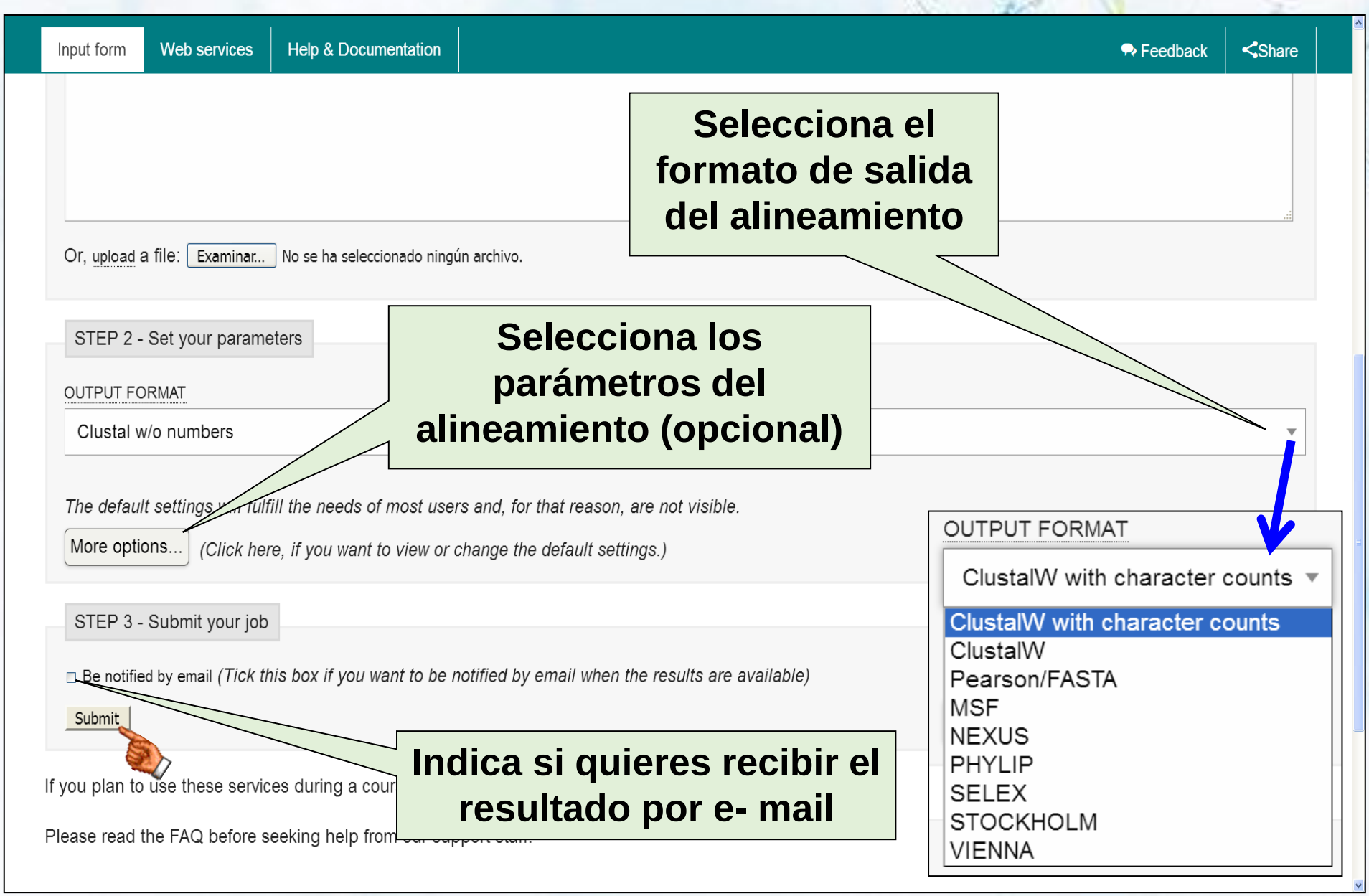
2. Clustal W uses k-tuple sequence distances and neighbour joining (or something) to create the guide tree in (I guess) $O(n^2)$ time. Clustal Omega uses a different heuristic to create guide trees in O(n log n) time. It scales much better to lots of sequences.

CÓMO SE HACE UN AMS CON CLUSTAL OMEGA

- 1.- https://www.ebi.ac.uk/Tools/msa/clustalo/
- 2.- Selecciona el tipo de secuencias a alinear
- 3.- Introduce las secuencias en el campo
- 4.- Selecciona el formato de salida para los resultados
- 5.- Selecciona los parámetros del alineamiento (optativo)
- 6.- Indica si quieres recibir los resultados por e-mail
- 7.- Submit

Introduce las secuencias

Selecciona los parámetros del alineamiento y envíalo



Clustal Omega está trabajando …

EMBL-EBI | Services | Research | Training | Industry | About us | EMBL-EBI Hinxton

Clustal Omega

Input form | Web services | Help & Documentation | Feedback | Share

Tools > Multiple Sequence Alignment > Clustal Omega

Service Retirement

We remind you that it is not long until the EBI's Wise2DBA and Promoterwise services are retired on 15th April 2018. Alternatives can be found at Exonerate, BWA or BLAT. If you have any concerns, please contact us via support.

Your job is currently running... please be patient

The result of your job will appear in this browser window.

Job ID: clustalo-I20180416-094233-0863-35400310-p2m

Please note the following

- You may press Shift+Refresh or Reload on your browser at any time to check if results are ready.
- You may bookmark this page to view your results later if you wish.
- Results are stored for 7 days.

EMBL-EBI

Services
By topic
By name (A-Z)
Help & Support

Research
Publications
Research groups
Postdocs & PhDs

Training
Train at EBI
Train outside EBI
Train online
Contact organisers

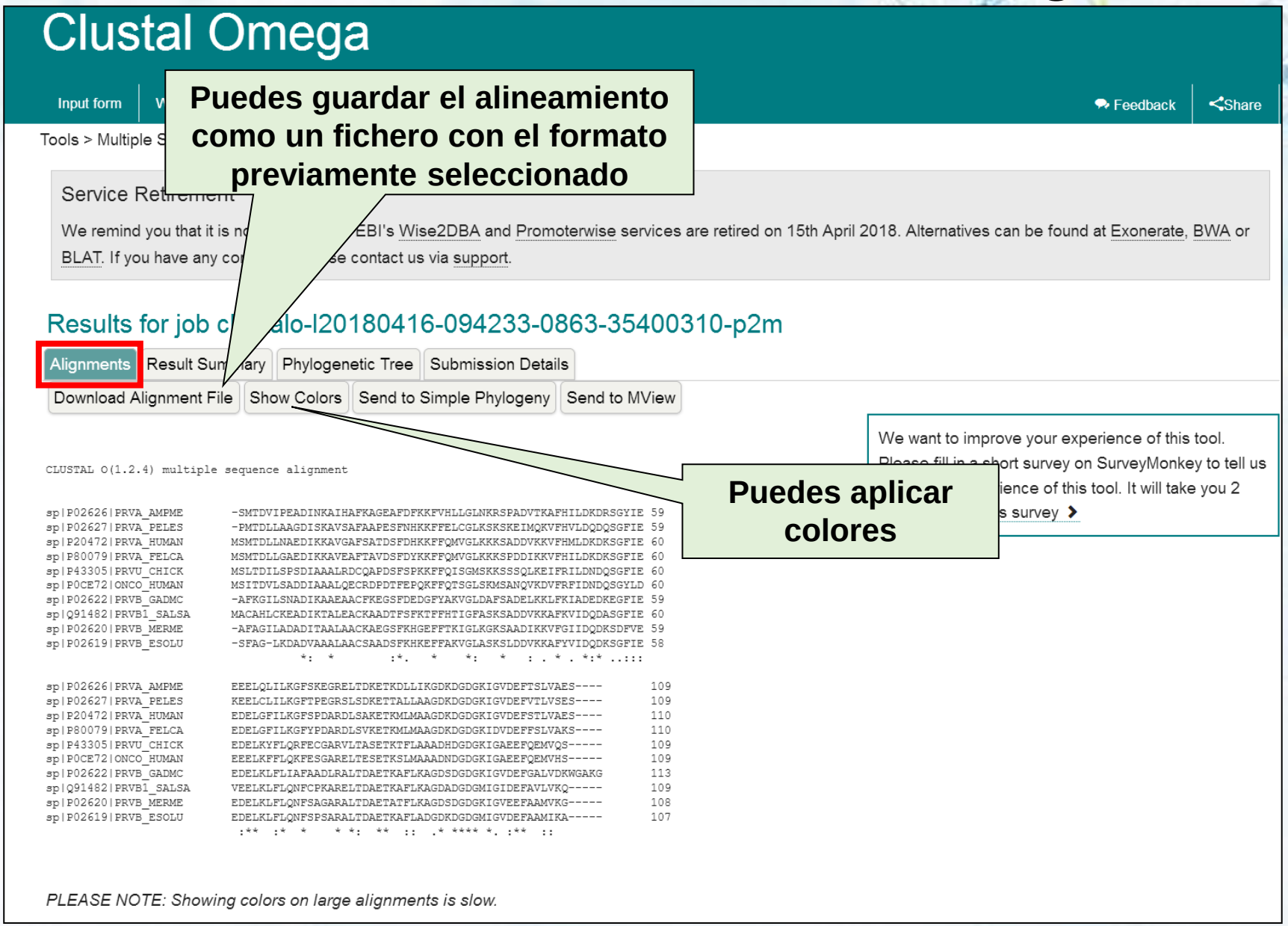
Industry
Members Area
Workshops
SME Forum
Contact Industry programme

