Genómica comparativa

Fundamentos y herramientas bioinformáticas para análisis genómicos

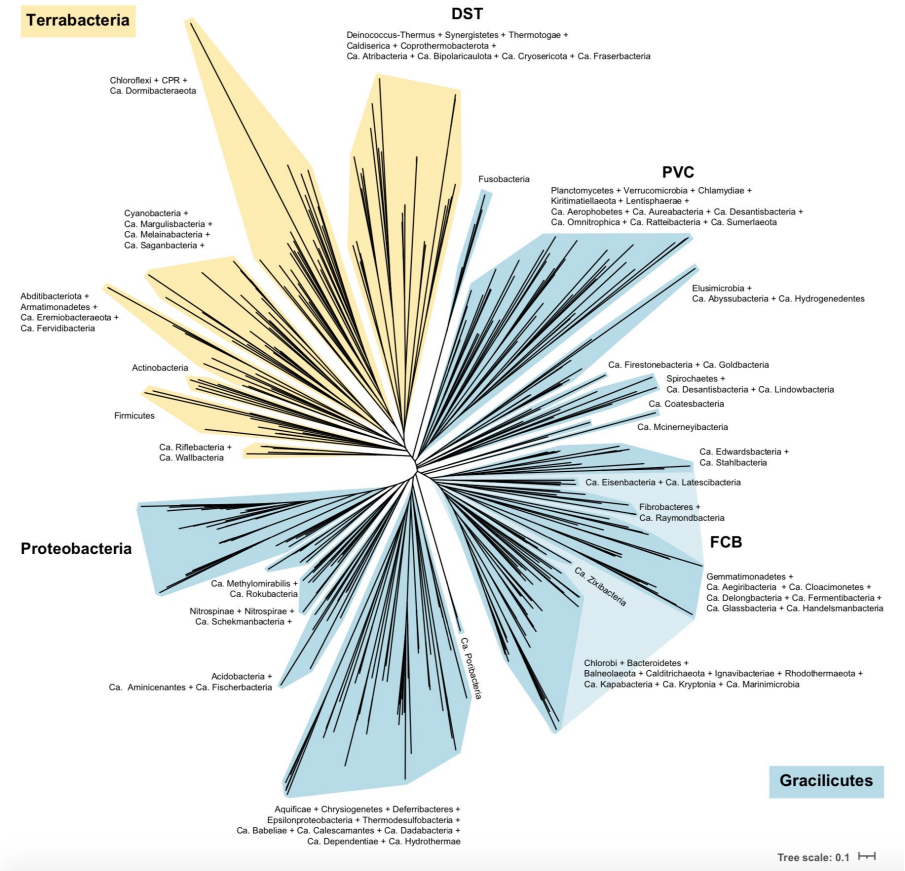
Daniela Megrian
Unidad de Bioinformática

16/10/2025

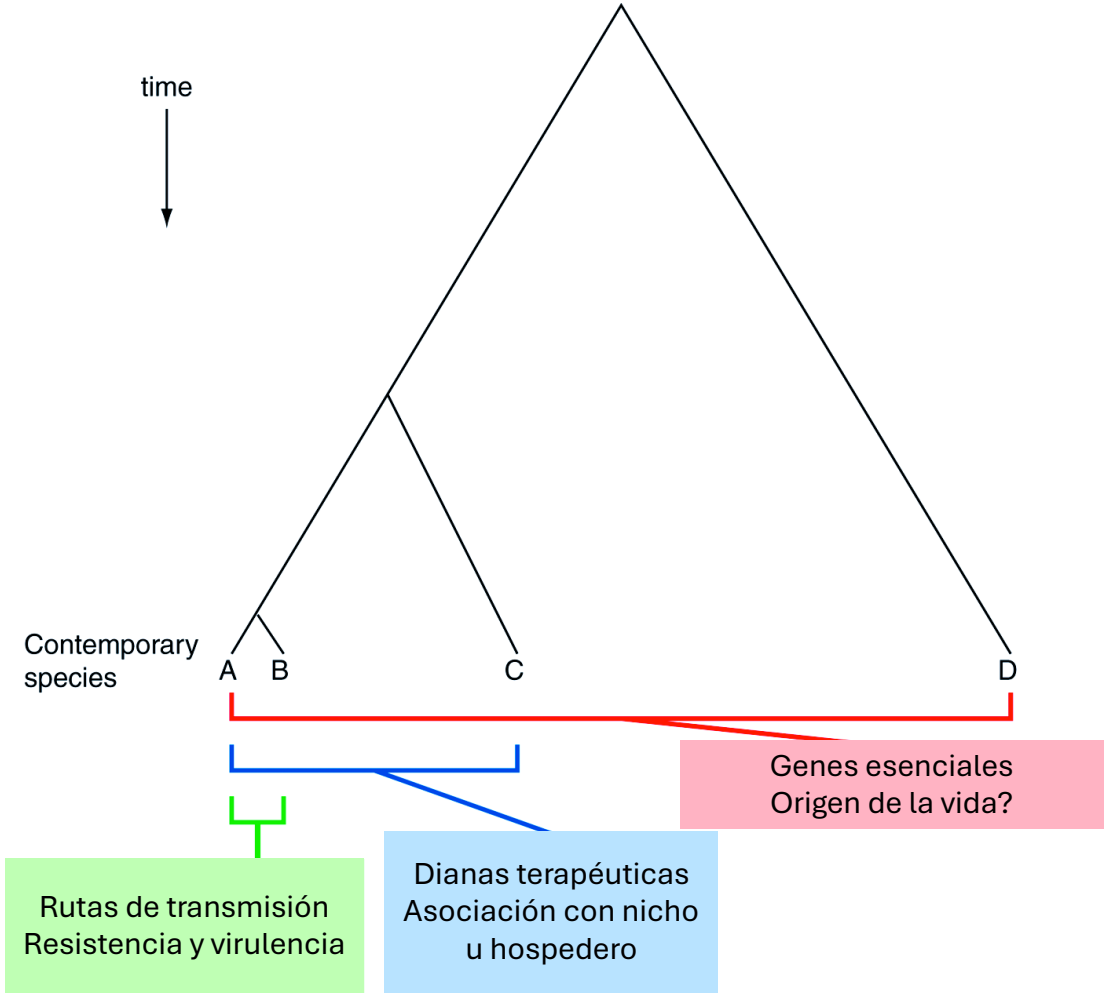
¿Qué es la genómica comparativa?

