Alineamiento de secuencias de proteína y filogenias

Bruno Contreras-Moreira e Inmaculada Yruela
http://www.eead.csic.es/compbio
Fundación ARAID y Estación Experimental de Aula Dei/CSIC, Zaragoza, España

---

Este curso es un tutorial paso a paso sobre cómo definir una familia de proteínas, por medio de alineamientos, que luego podremos aprovechar para inferir una filogenia molecular. A lo largo del texto hay referencias a la literatura especializada y algunos enlaces a recursos externos recomendados.

El texto está en español, aunque encontrarás enlaces y figuras en inglés. Para completar el estudio de estos temas recomendamos estos libros:

- Bioinformática con Ñ
- The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing

y el blog de nuestro laboratorio, donde de vez en cuando tocamos temas relacionados:

- #!/perl/bioinfo

Este material fue creado inicialmente para el curso de verano de la Universidad de Zaragoza "Estructura y Función de Proteínas".

Análisis jerárquico de la estructura de proteínas

Desde el punto de vista de la estructura, las proteínas pueden analizarse de forma jerárquica en 4 niveles, partiendo de la estructura primaria, su secuencia, hasta llegar a su estructura cuaternaria, cuando interaccionan con otras moléculas.

Estructura primaria

La estructura primaria de una proteína se corresponde con la secuencia lineal de aminoácidos codificada en su gen correspondiente, y suele representarse por medio de una cadena donde cada letra identifica a un aminoácido o residuo. Por ejemplo, los primeros 30 aminoácidos de la insulina de la mosca *Drosophila melanogaster* son:

```
MFSQHNGAAV HGLRLQSLLI AAMLTAAMAM...
```

En formato FASTA sería:

```
>Insulina [Drosophila melanogaster]
MFSQHNGAAV HGLRLQSLLI AAMLTAAMAM...
```

donde por ejemplo M es metionina, Q es glutamina o A es alanina (ver tabla 1). El sentido de la cadena es desde el extremo amino-terminal hacia el carboxilo-terminal.

Tabla 1: Nomenclatura de los 20 aminoácidos esenciales

A	ALA	Alanina	M	MET	Metionina
C	CYS	Cisteína	N	ASN	Asparagina
D	ASP	Aspartato	P	PRO	Prolina
E	GLU	Glutamato	Q	GLN	Glutamina
F	PHE	Fenilalanina	R	ARG	Arginina
G	GLY	Glicina	S	SER	Serina
H	HIS	Histidina	T	THR	Treonina
I	ILE	Isoleucina	V	VAL	Valina
K	LYS	Lisina	W	TRP	Triptófano
L	LEU	Leucina	Y	TYR	Tirosina
X	-	desconocido			

Estructura secundaria

Para neutralizar las cargas polares del esqueleto peptídico, las proteínas adoptan conformaciones que maximizan la formación de puentes de hidrógeno, gracias a la libertad de giro de los enlaces situados inmediatamente antes y después del enlace peptídico. Esto lo hacen principalmente formando α-hélices dextrógiras, láminas β, como se muestra en las figuras 1 y 2, giros de varios tipos, y a veces regiones desordenadas.

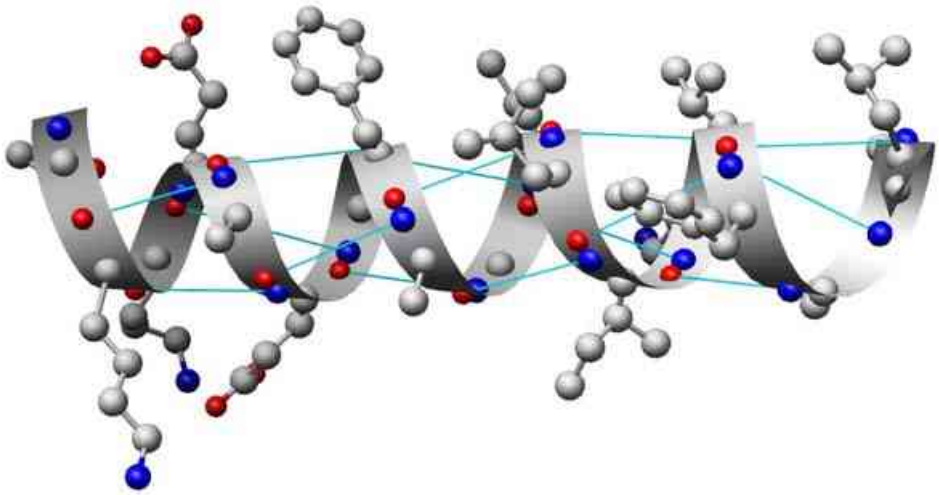

Figura 1: Alfahélice mostrando cadenas laterales y puentes de hidrógeno, tomada de http://structuralbioinformatics.com.