About EMBL-EBI
Contact us
Events
Jobs
News
People & groups

EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. +44 (0)1223 49 44 44

Intranet

Resultados de Clustal Omega



Puedes guardar el alineamiento como un fichero con el formato previamente seleccionado

Puedes aplicar colores

Resultados de Clustal Omega

EMBL-EBI | Services | Research | Training | Industry | About us

Clustal Omega

Input form | Web services | Help & Documentation

Tools > Multiple Sequence Alignment > Clustal Omega

Resumen de los resultados

We remind you ... until the EBI's WISEZDBA and Promoterwise services are ... BLAT. If you have ... please contact us via support.

Results for job c...talo-I20180416-094233-0863-3540031...

Alignments | Result Summary | Phylogenetic Tree | Submission Details

Download Alignment File | Hide Colors | Send to Simple Phylogeny | Send to MView

```
CLUSTAL O(1.2.4) multiple sequence alignment


sp|P02626|PRVA_AMPME       -SMTDVIPEADINKAIHAFKAGEAFDFKKFVHLLGLNKRSPADVTKAFHILDKDRSGYIE 59
sp|P02627|PRVA_PELES       -PMTDLLAAGDISKAVSAFAAPESFNHKKFFELCGLKSKSKEIMQKVFHVLDQDQSGFIE 59
sp|P20472|PRVA_HUMAN       MSMTDLLNAEDIKKAVGAFSATDSFDHKKFFQMVGLKKKSADDVKKVFHMLDKDKSGFIE 60
sp|P80079|PRVA_FELCA       MSMTDLLGAEDIKKAVEAFTAVDSFDYKKFFQMVGLKKKSPDDIKKVFHILDKDKSGFIE 60
sp|P43305|PRVU_CHICK       MSLTDILSPSDIAAALRDCQAPDSFSPKKFFQISGMSKKSSSQLKEIFRILDNDQSGFIE 60
sp|P0CE72|ONCO_HUMAN       MSITDVLSADDIAAALQECRDPDTFEPQKFFQTSGLSKMSANQVKDVFRFIDNDQSGYLD 60
sp|P02622|PRVB_GADMC       -AFKGILSNADIKAAEAACFKEGSFDEDGFYAKVGLDAFSADELKKLFKIADEDKEGFIE 59
sp|Q91482|PRVB1_SALSA      MACAHLCKEADIKTALEACKAADTFSFKTFFHTIGFASKSADDVKKAFKVIDQDASGFIE 60
sp|P02620|PRVB_MERME       -AFAGILADADITAALAACKAEGSFKHGEFFTKIGLKGKSAADIKKVFGIIDQDKSDFVE 59
sp|P02619|PRVB_ESOLU       -SFAG-LKDADVAAALAACSAADSFKHKEFFAKVGLASKSLDDVKKAFYVIDQDKSGFIE 58
                                *:  *       :*.   *    *:   *   : . * . *:* ..:::


sp|P02626|PRVA_AMPME       EEELQLILKGFSKEGRELTDKETKDLLIKGDKDGDGKIGVDEFTSLVAES----       109
sp|P02627|PRVA_PELES       KEELCLILKGFTPEGRSLSDKETTALLAAGDKDGDGKIGVDEFVTLVSES----       109
sp|P20472|PRVA_HUMAN       EDELGFILKGFSPDARDLSAKETKMLMAAGDKDGDGKIGVDEFSTLVAES----       110
sp|P80079|PRVA_FELCA       EDELGFILKGFYPDARDLSVKETKMLMAAGDKDGDGKIDVDEFFSLVAKS----       110
sp|P43305|PRVU_CHICK       EDELKYFLQRFECGARVLTASETKTFLAAADHDGDGKIGAEEFQEMVQS-----       109
sp|P0CE72|ONCO_HUMAN       EEELKFFLQKFESGARELTESETKSLMAAADNDGDGKIGAEEFQEMVHS-----       109
sp|P02622|PRVB_GADMC       EDELKLFLIAFAADLRALTDAETKAFLKAGDSDGDGKIGVDEFGALVDKWGAKG       113
sp|Q91482|PRVB1_SALSA      VEELKLFLQNFCPKARELTDAETKAFLKAGDADGDGMIGIDEFAVLVKQ-----       109
sp|P02620|PRVB_MERME       EDELKLFLQNFSAGARALTDAETATFLKAGDSDGDGKIGVEEFAAMVKG-----       108
sp|P02619|PRVB_ESOLU       EDELKLFLQNFSPSARALTDAETKAFLADGDKDGDGMIGVDEFAAMIKA-----       107
                            :**  :*  *    * *:  **  ::  .* **** *. :**  ::

sp|P02619|PRVB_ESOLU       EDELKLFLQNFSPSARALTDAETKAFLADGDKDGDGMIGVDEFAAMIKA-----       107
                            :**  :*  *    * *:  **  ::  .* **** *. :**  ::
```

What do the colours mean when I show them on protein alignments?

This protein-only option colours the residues according to their physicochemical properties:

Residue	Colour	Property
AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RK	MAGENTA	Basic - H
STYHCNGQ	GREEN	Hydroxyl + sulfhydryl + amine + G
Others	Grey	Unusual amino/imino acids etc

Código de colores

We want to imp... perience of this tool. Pleas... about ... minutes. Take this survey ›

Resultados de Clustal Omega

Puedes utilizar Jalview, una herramienta que permite visualizar y/o editar el alineamiento

Clustal Omega

Input form | Web services | Help & Documentation

Tools > Multiple Sequence Alignment > Clustal Omega

Service Retirement

We remind you that it is not long until the EBI's Wise2DBA and ...rwise services BLAT. If you have any concerns, please contact us via support.

Results for job clustalo-I20180416-094233-0...-354003...