Comparación de la información genética entre organismos para comprender la **evolución**, la **estructura** y la **función** de los **genes**, las **proteínas** y las **regiones no codificantes**.

Avances en las tecnologías de secuenciación y en los algoritmos de ensamblado.



Diferentes distancias filogenéticas – diferentes preguntas

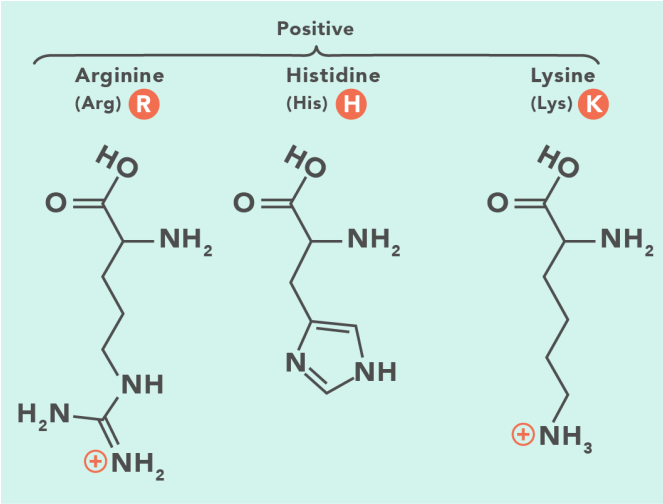


Homología – base de la genómica comparativa

Dos secuencias son homólogas – provienen de un ancestro común.

Identidad: posiciones exactamente iguales en un alineamiento
Similitud: considera sustituciones conservativas.

```
Secuencia 1: M A T H K W L K
Secuencia 2: M A T H R W L K
             | | | | ≈ | | |
```

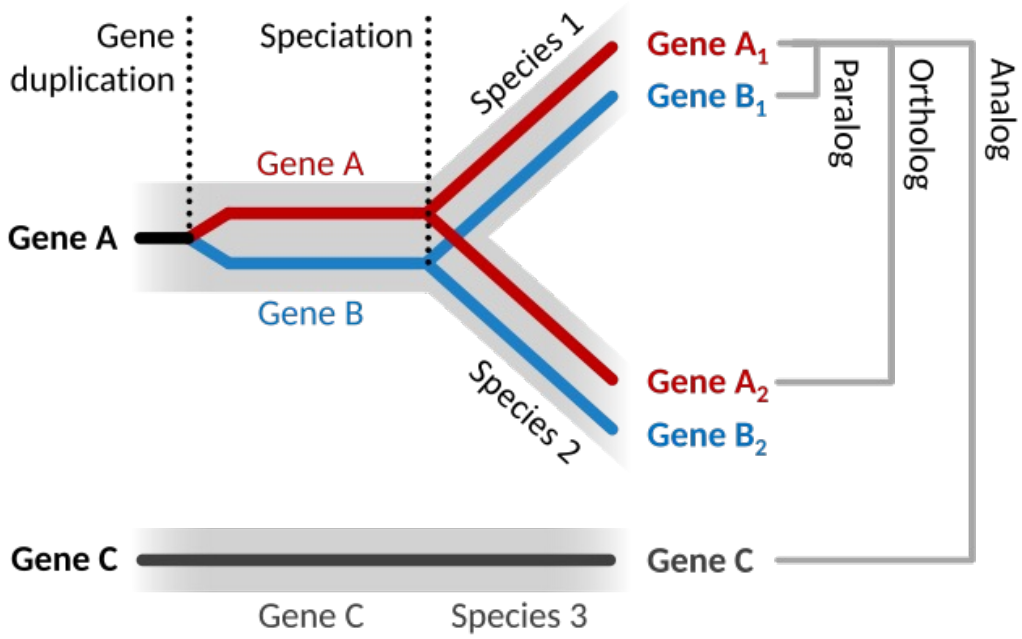


Identidad y similitud son propiedades del alineamiento.
La homología es una relación evolutiva binaria inferida (secuencia, estructura o función)

Ortólogos y parálogos

Ortólogos – genes que divergen a partir de la **especiación**. Suelen conservar la **función ancestral**.