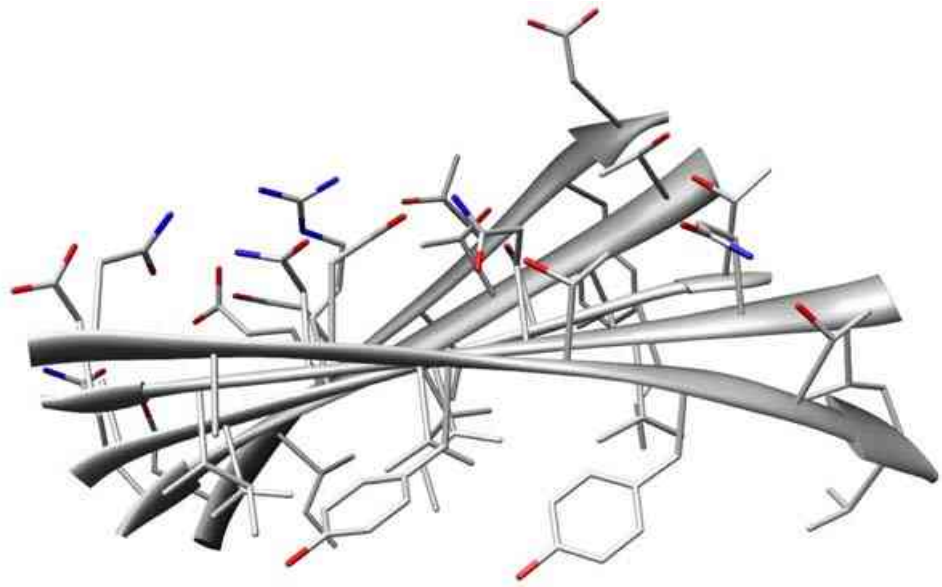

Figura 2: Diagrama de láminas beta antiparalelas mostrando la disposición de las cadenas laterales, tomado de http://structuralbioinformatics.com.

La estructura secundaria de las proteínas se puede codificar de manera similar a la secuencia primaria, asignando a cada residuo una letra que identifica el estado de estructura secundaria en que se encuentra. Se suele identificar a los residuos de una α-hélice con H, los de una lámina β con E (de extendida) y los demás con C, del inglés *coil* . Cuando la estructura secundaria es de especial interés se pueden hacer subclases del estado C, como T (del inglés *turn*) o B (de horquilla β, *hairpin*).

La misma secuencia que vimos antes podría tener esta estructura secundaria simplificada:

```
MFSQHNGAAV HGLRLQSLLI AAMLTAAMAM...
EEEECCEEEE HHHHHHHHHH CCCCCCCCCC...
```

Estructura terciaria y cuaternaria

La mayoría de las proteínas forman glóbulos compactos al plegarse, compuestos de elementos de estructura secundaria conectados por lazos (*loops* en inglés). Las unidades globulares de plegamiento pueden llamarse dominios (Porter & Rose, 2012), y tienen en su interior sobre todo cadenas laterales hidrofóbicas (Isom *et al.*, 2010), mostrando hacia el solvente los lazos (Branden & Tooze, 1999). En raras

ocasiones, una proteína puede formar nudos al plegarse (Potestio *et al.*, 2010; King *et al.*, 2010) o agregarse como amiloides (Schnabel, 2010).

Las clasificaciones estructurales de proteínas se hacen generalmente a nivel de dominios, como en el caso de SCOP, CATH o la taxonomía de Richardson, aunque también se han propuesto otros esquemas, como la tabla periódica de Taylor (2002).

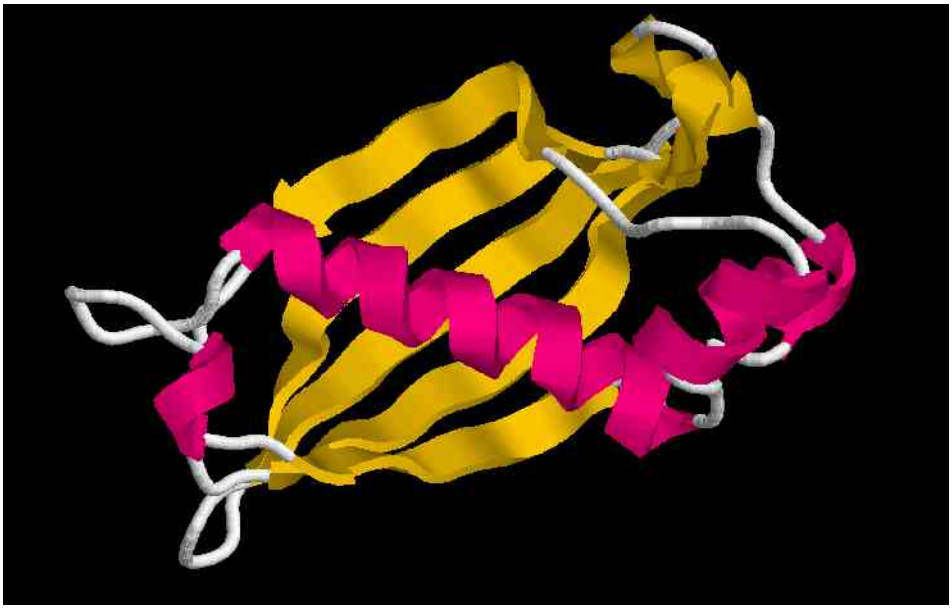

Figura 3: Estructura terciaria de una proteína (tioesterasa), con los elementos de estructura secundaria coloreados
(amarillo para la láminas betas, rosa para alfa-hélices y blanco para los lazos). Figura exportada con el programa RasMol.

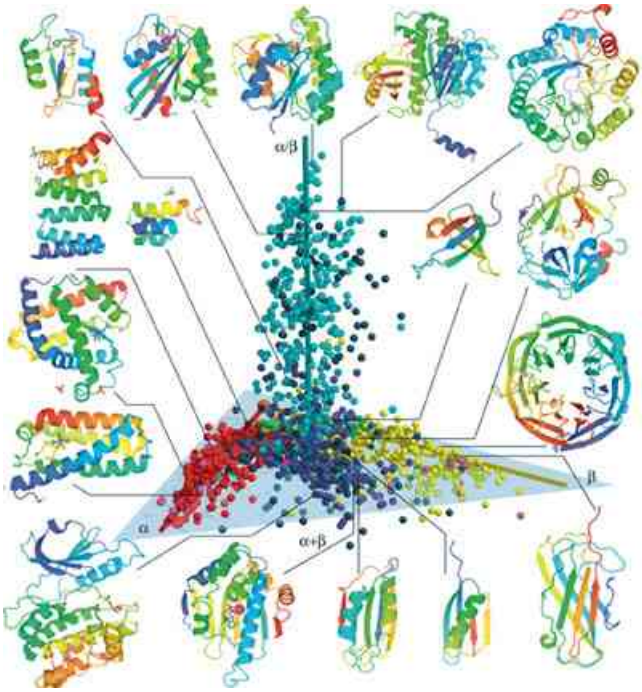

Figura 4: Distribución de plegamientos de proteínas de las 4 clases principales de SCOP, tomada de Hou *et al.* (2005).