Alignments | Result Summary | Phylogenetic Tree | Submission Details

Input Sequences

clustalo-I20180416-094233-0863-35400310-p2m.input

Tool Output

clustalo-I20180416-094233-0863-35400310-p2m.output

Alignment in CLUSTAL format with base/residue numbering

clustalo-I20180416-094233-0863-35400310-p2m.clustal_num

Phylogenetic Tree

clustalo-I20180416-094233-0863-35400310-p2m.ph

Percent Identity Matrix

clustalo-I20180416-094233-0863-35400310-p2m.pim

Jalview

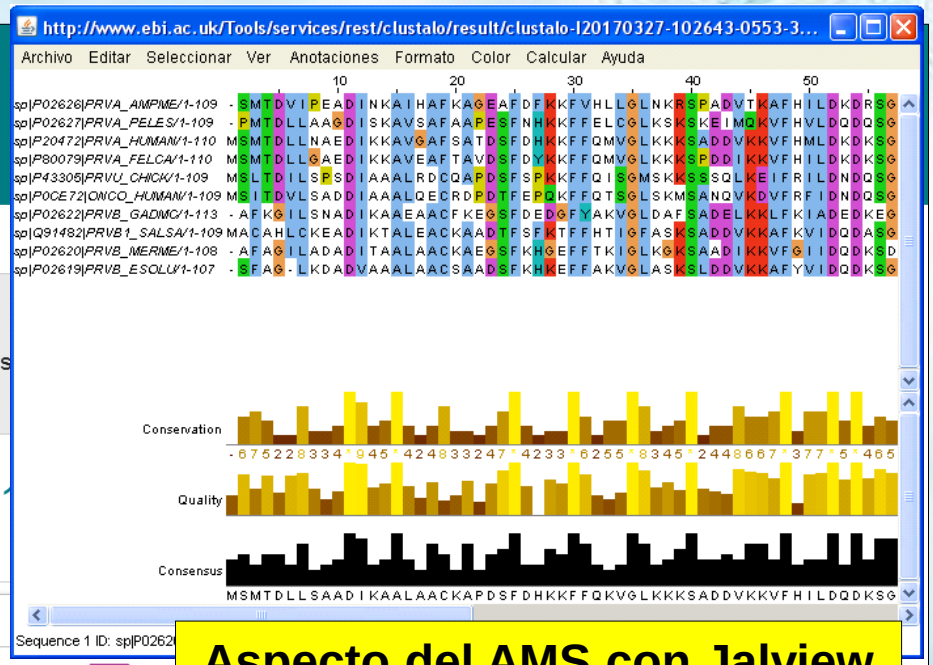
View result with Jalview

Simple Phylogeny

Send to Simple Phylogeny

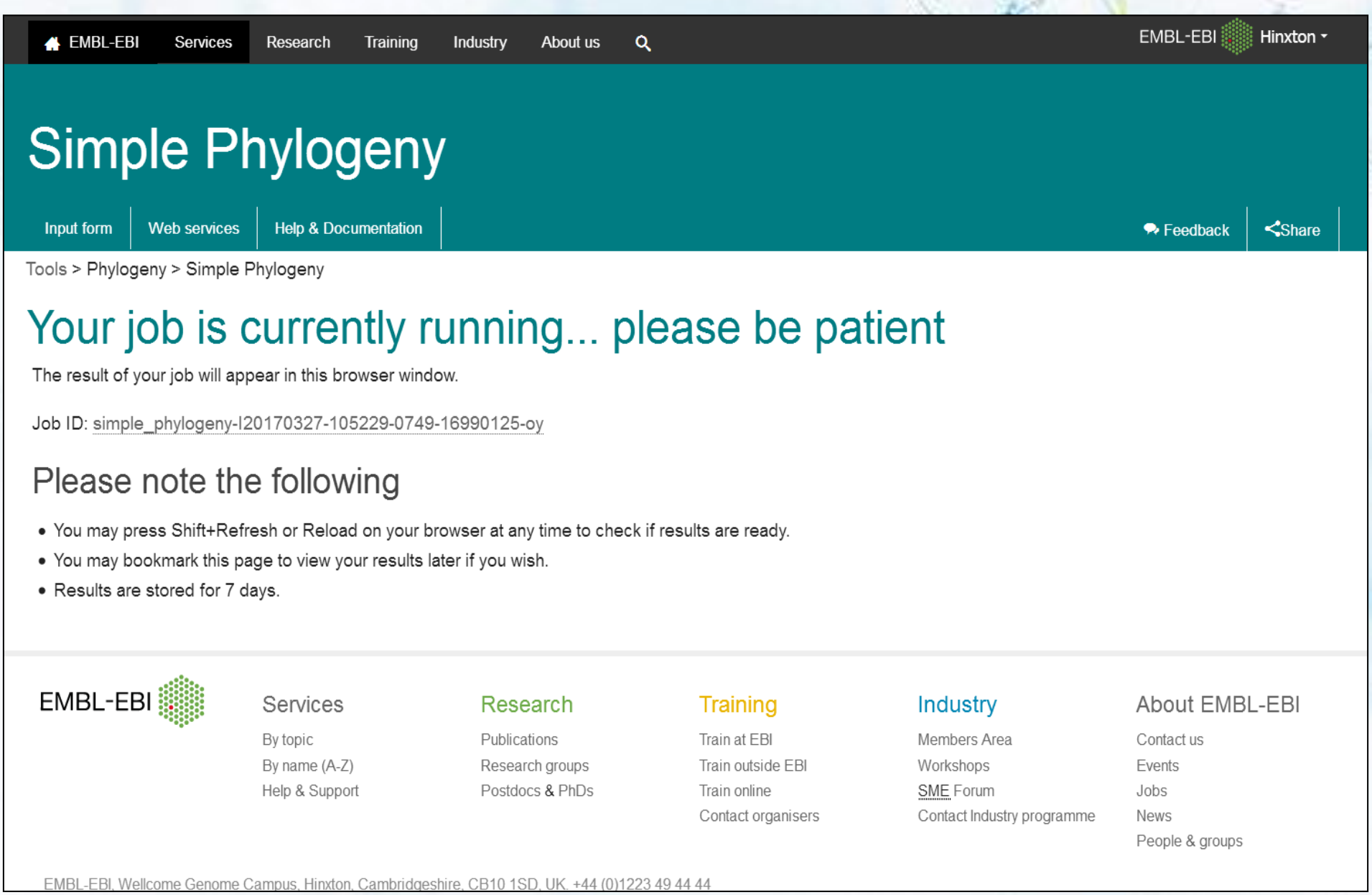
MView

Send to MView



Aspecto del AMS con Jalview

Simple Philogeny está trabajando …



Resultado de Simple Phylogeny

```
Percent Identity  Matrix - created by Clustal2.1

 1: sp|P02626|PRVA_AMPME   100.00   60.55   62.39   61.47   42.59   42.59   44.04   50.00   44.44   47.66
 2: sp|P02627|PRVA_PELES    60.55  100.00   67.89   64.22   45.37   41.67   43.12   49.07   47.22   50.47
 3: sp|P20472|PRVA_HUMAN    62.39   67.89  100.00   87.27   52.29   50.46   49.54   50.46   50.00   57.01
 4: sp|P80079|PRVA_FELCA    61.47   64.22   87.27  100.00   51.38   47.71   49.54   48.62   48.15   52.34
 5: sp|P43305|PRVU_CHICK    42.59   45.37   52.29   51.38  100.00   67.89   47.22   47.71   52.78   52.34
 6: sp|P0CE72|ONCO_HUMAN    42.59   41.67   50.46   47.71   67.89  100.00   39.81   45.87   46.30   45.79
 7: sp|P02622|PRVB_GADMC    44.04   43.12   49.54   49.54   47.22   39.81  100.00   52.78   63.89   59.81
 8: sp|Q91482|PRVB1_SALSA   50.00   49.07   50.46   48.62   47.71   45.87   52.78  100.00   59.26   68.22
 9: sp|P02620|PRVB_MERME    44.44   47.22   50.00   48.15   52.78   46.30   63.89   59.26  100.00   71.03
10: sp|P02619|PRVB_ESOLU    47.66   50.47   57.01   52.34   52.34   45.79   59.81   68.22   71.03  100.00
```

Distancias entre las secuencias

https://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/

Simple Phylogeny

Input form | Web services | Help & Documentation

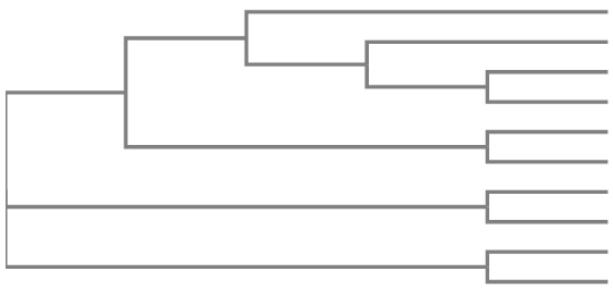
Tools > Phylogeny > Simple Phylogeny

Results for job simple_phylogeny-I20

Phylogenetic Tree | Result Summary | Submission Details

Phylogram

Branch length: Cladogram Real



sp|P02626|PRVA_AMPME 0.20841
sp|P02627|PRVA_PELES 0.18744
sp|P20472|PRVA_HUMAN 0.05325
sp|P80079|PRVA_FELCA 0.07402
sp|P43305|PRVU_CHICK 0.14003
sp|P0CE72|ONCO_HUMAN 0.18107
sp|P02622|PRVB_GADMC 0.21599
sp|P02620|PRVB_MERME 0.14512
sp|Q91482|PRVB1_SALSA 0.18581
sp|P02619|PRVB_ESOLU 0.13195