Parálogos – genes que divergen a partir de una duplicación génica dentro de la misma especie. Pueden experimentar **neofuncionalización** o **subfuncionalización**.



Introducción a los algoritmos de alineamiento

Proceso central de la genómica comparativa – alineamiento de secuencias.

Asignación de nucleótidos o aminoácidos de una secuencia a otra.

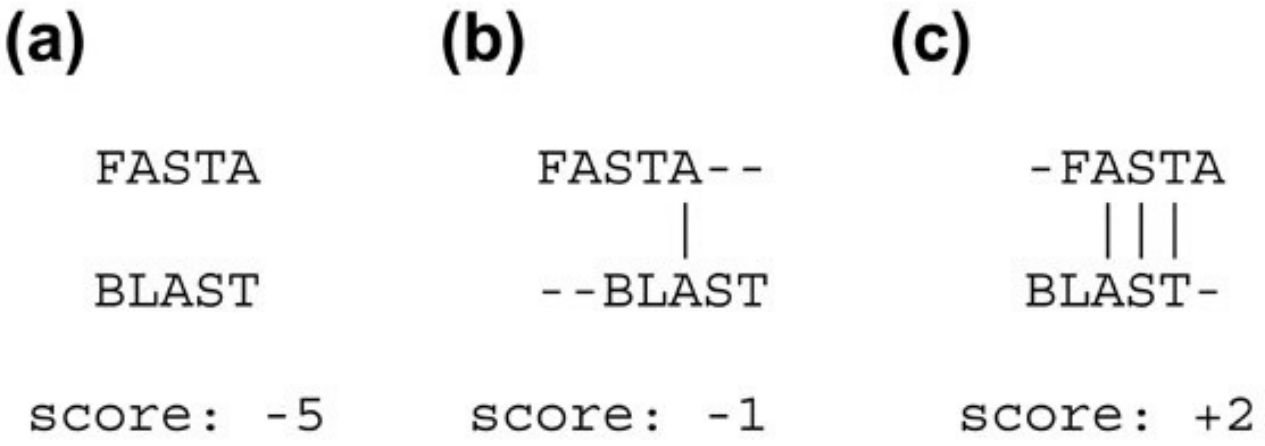
¿Qué tan similares son dos secuencias?

Necesitamos:

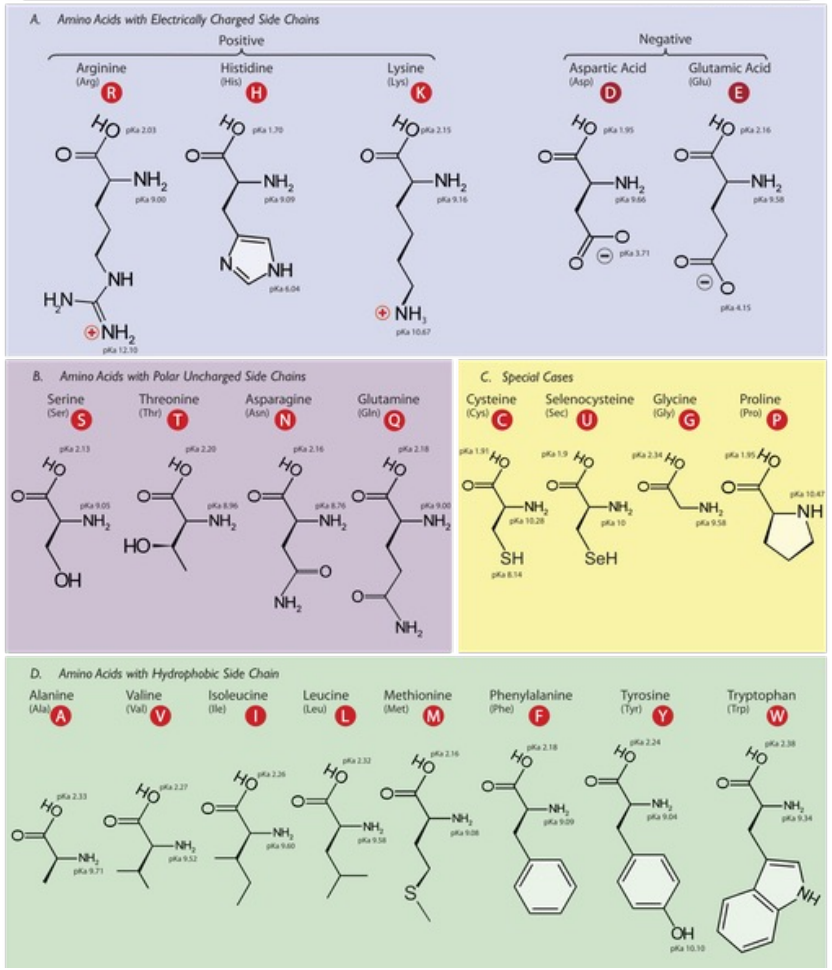
- **Sistema de puntuación**
- **Penalidades para gaps**
- **Método computacional**