Un tipo especial de estructura terciaria es de las proteínas transmembranales, que se pliegan y ejercen su papel dentro de una bicapa lipídica en vez de una solución más o menos acuosa:

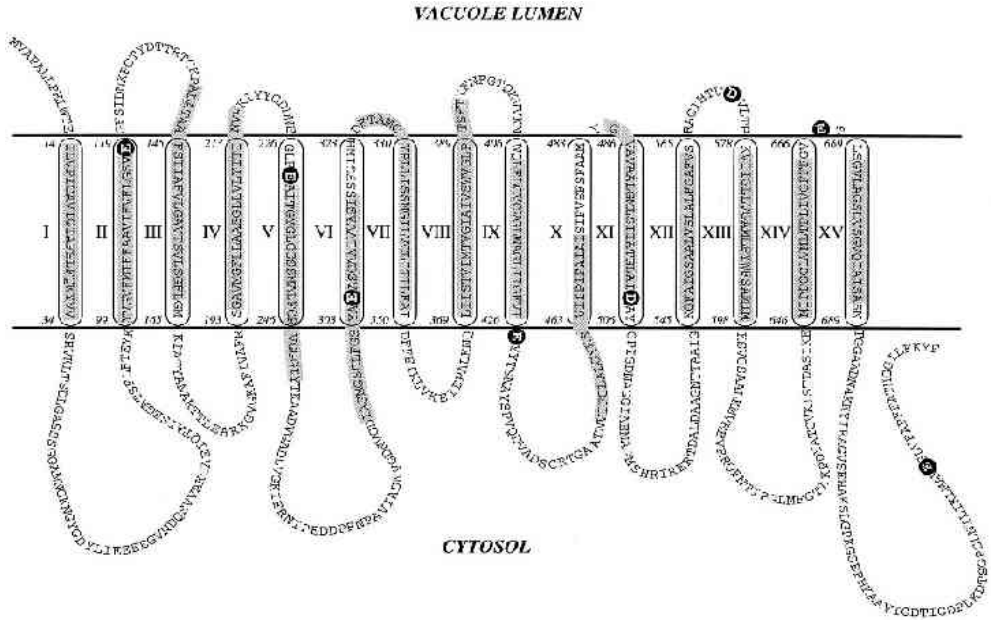

Figura 5: Modelo de topología transmembrana tomado de Zhen *et al.* (1997).

Demarcación de dominios en proteínas

Las proteínas son máquinas moleculares que ejercen una gran variedad de funciones. De hecho, una misma proteína puede desempeñar distintos papeles a la vez. Para poder manejar esta complejidad, a la unidad funcional y evolutiva le llamamos *dominio* . Esta definición es cómoda, pero no ha evitado que haya controversias, porque la dinámica evolutiva de las proteínas no es discreta (Pascual-García *et al.*, 2009). Sin embargo, desde un punto de vista práctico, podemos generalizar y decir que a cada dominio le corresponde una función molecular. En la figura 6 vemos que un dominio se puede normalmente reducir a un conjunto de elementos de estructura secundaria que se pliegan juntos.

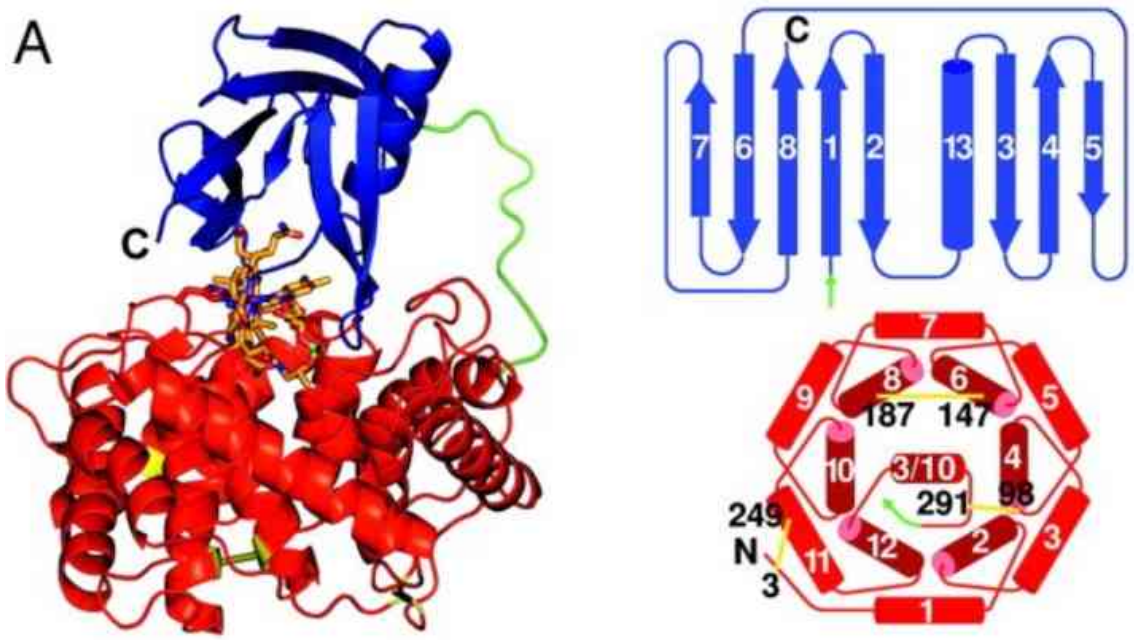

Figura 6: Estructura terciaria de una proteína con dos dominios (azul y rojo), con su topología a la derecha. Figura adaptada de Wuerges *et al.* (2006).