Formato reconocible por el ordenador, que se puede cortar y pegar en otras herramientas bioinformáticas

The Newick tree format

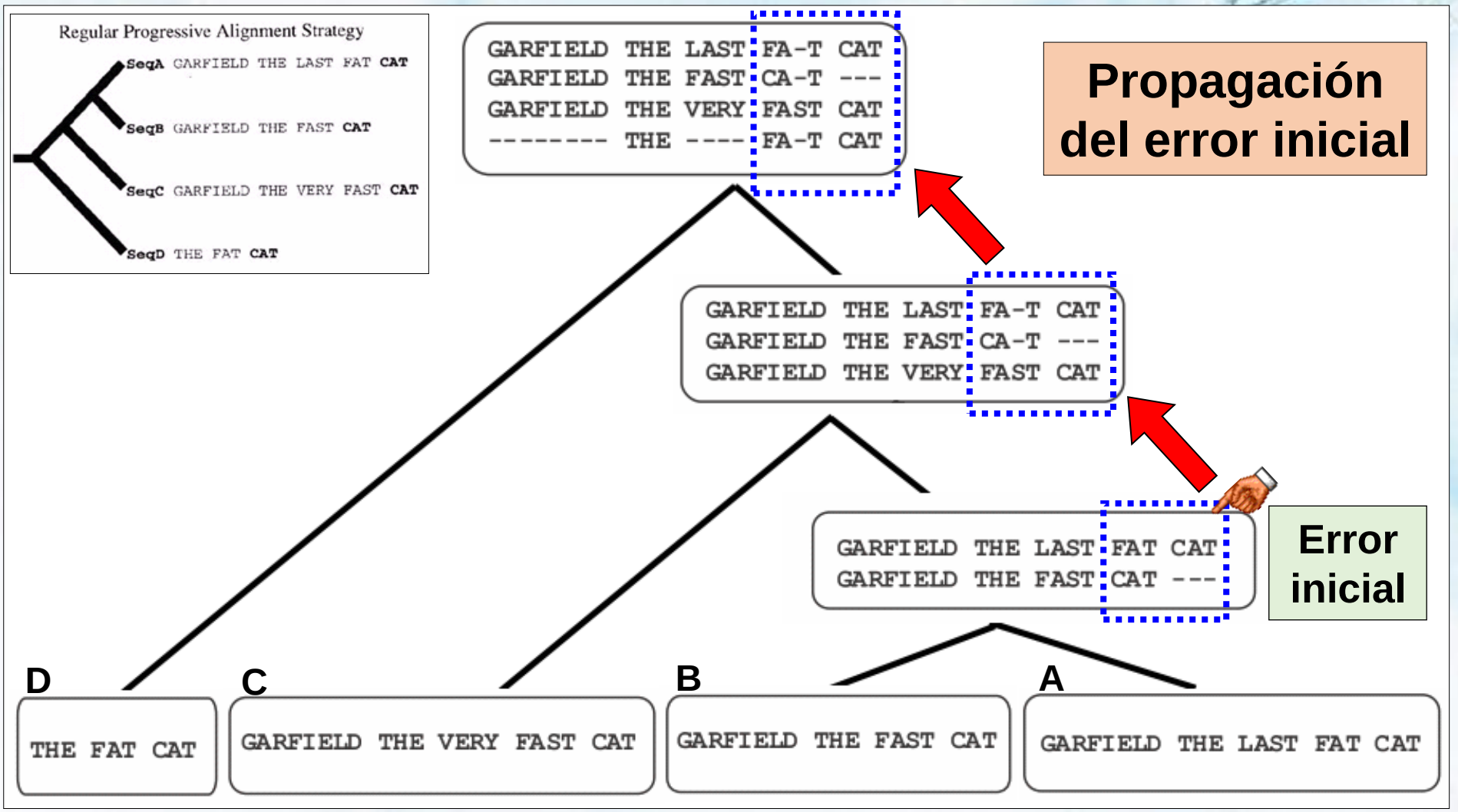
Phylogenetic Tree

View Phylogenetic Tree File

```
(
(
(
sp|P02626|PRVA_AMPME:0.20841,
(
sp|P02627|PRVA_PELES:0.18744,
(
sp|P20472|PRVA_HUMAN:0.05325,
sp|P80079|PRVA_FELCA:0.07402)
:0.08837)
:0.00948)
:0.07931,
(
sp|P43305|PRVU_CHICK:0.14003,
sp|P0CE72|ONCO_HUMAN:0.18107)
:0.11842)
:0.05235,
(
sp|P02622|PRVB_GADMC:0.21599,
sp|P02620|PRVB_MERME:0.14512)
:0.02971,
(
sp|Q91482|PRVB1_SALSA:0.18581,
sp|P02619|PRVB_ESOLU:0.13195)
:0.02366);
```

El gran inconveniente de ClustalW

The major problem with the progressive alignment method described above is that errors in the initial alignments of the most closely related sequences are propagated to the msa.

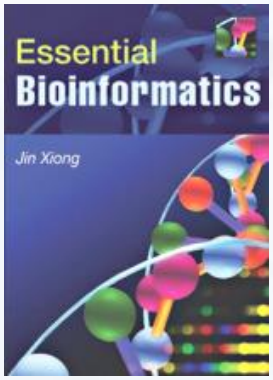


Métodos iterativos

Fundamento de los métodos iterativos

ITERATIVE METHODS OF MULTIPLE SEQUENCE ALIGNMENT

The major problem with the progressive alignment method described above is that errors in the initial alignments of the most closely related sequences are propagated to the msa. This problem is more acute when the starting alignments are between more distantly related sequences. Iterative methods attempt to correct for this problem by repeatedly realigning subgroups of the sequences and then by aligning these subgroups into a global alignment of all of the sequences. The objective is to improve the overall alignment score, such as a sum of pairs score. Selection of these groups may be based on the ordering of the sequences on a phylogenetic tree predicted in a manner similar to that of progressive alignment, separation of one or two of the sequences from the rest, or a random selection of the groups.



Iterative Alignment

The iterative approach is based on the idea that an optimal solution can be found by repeatedly modifying existing suboptimal solutions. The procedure starts by producing a low-quality alignment and gradually improves it by iterative realignment through well-defined procedures until no more improvements in the alignment scores can be achieved. Because the order of the sequences used for alignment is different in each iteration, this method may alleviate the "greedy" problem of the progressive strategy. However, this method is also heuristic in nature and does not have guarantees for finding the optimal alignment.

1. Construcción de un alineamiento multiple por cualquier método rápido
2. Mejora del alineamiento usando diversos métodos matemáticos
3. Se asigna un score a cada uno de los resultados de los alineamientos
4. Las rondas de alineamientos terminan cuando ya no puede mejorarse el Score

JMB

J. Mol. Biol. (1996) **264**, 823–838

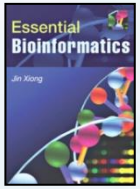
Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinement as Assessed by Reference to Structural Alignments

Osamu Gotoh

Department of Biochemistry Saitama Cancer Center Research Institute, 818 Komuro, Ina-machi, Saitama 362, Japan

The relative performances of four strategies for aligning a large number of protein sequences were assessed by referring to corresponding structural alignments of 54 independent families. Multiple sequence alignment of a family was constructed by a given method from the sequences of known structures and their homologues, and the subset consisting of the sequences of known structures was extracted from the whole alignment and compared with the structural counterpart in a residue-to-residue fashion. Gap-opening and -extension penalties were optimized for each family and method. Each of the four multiple alignment methods gave significantly more accurate alignments than the conventional pairwise method. In addition, a clear difference in performance was detected among three of the four multiple alignment methods examined. The currently most popular progressive method ranked worst among the four, and the randomized iterative strategy that optimizes the sum-of-pairs score ranked next worst. The two best-performing strategies, one of which was newly developed, both pursue an optimal weighted sum-of-pairs score, where the pair weights were introduced to correct for uneven representations of subgroups in a family. The new method uses doubly nested iterations to make alignment, phylogenetic tree and pair weights mutually consistent. Most importantly, the improvement in accuracy of alignments obtained by these iterative methods over pairwise or progressive method tends to increase with decreasing average sequence identity, implying that iterative refinement is more effective for the generally difficult alignment of remotely related sequences. Four well-known amino acid substitution matrices were also tested in combination with the various methods. However, the effects of substitution matrices were found to be minor in the framework of multiple alignment, and the same order of relative performance of the alignment methods was observed with any of the matrices.