Ejemplo de alineamiento

match: 1
mismatch: -1
gap: 0



Matrices de sustitución



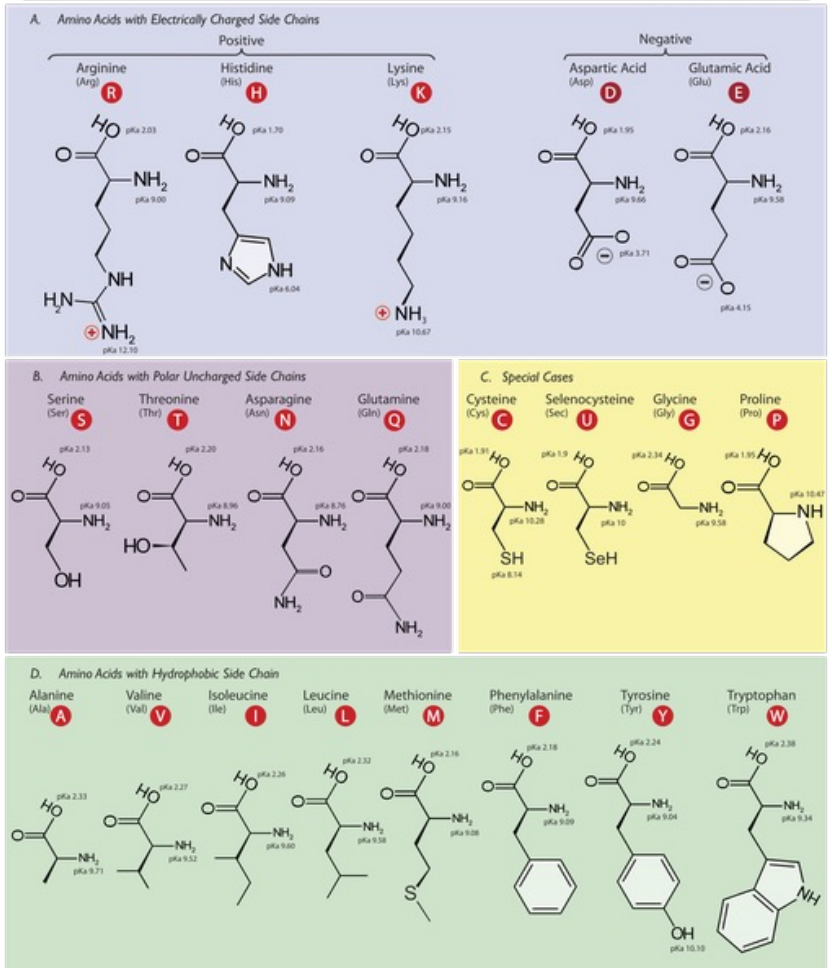

```
VEDAFYTLVREIRQHKLRKLNPPDESGPG
VEDAFYTLVREIRQYRMKKLNSSDDGTQG
VEDAFYTLVREIRQYRLKKISKEE-KTPG
**************:::::*:.  :    *
```

* Conserved sequence (identical)
: Conservative mutation
. Semi-conservative mutation
( ) Non-conservative mutation
\- Gap

Cuando comparamos secuencias de proteínas no todas las sustituciones no son iguales.

Matrices más usadas son PAM y BLOSUM.

Matrices BLOSUM



Construida a partir de bloques de alineamiento de familias proteicas.

BLOSUM62 - umbral de identidad 62

Más bajo – secuencias más diversas – relaciones más distantes.

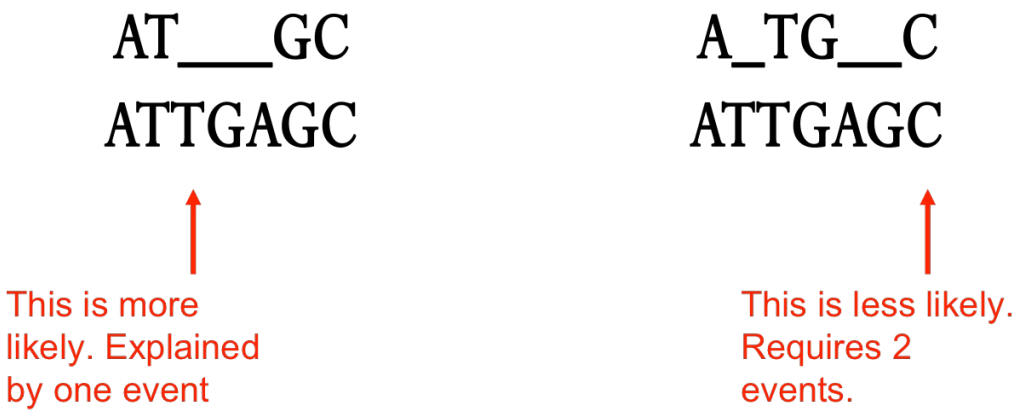
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	-1	0	-2	-2	0	6													
D	-3	1	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	0	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	-1	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	0	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Penalización de gaps

Además de sustituciones – mutaciones incluyen **inserciones y deleciones**.

En un alineamiento se representan como gaps (huecos).

Penalización de apertura + penalización de extensión (menor).
Inserciones o deleciones probablemente afecten varios residuos contiguos.



Programación dinámica: Needleman-Wunsch y Smith-Waterman

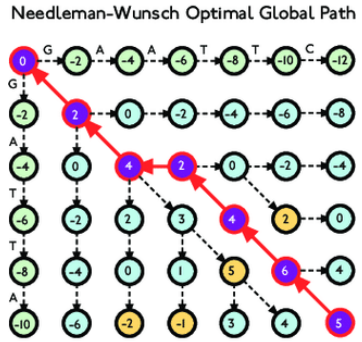
Mejor alineamiento de forma automática.