Los recursos de referencia para la definición de dominios de proteínas han sido tradicionalmente las bases de datos SCOP y Pfam. Ambas bases de datos dependen de la curación manual de datos por parte de expertos, pero siguen enfoques distintos:

- El material de partida de SCOP son estructuras contenidas en el Protein Data Bank; define

dominios en base a criterios estructurales y evolutivos.

- Pfam define familias de secuencias o dominios en base alineamientos múltiples para facilitar su localización en otras proteínas.

Como muestra del tamaño del repertorio de dominios conocidos, la versión 28.0 de Pfam contiene 16230 familias, mientras que SCOPe v2.05 contiene 4756.

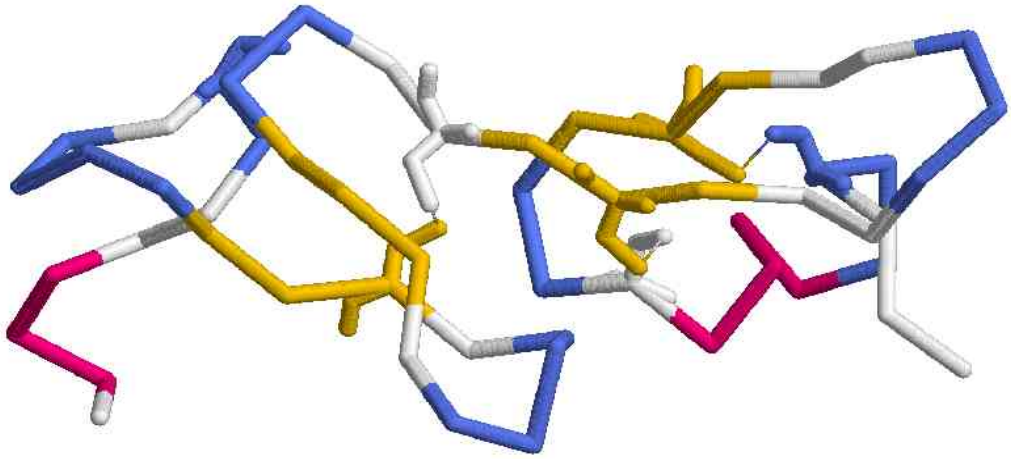

Figura 7: Definición de un dominio EGF contenido en la estructura 1W8B según SCOP.
El dominio incluye 4 láminas beta (en amarillo) unidas con 3 puentes disulfuro entre 6 residuos de cisteína.

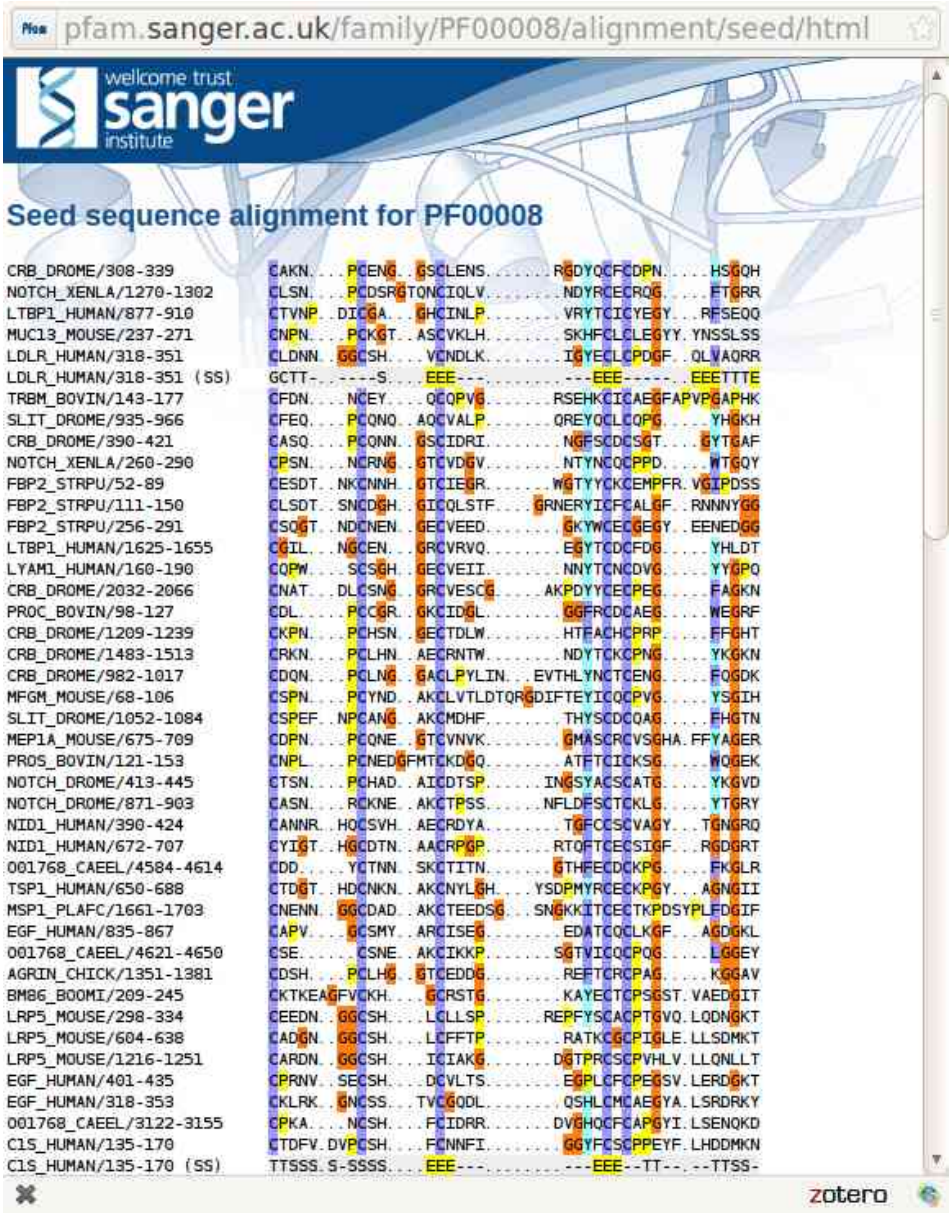

Figura 8: Alineamiento semilla de la familia/dominio EGF en Pfam.