PRRN: el artículo original (1996)



El programa PRRN

PRRN (http://prrn.ims.u-tokyo.ac.jp/) is a web-based program that uses a double-nested iterative strategy for multiple alignment. It performs multiple alignment through two sets of iterations: inner iteration and outer iteration. In the *outer iteration*, an initial random alignment is generated that is used to derive a UPGMA tree (see Chapter 11). Weights are subsequently applied to optimize the alignment. In the *inner iteration*, the sequences are randomly divided into two groups. Randomized alignment is used for each group in the initial cycle, after which the alignment positions in each group are fixed. The two groups, each treated as a single sequence, are then aligned to each other using global dynamic programming. The process is repeated through many cycles until the total SP score no longer increases. At this point, the resulting alignment is used to construct a new UPGMA tree. New weights are applied to optimize alignment scores. The newly optimized alignment is subject to further realignment in the inner iteration. This process is repeated over many cycles until there is no further improvement in the overall alignment scores (Fig. 5.4).

Algoritmo del programa PRRN (1)

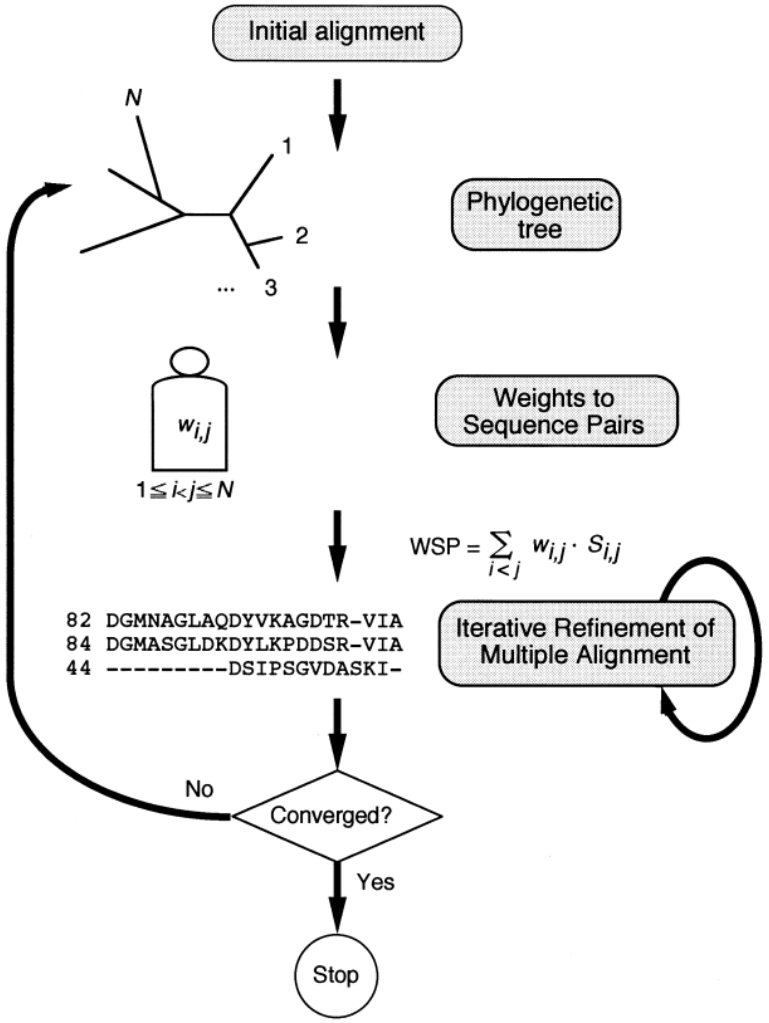


Figure 9. Schematic diagram of the procedures of the doubly nested randomized iterative (DNR) method for multiple sequence alignment. The details of the procedures are explained in the text.

Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinement as Assessed by Reference to Structural Alignments

Osamu Gotoh

J. Mol. Biol. (1996) **264**, 823–838

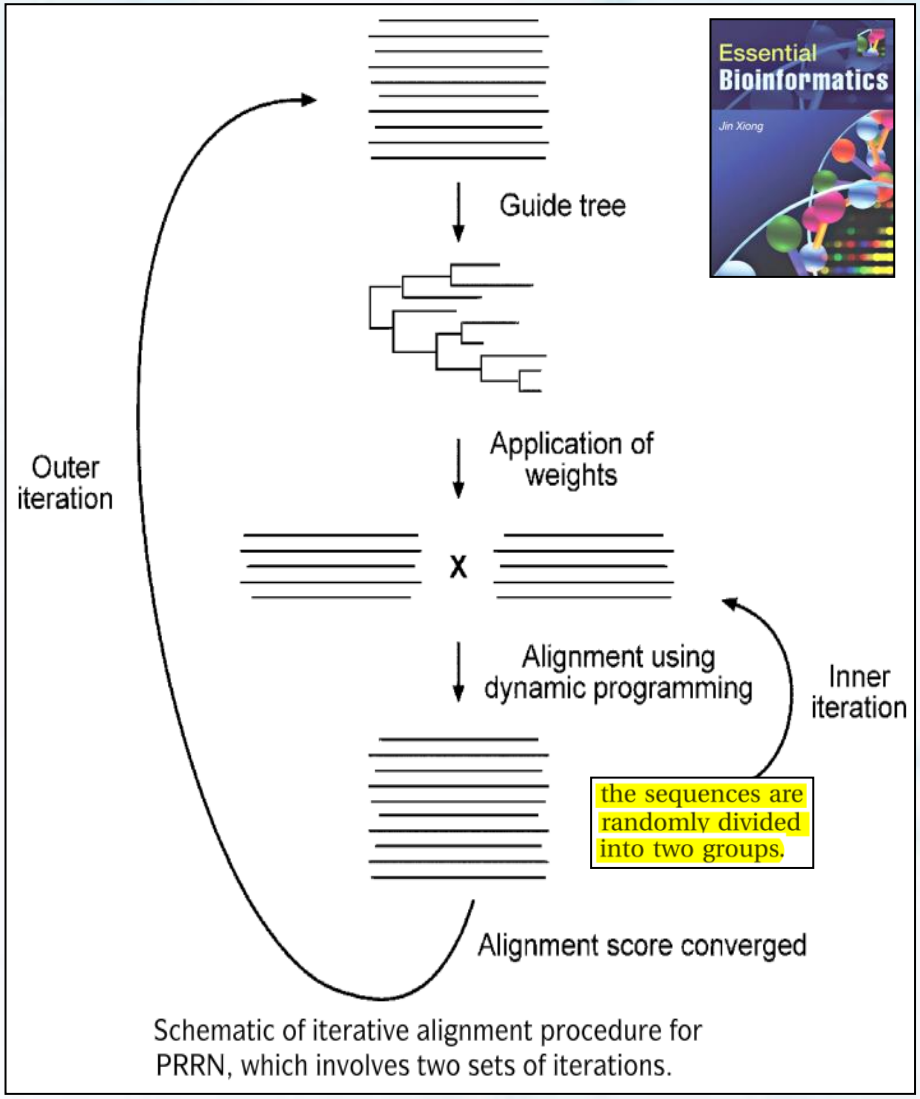
It starts with a preliminary multiple alignment, which may be obtained by any simpler method.