Programación dinámica convierte alineamiento de secuencias completas en subproblemas.

Needleman-Wunsch (1970) – mejor **alineamiento global** entre secuencias.

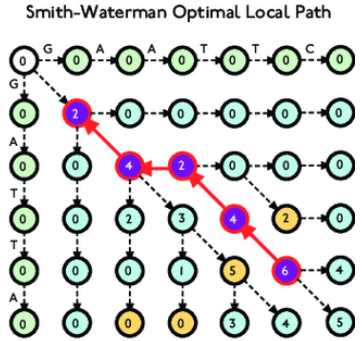
Smith-Waterman (1981) – identifica mejor **alineamiento local**

Garantiza encontrar alineamiento de mayor puntuación – complejidad computacional O(nxm).



Global Alignment (Score=5)

```
GAATTC
GA-TTA
```



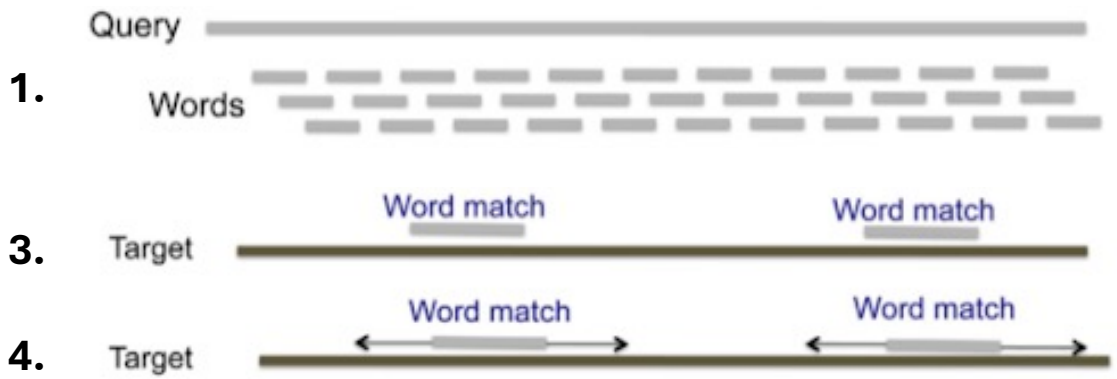
Local Alignment (Score=6)

```
GAATTC
GA-TT
```

BLAST – búsquedas heurísticas

Basic Local Alignment Search Tool - **Aproxima mejor alineamiento** sin garantizarlo.

1. **Divide secuencia en palabras cortas** (3 aa para proteínas).
2. Generar palabras de alta puntuación que podrían hacer match con palabras del query.
3. **Busca matches exactos de esas palabras en base de datos preindexada.**
4. **Cuando encuentra un match intenta extender el alineamiento.**
5. Calcula significancia estadística de cada alineamiento.