Comparación de secuencias de proteína y conservación de su estructura y función

Hay muchas maneras de comparar proteínas, pero entre ellas posiblemente la más natural sea el alineamiento de sus secuencias. Esto se debe, por un lado, a que la estructura primaria determina en gran medida el plegamiento, y por otro lado a que es la manera más sencilla de manipular proteínas en el cuaderno y el ordenador.

Al alinear dos o más secuencias establecemos exactamente qué residuos de unas se corresponden con los de otras. Esta correspondencia es obvia cuando las secuencias son muy parecidas, pero se va diluyendo a medida que las secuencias acumulan mutaciones. La medida más utilizada para medir esta semejanza o divergencia es el % de identidad de secuencia, que se calcula como se muestra en la figura 9. Otra medida utilizada para este fin son los valores esperados o *E-values* de programas como BLASTP.

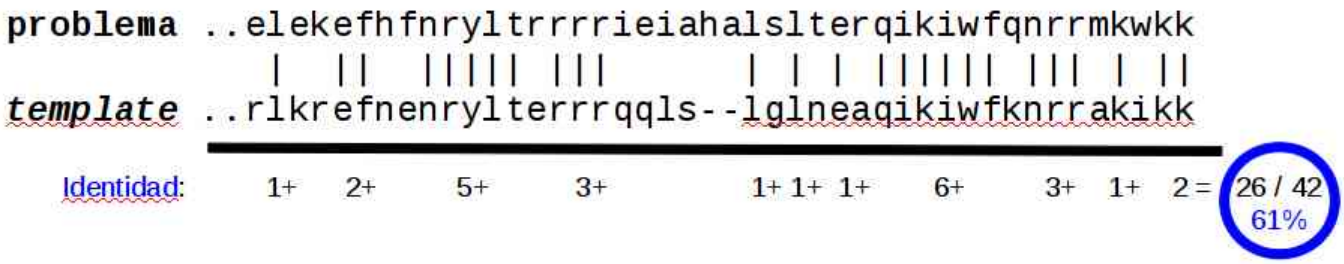

Figura 9: Ejemplo de alineamiento pareado entre una secuencia problema y otra, de estructura conocida, a la que llamamos molde, subject o template en inglés, detallando el cálculo de la identidad de secuencia, que es la proporción de posiciones de un alineamiento que son idénticas.

Chothia & Lesk (1986) analizaron por vez primera la relación entre la secuencia y la estructura de las proteínas, que se puede resumir en esta figura:

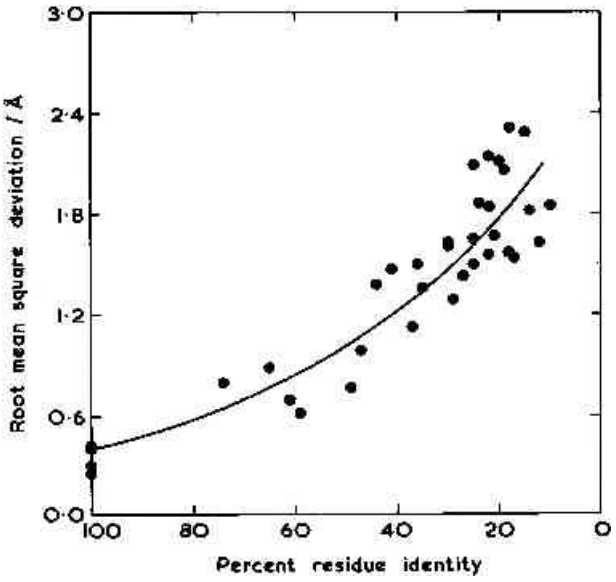

Figura 10: Correlación no lineal entre la conservación de secuencia y estructura de las proteínas, tomada de Chothia & Lesk (1986).

En este artículo se hace la observación de que a una determinada conservación entre las secuencias A y B, calculada por medio de un alineamiento, le corresponde una mayor o menor divergencia en la comparación de sus estructuras terciarias, medida en términos de desviaciones (RMSD) en las posiciones de sus residuos, dependiendo de si las mutaciones ocurren en el interior (*core*) o exterior del plegamiento.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(equiv_i^A - equiv_i^B)}{n}} \qquad (0.1)$$

Además, éste y otros trabajos posteriores, como el de Illergard *et al.* (2009), sugieren que la estructura es una propiedad de las proteínas que se conserva en mayor medida que la secuencia durante la historia evolutiva. Lo excepcional es encontrar secuencias parecidas con grandes diferencias estructurales (Kosloff & Kolodny, 2008).

Algoritmos de alineamiento y matrices de sustitución de aminoácidos

En el ejemplo 9 ya vimos cómo se pueden alinear dos secuencias. Ahora revisaremos los tipos fundamentales de algoritmos de alineamiento, que deberemos conocer para aplicar según nuestras necesidades:

- Alineamiento pareado. Alinea dos secuencias entre si.
- Alineamiento múltiple. Alinea tres o más secuencias entre si.
- Alineamiento global. Alinea secuencias completas, incluyendo en el alineamiento resultante todos los residuos de todas las secuencias de entrada.
- Alineamiento local. Optimiza el alineamiento de las subsecuencias que se parecen y descarta el resto.