A set of weights assigned to all the pairs of sequences is calculated by the three-way algorithm (Gotoh, 1995) guided by the phylogenetic tree constructed from the distance values between members in the initial alignment.

The first cycle is the same as that of the RIW method but with only a single series of iteration. After the convergence, a new phylogenetic tree and pair weights are calculated, and the second cycle is started. In this way, the doubly nested iterations are continued until no change in the total WSP score is observed.

Algoritmo y página web con la herramienta PRRN

http://www.genome.jp/tools/prrn/



Schematic of iterative alignment procedure for PRRN, which involves two sets of iterations.

Multiple Sequence Alignment by PRRN

CLUSTALW | MAFFT | PRRN

Help

Specify Parameters :

Sequence Type :	PROTEIN / DNA
Score Matrix :	PAM250
Gap Penalty (PROTEIN) :	Open: 9.0, Extend: 2.0, Background: 1.0 (0 .. 100)
Gap Penalty (DNA) :	Open: 5.0, Extend: 2.0, Background: 2.0 (0 .. 100)
Output Format :	Native
Method :	Progressive (pairwise alignment) + Iterative refinement
Window Size :	100 (0 .. 1000)
Score Matrix (Conserved Regions) :	PAM100
Threshold Value (Conserved Regions) :	25.0 (0 .. 1000)

Enter your sequences (FASTA format) :

Restrictions of the sequence size:
1. [The number of sequences] <= 200
2. [The maximum length] <= 2000
3. [The number of sequences] x [The maximum length] <= 50000

File Upload [Examinar...] No se ha seleccionado ningún archivo.
Copy & Paste

[submit] [reset]

KEGG GenomeNet Kyoto University Bioinformatics Center

MUSCLE: MUltiple **S**equence **C**omparison by **L**og- **E**xpectation

MUSCLE: multiple sequence alignment with high accuracy and high throughput

Robert C. Edgar*

195 Roque Moraes Drive, Mill Valley, CA 94941, USA

Received January 19, 2004; Revised January 30, 2004; Accepted February 24, 2004

ABSTRACT

We describe MUSCLE, a new computer program for creating multiple alignments of protein sequences. Elements of the algorithm include fast distance estimation using *k*mer counting, progressive alignment using a new profile function we call the log-expectation score, and refinement using tree-dependent restricted partitioning. The speed and accuracy of MUSCLE are compared with T-Coffee, MAFFT and CLUSTALW on four test sets of reference alignments: BAliBASE, SABmark, SMART and a new benchmark, PREFAB. MUSCLE achieves the highest, or joint highest, rank in accuracy on each of these sets. Without refinement, MUSCLE achieves average accuracy statistically indistinguishable from T-Coffee and MAFFT, and is the fastest of the tested methods for large numbers of sequences, aligning 5000 sequences of average length 350 in 7 min on a current desktop computer. The MUSCLE program, source code and PREFAB test data are freely available at http://www.drive5.com/muscle.

variant on this strategy is used by T-Coffee (5), which aligns profiles by optimizing a score derived from local and global alignments of all pairs of input sequences. Misalignments by progressive methods are sometimes readily apparent (Fig. 1), motivating further processing (refinement). For a recent review of multiple alignment methods, see Notredame (6). Here we describe MUSCLE (multiple sequence comparison by log-expectation), a new computer program for multiple protein sequence alignment.

Se pueden alinear 5000 secuencias de 350 caracteres en 7 minutos con un ordenador personal.

MUSCLE ALGORITHM

Here we give an overview of the algorithm; a more detailed

Distance measures and guide tree estimation

MUSCLE uses two distance measures for a pair of sequences: a *k*mer distance (for an unaligned pair) and the Kimura distance (for an aligned pair). A *k*mer is a contiguous subsequence of length *k*, also known as a word or *k*-tuple.

***Nucleic Acids Research* 2004, 32: 1792-1797**

MUSCLE: el algoritmo

Crunching large datasets with MUSCLE

MUSCLE is an efficient package for making fast, high-quality multiple sequence alignments. MUSCLE is ideal if you want to align several hundred sequences.

MUSCLE permite alinear miles de secuencias

El proceso se lleva a cabo en 3 etapas:

1.- AMS progresivo (borrador)

2.- AMS progresivo (mejorado)

3.- Refinado

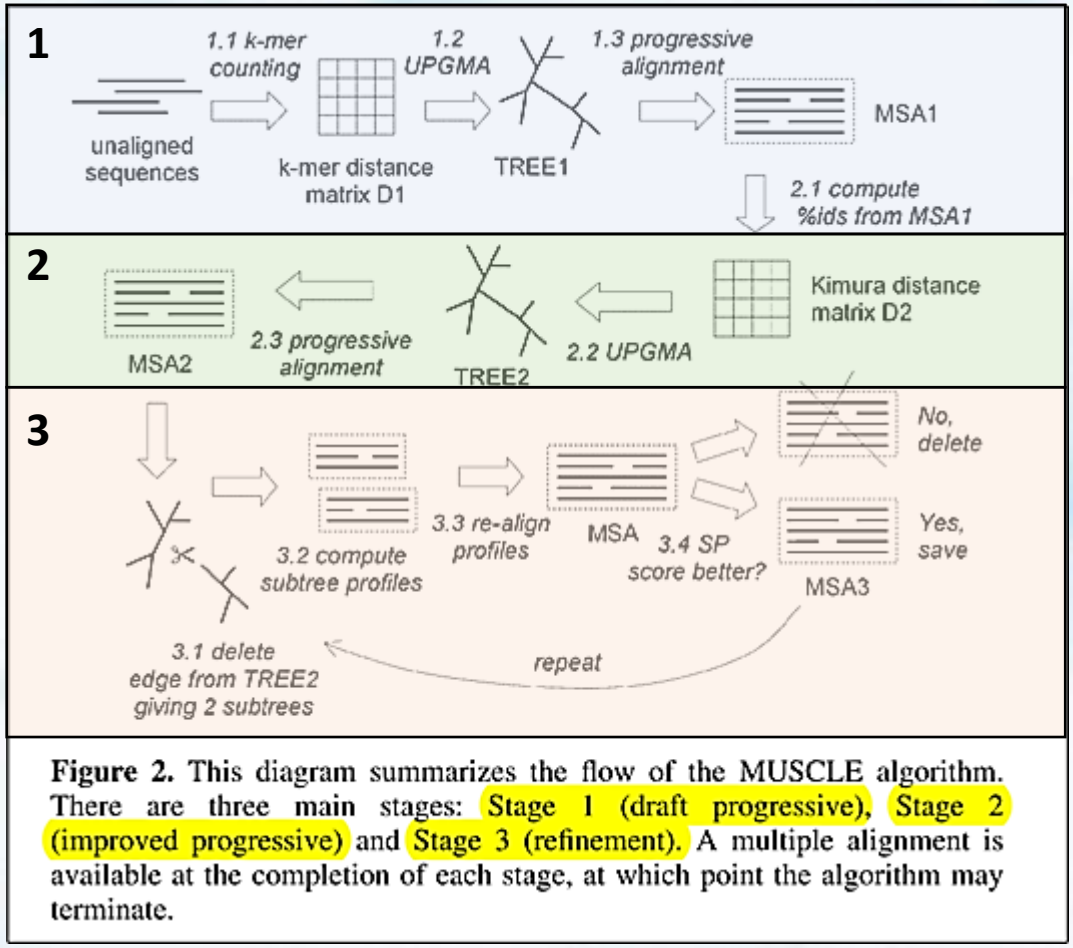


Figure 2. This diagram summarizes the flow of the MUSCLE algorithm. There are three main stages: Stage 1 (draft progressive), Stage 2 (improved progressive) and Stage 3 (refinement). A multiple alignment is available at the completion of each stage, at which point the algorithm may terminate.

El algoritmo de MUSCLE: Etapa 1 (MSA provisional)

Stage 1, Draft progressive. The goal of the first stage is to produce a multiple alignment, emphasizing speed over accuracy.