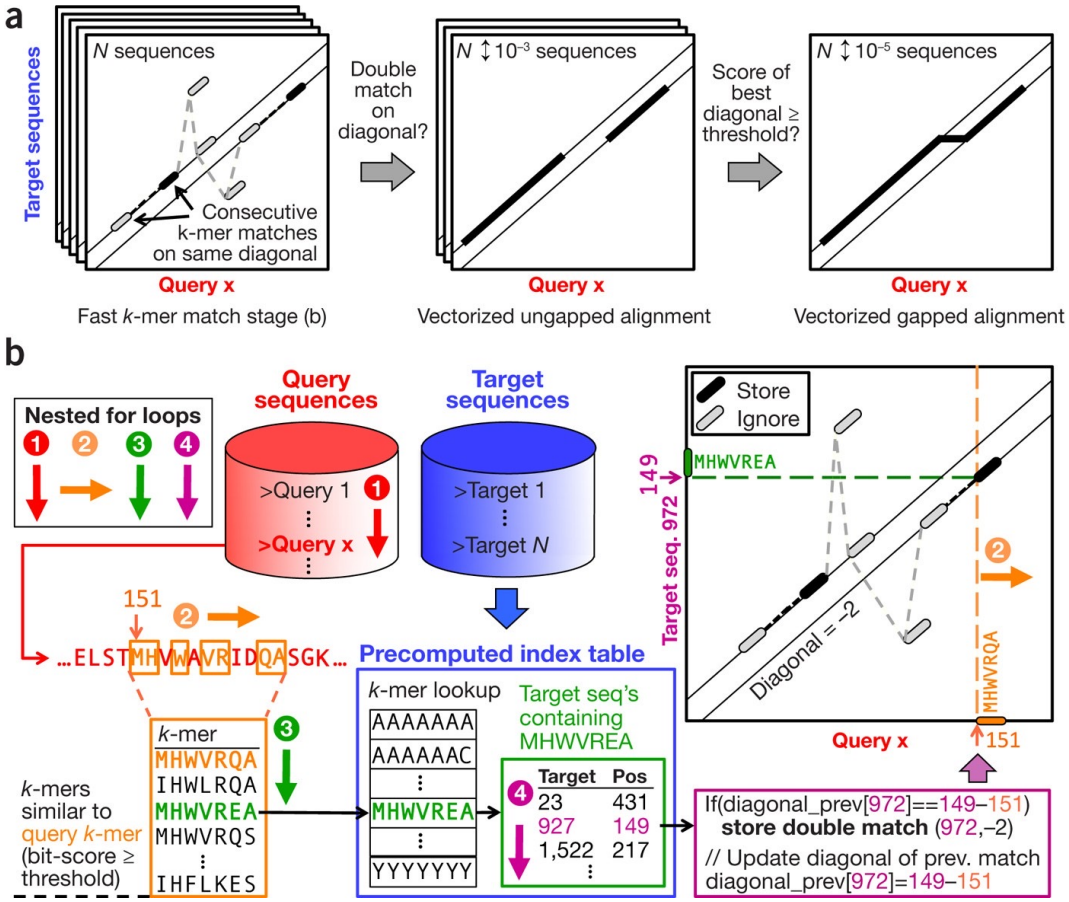
Diamond y MMseqs2

Crecimiento exponencial de base de datos.

Diamond – algoritmo similar a BLAST con indexación más eficiente.

MMseqs2 – pipeline multi-etapa con filtrado muy eficiente por k-mers.

BLAST, Diamond y MMseqs2 usan seed and extend.

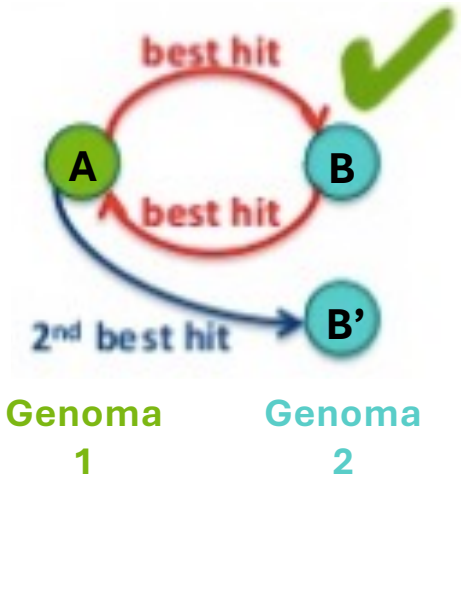


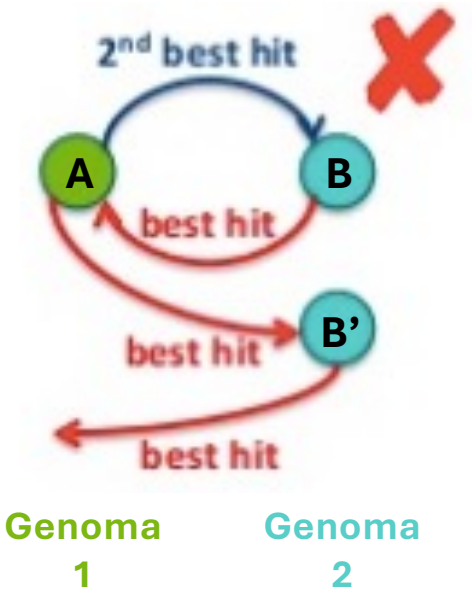
Métodos de identificación de ortólogos
Reciprocal Best Hits (RBH)

Método más simple – si el gen A en el genoma 1 encuentra al gen B en el genoma 2 como su mejor hit y viceversa, entonces A y B son probablemente ortólogos.

Funciona bien cuando:

- No hay duplicaciones génicas recientes
- Tasas de evolución relativamente uniformes
- Genomas completos