Estos algoritmos utilizan métricas para puntuar el emparejamiento de aminoácidos de proteínas distintas. Para ello usan matrices de sustitución, de las que probablemente la más conocida sea BLOSUM62, la que usa BLASTP:

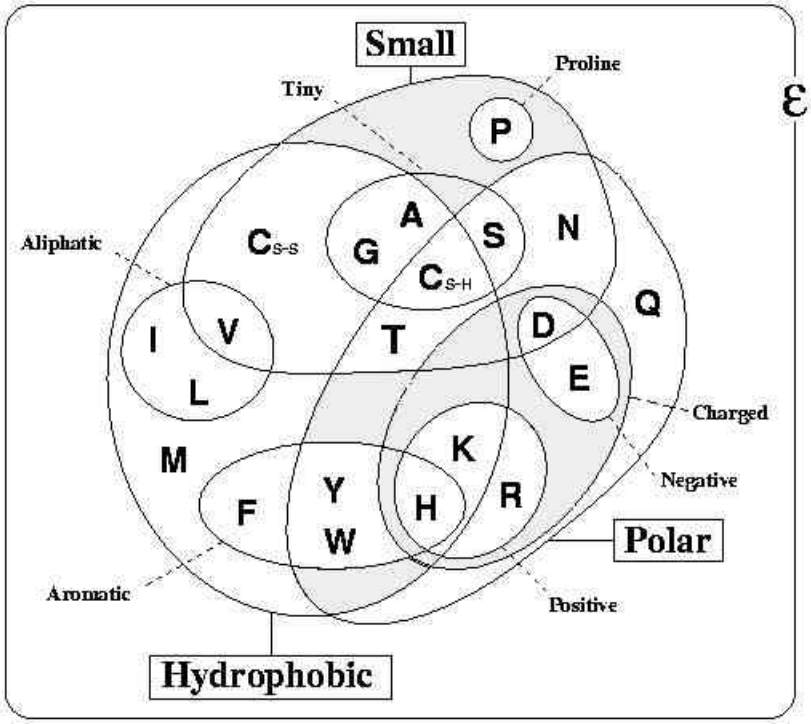

Figura 11: Clasificación de los 20 aminoácidos naturales, tomada de http://www.russelllab.org/aas.

Tabla 2: BLOSUM62, contiene estimaciones del valor evolutivo del cambio de un aminoácido por otro en una proteína.

Por ejemplo, una mutación de alanina (A) a aspártico (D) está penalizada con un -2.

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

Empleando matrices como BLOSUM62 los algoritmos pueden calcular la *similitud* entre secuencias, que es la suma de puntuaciones de parejas de residuos alineados tras restar penalizaciones por inserciones y deleciones (*indels*):

```
V L F L I V I D
V L Y L V - I I

Identidad  =  1+1+0+1+0+0+1+0  =    4

Similitud  =  4+4+3+4+3-2+4-3  =   17
```

Figura 12: Ejemplo de cálculo de identidad y similitud entre dos péptidos.

Alineamiento múltiple (normalmente global)

Un alineamiento múltiple es una manera sistemática de comparar simultáneamente las secuencias de diferentes proteínas de una misma familia, que normalmente comparten dominios en el mismo orden. Es por tanto normalmente un alineamiento global. El algoritmo clásico para alinear globalmente dos secuencias es el de Needleman & Wunsch (1970):

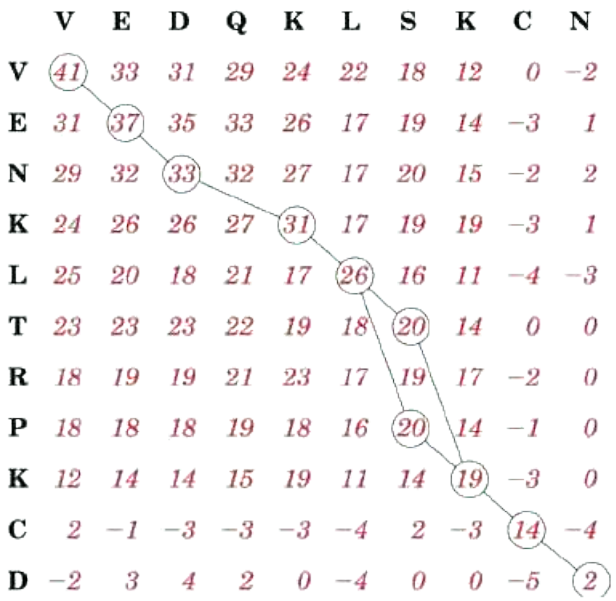

Figura 13: Ejemplo de aplicación del algoritmo de Needleman-Wunsch para alinear dos péptidos de diferente longitud. El diagrama muestra la etapa final, donde se recorre la matriz de programación dinámica desde la casilla con mayor puntuación cercana a la esquina inferior derecha hasta la esquina opuesta.

Para dos secuencias de longitud L y M, este algoritmo tiene un coste cuadrático porque debe calcular LxM puntuaciones. La extensión de Needleman-Wunsch a más secuencias es por tanto muy costosa y en la práctica se toman atajos, o heurísticas en la jerga. El software más popular para el cálculo de alineamientos múltiples, posiblemente Clustal Omega, primero calcula alineamientos pareados de todas las secuencias a alinear y con ellos calcula un árbol guía que define en qué orden se van a ir añadiendo las secuencias al alineamiento múltiple de manera progresiva:

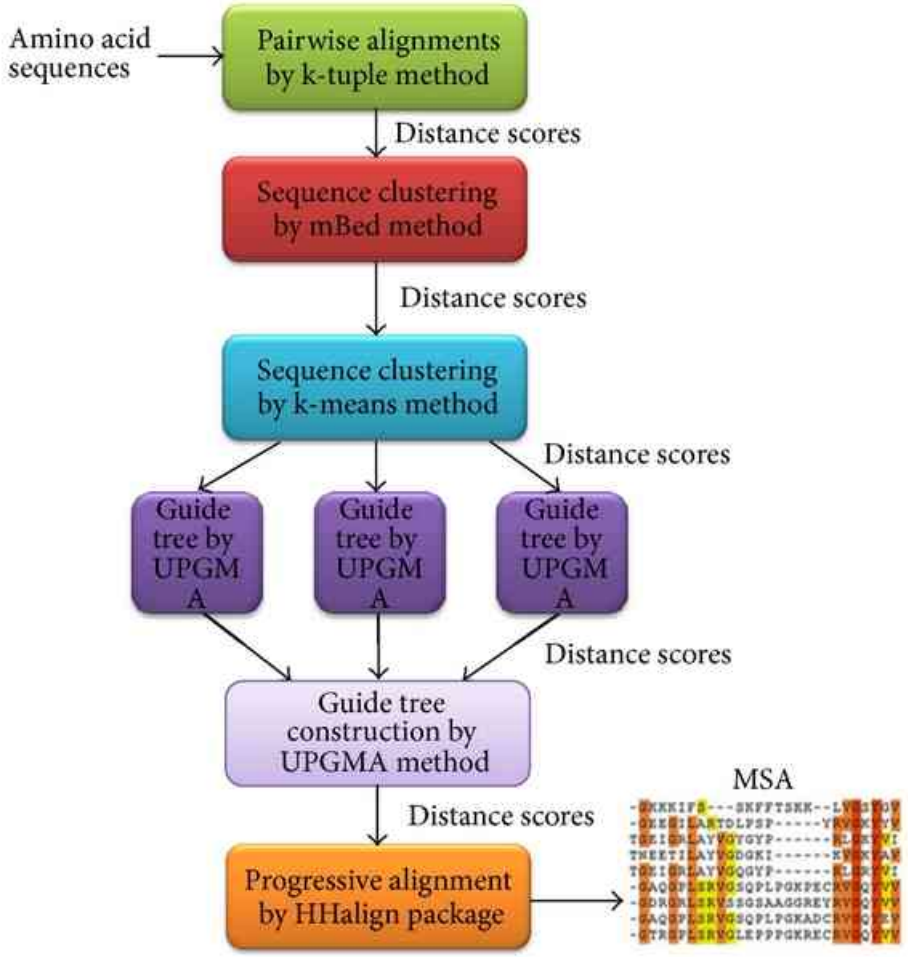

Figura 14: Diagrama de flujo de Clustal Omega,
tomado de http://www.hindawi.com/journals/isrn/2013/615630/.

Alineamiento pareado (normalmente local)

Los alineamientos pareados se utilizan habitualmente para comparar cuantitativamente dos secuencias y obtener una puntuación que podamos comparar con la de otras secuencias. Tienen muchas aplicaciones, pero destacamos tres:

- Buscar, dentro de una colección de secuencias, a cuál se parece más una proteína problema.
- Comprobar si dos proteínas comparten un dominio aunque el resto de la secuencia sea diferente.
- Obtener información funcional a partir de las anotaciones de secuencias parecidas en una base de datos.

El algoritmo fundamental para calcular el alineamiento local óptimo entre dos secuencias es el de Smith & Waterman (1981), que es esencialmente una modificación del de Needleman & Wunsch (1970). En la práctica, dado el tamaño de las colecciones de secuencias de referencia como Uniref o nr, este algoritmo es poco eficiente y se usan heurísticas como BLASTP, que es ya una herramienta de uso universal, que produce alineamientos como éste:

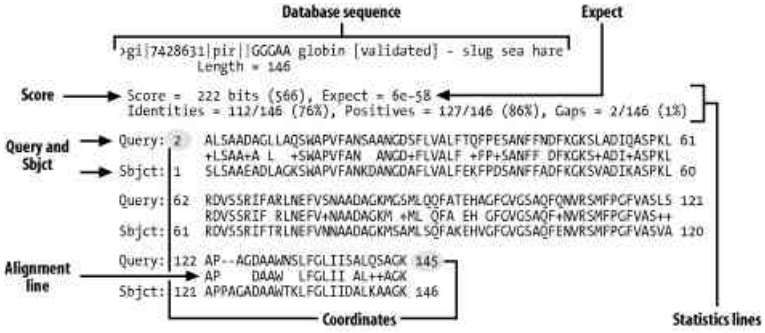

Figura 15: Ejemplo de alineamiento de BLASTP, tomado de http://etutorials.org.

Para interpretar correctamente un alineamiento local de BLASTP debemos conocer estos conceptos:

- Score. Es la similitud de las dos secuencias comparadas, normalizada y en bits.
- La secuencia alineada de la base de datos. La información mostrada contiene anotaciones funcionales y otros nombres con el que se conoce la misma secuencia.
- Expect o *E-value* . Es una estima del número de alineamientos al azar que se esperarían en esa base de datos con el mismo *Score* .

Modelos evolutivos de proteínas

Los alineamientos, además de comparar secuencias, pueden usarse como estimadores de homología. A diferencia de funciones cuantitativas como la identidad o la similitud, la homología es una propiedad cualitativa binaria: se es homólogo o no. Como la evolución ocurre generalmente a nivel de ADN, realmente se habla de genes homólogos cuando se supone que son descendientes de un mismo gen ancestral, pero por extensión hablamos también de proteínas homólogas.