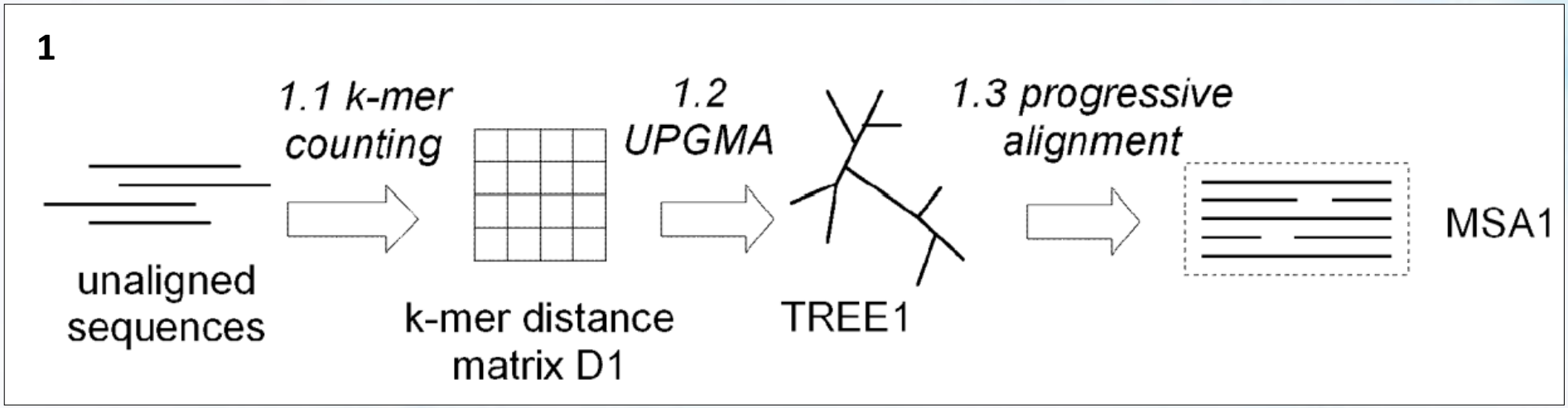
1.1 The *k*mer distance is computed for each pair of input sequences, giving distance matrix D1.

1.2 Matrix D1 is clustered by UPGMA, producing binary tree TREE1.

1.3 A progressive alignment is constructed by following the branching order of TREE1. At each leaf, a profile is constructed from an input sequence. Nodes in the tree are visited in prefix order (children before their parent). At each internal node, a pairwise alignment is constructed of the two child profiles, giving a new profile which is assigned to that node. This produces a multiple alignment of all input sequences, MSA1, at the root.

The *k*mer distance is derived from the fraction of *k*mers in common. This measure does not require an alignment, giving a significant speed advantage.

UPGMA (**U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean)



El algoritmo de MUSCLE: Etapa 2 (MSA mejorado)

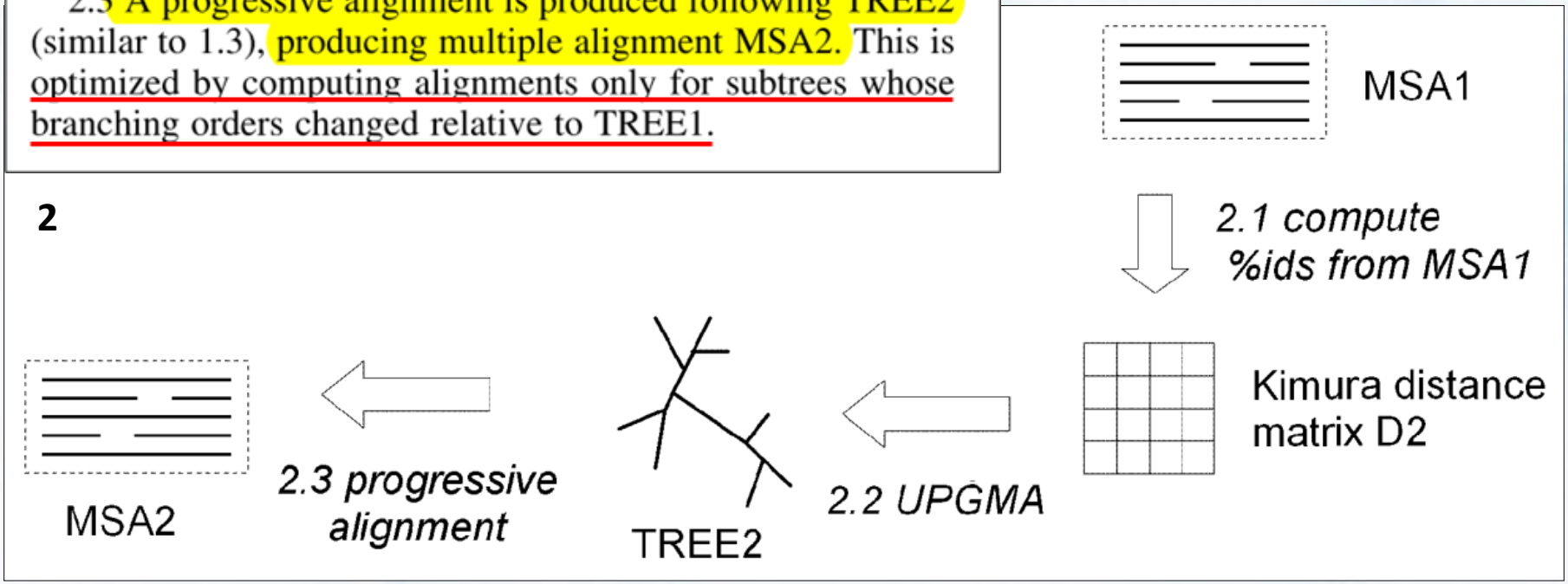
Stage 2, Improved progressive. The main source of error in the draft progressive stage is the approximate *k*mer distance measure, which results in a suboptimal tree. MUSCLE therefore re-estimates the tree using the Kimura distance, which is more accurate but requires an alignment.

2.1 The Kimura distance for each pair of input sequences is computed from MSA1, giving distance matrix D2.

2.2 Matrix D2 is clustered by UPGMA, producing binary tree TREE2.

2.3 A progressive alignment is produced following TREE2 (similar to 1.3), producing multiple alignment MSA2. This is optimized by computing alignments only for subtrees whose branching orders changed relative to TREE1.

The approximate *k*mer distance results in a suboptimal tree. Now is time to improve it.



El algoritmo de MUSCLE: Etapa 3 (Refinado)

Stage 3, Refinement.

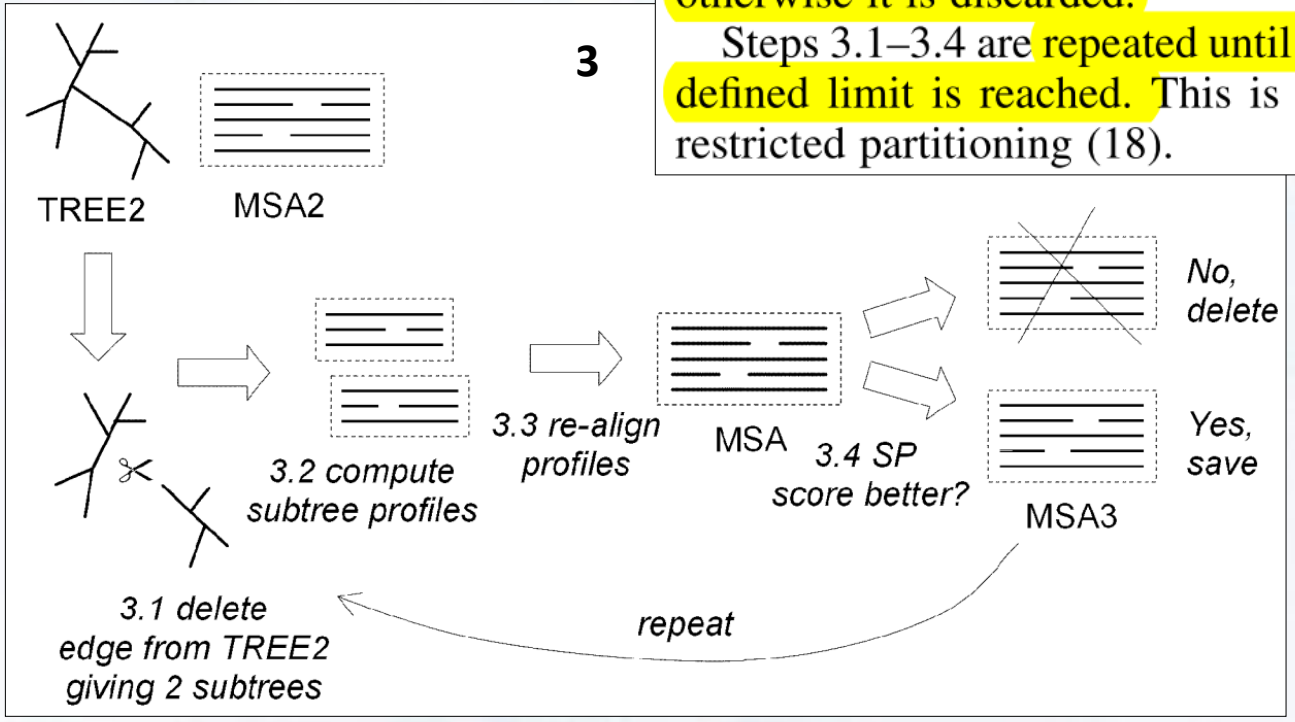
3.1 An edge is chosen from TREE2 (edges are visited in order of decreasing distance from the root).