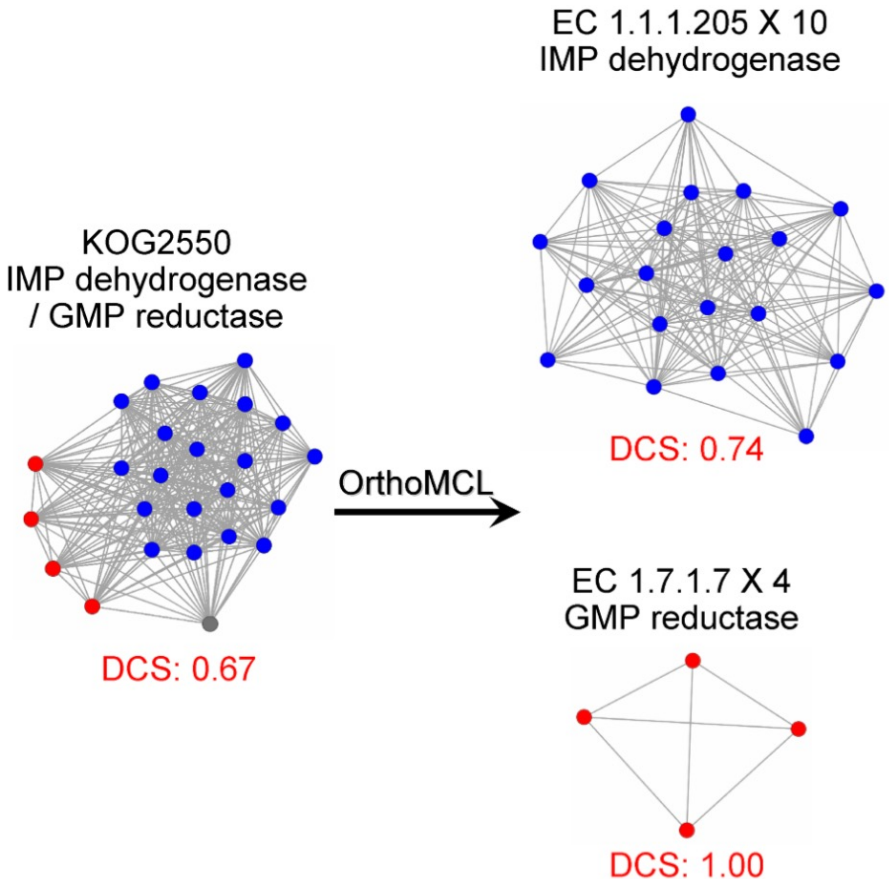




OrthoMCL

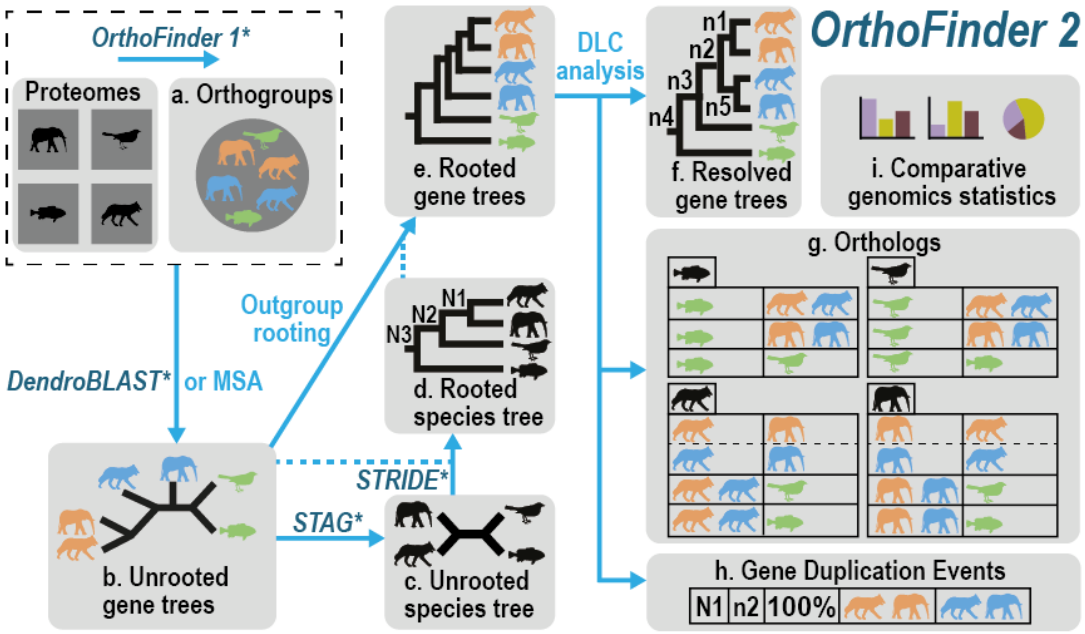
Agrupa proteínas en “**orthogroups**” (grupos de ortología):

1. **Alinea todo contra todo** con BLAST/DIAMOND.
2. **Identifica RBH** y detecta “**in-paralogs**” intra especie.
3. Construye **grafo con pesos derivados del score**.
4. **Clusteriza con MCL** (Markov Cluster Algorithm) – separa ortólogos e incorpora in-paralogs recientes.



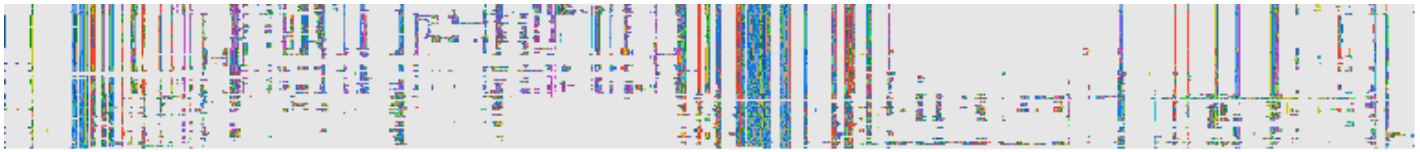
OrthoFinder2

1. **Alinea todo contra todo.**
2. **Construye grafo de similitud.**
3. Aplica MCL sobre el grafo.
4. **Para cada orthogroup infiere un árbol.**
5. Estima árbol de especies (STAG - combina señal de muchos árboles génicos)
6. Enraiza el árbol (STRIDE).
7. **Reconciliación gen-especie.**



No es tan sencillo...