3.2 TREE2 is divided into two subtrees by deleting the edge. The profile of the multiple alignment in each subtree is computed.

3.3 A new multiple alignment is produced by re-aligning the two profiles.

3.4 If the SP score is improved, the new alignment is kept, otherwise it is discarded.

Steps 3.1–3.4 are repeated until convergence or until a user-defined limit is reached. This is a variant of tree-dependent restricted partitioning (18).



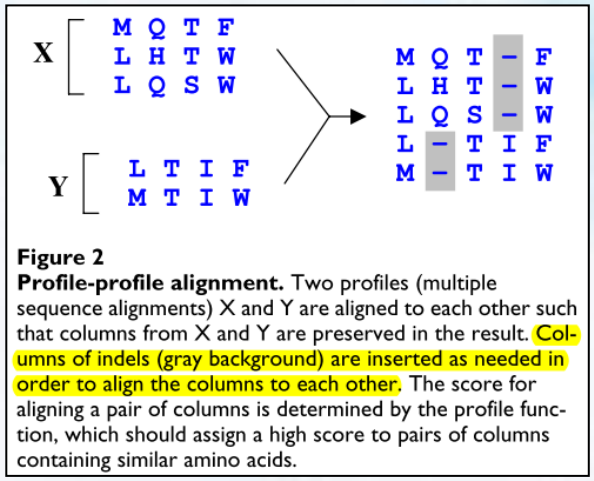
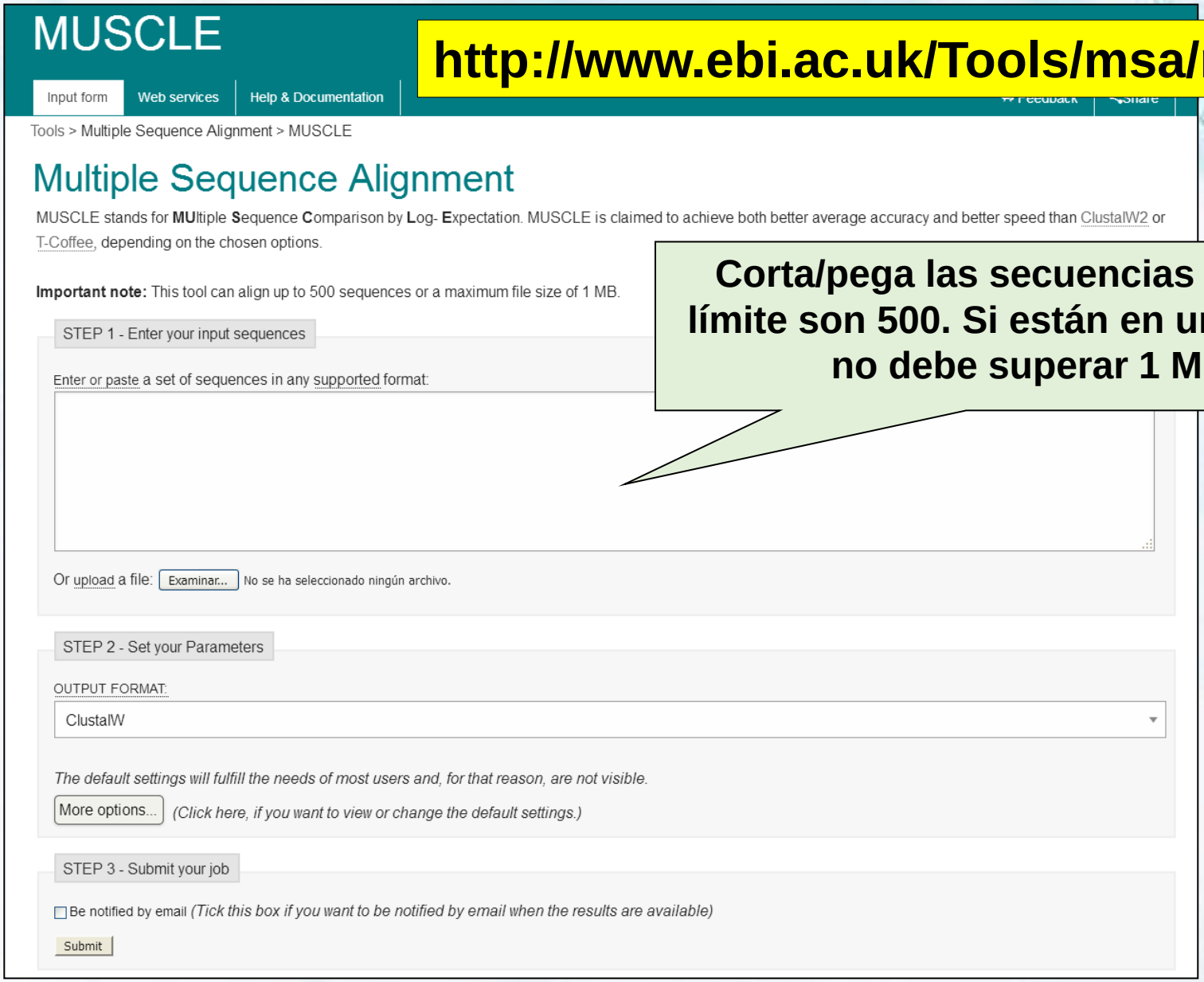


Figure 2
Profile-profile alignment. Two profiles (multiple sequence alignments) X and Y are aligned to each other such that columns from X and Y are preserved in the result. Columns of indels (gray background) are inserted as needed in order to align the columns to each other. The score for aligning a pair of columns is determined by the profile function, which should assign a high score to pairs of columns containing similar amino acids.

La herramienta MUSCLE (EBI)



Home | Software | Services | About | Contact

MUSCLE

http://www.drive5.com/muscle/

MUSCLE has been cited by
29,921 papers
Google scholar
Last updated 08 Apr 2019

Downloads

Documentation

Support

Popular multiple alignment software

MUSCLE is one of the most widely-used methods in biology. On average, MUSCLE is cited by ten new papers every day.

Fast, accurate and easy to use

MUSCLE is one of the best-performing multiple alignment programs according to published benchmark tests, with accuracy and speed that are consistently better than CLUSTALW. MUSCLE can align hundreds of sequences in seconds. Most users learn everything they need to know about MUSCLE in a few minutes—only a handful of command-line options are needed to perform common alignment tasks.

Papers

There are two papers. The first (NAR) introduced the algorithm, and is the primary citation if you use the program. The second (BMC Bioinformatics) gives more technical details, including descriptions of non-default options.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Res.* **32**(5):1792-1797 [Link to PubMed].

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity *BMC Bioinformatics*, (**5**) 113 [Link to PubMed].

El programa MUSCLE (EBI)