Proteína **GlpR**
Actinobacteria

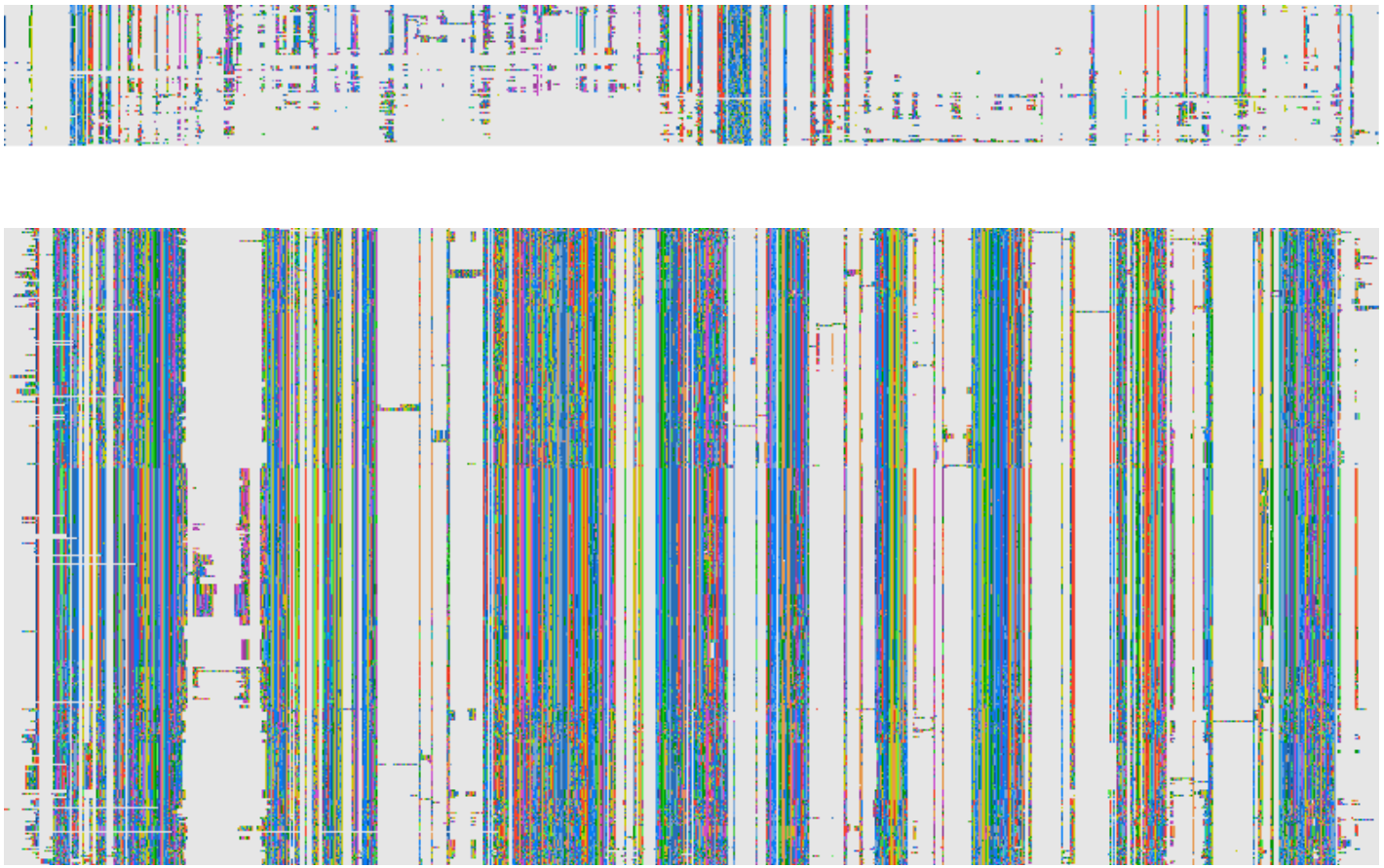


No es tan sencillo...

Proteína **GlpR**
Actinobacteria

Proteína **Glp**
Actinobacteria

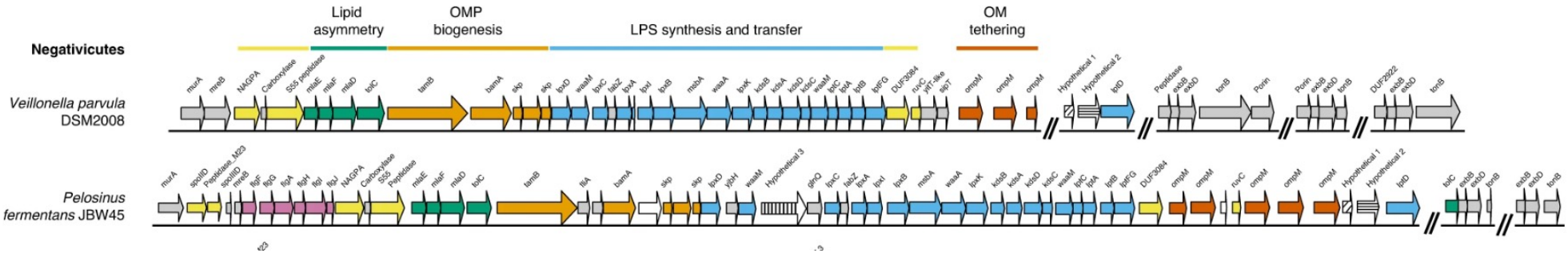
Proteína **MoeA**
Actinobacteria



Sintenia

Se refiere a la conservación del orden (y orientación) de los genes en cromosomas de especies relacionadas.

Operones en bacterias.



Pangenómica bacteriana

Pangenoma – **conjunto completo de familias génicas** observadas en un grupo de genomas definido.
Reconstrucción funcional y ecológica.
Vigilancia epidemiológica.
Más diversidad filogenética, más grande y heterogéneo – más difícil interpretación.

Core genome – **familias génicas presentes en todos los genomas** del grupo.
Base para filogenias de especies robustas.
Anotación y transferencia de función.