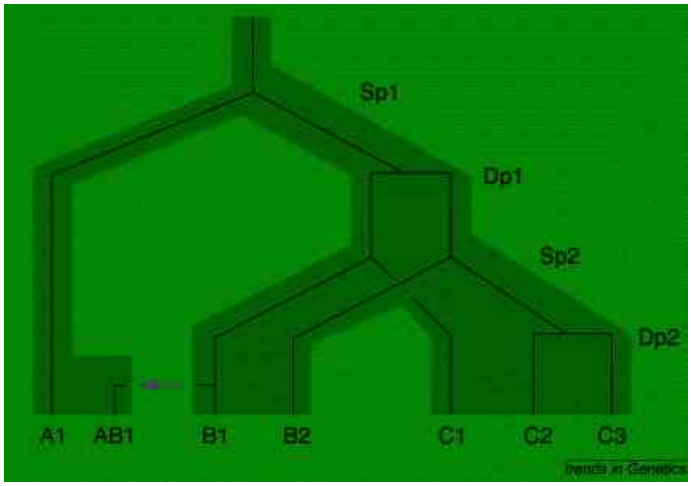

Figura 16: Ejemplo de relaciones de homología entre genes de especies hermanas, tomado de Fitch (2000).

El interés por la historia evolutiva de las proteínas en nuestro contexto se debe a que en general las proteínas homólogas conservan su estructura (ver Figura 10) y, más concretamente, a que las ortólogas suelen conservar sus patrones de expresión y su función biológica (Gabaldon & Koonin, 2013).

Mientras que niveles altos de similitud son normalmente evidencia de homología, sobre todo con identidades por encima de la zona de penumbra del 20-30% de identidad, la ortología se determina con mayor fiabilidad calculando filogenias moleculares, sobre todo entre genomas eucariotas. En las

siguientes secciones veremos brevemente como calcular filogenias.

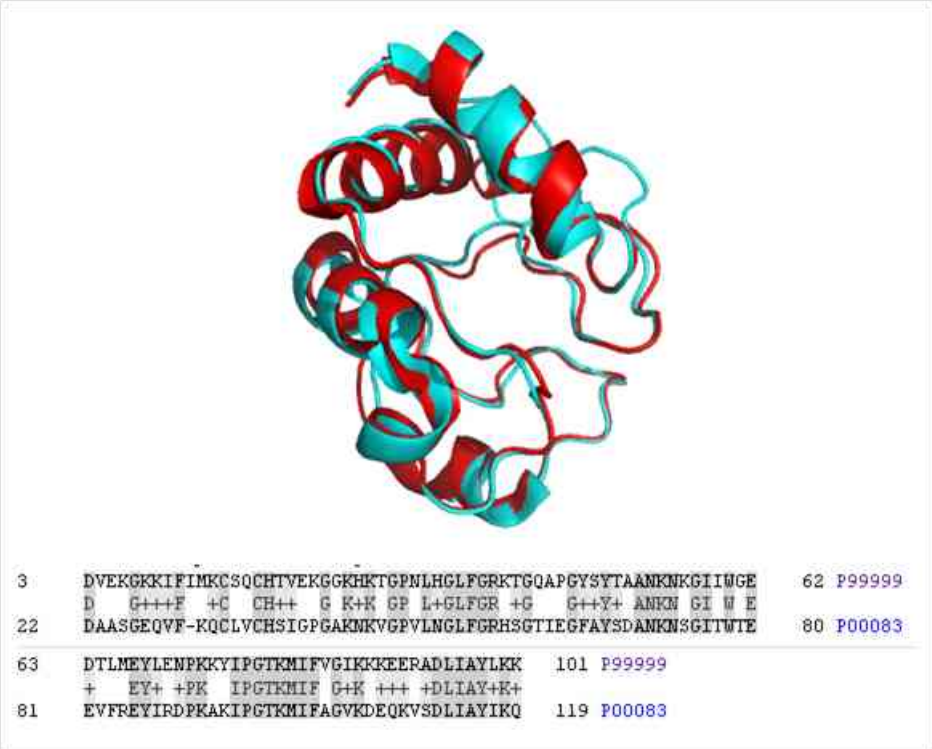

Figura 17: Conservación de la estructura y la secuencia del citocromo C humano y el citocromo C2 de la bacteria Rhodopseudomonas viridis.

Cómo evolucionan las secuencias de genes y proteínas

Algunos mecanismos que provocan cambios en las secuencias de genes son:

- mutación: cuando el ADN se replica se pueden producir errores al realizar la copia; también pueden introducirse mutaciones por efecto de agentes externos (mutagénicos) como por ejemplo la luz ultravioleta.

- duplicación: cuando un gen se duplica se abre una puerta para la adquisición de nuevas funciones biológicas. Las mutaciones en el nuevo gen son más fácilmente tolerables. Estos genes pueden degenerar convirtiéndose en pseudogenes.

- barajado de dominios: muchas proteínas están constituidas por dominios. Mediante recombinación se pueden barajar y producir nuevas combinaciones.

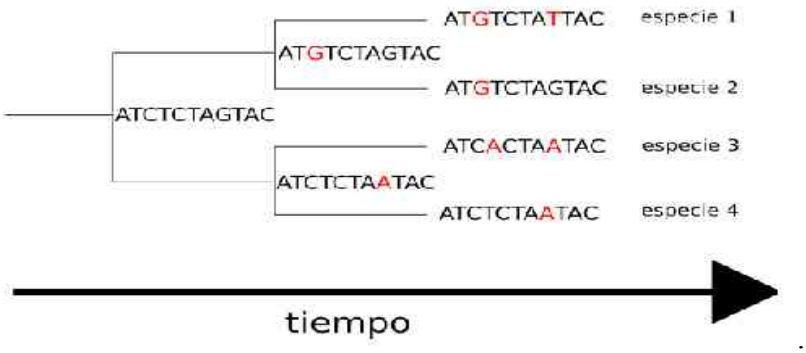

Figura 18: Mutaciones sobre secuencias en las ramas de una filogenia molecular.

Conocer los mecanismos que alteran las secuencias génicas a lo largo de la historia es importante para poder modelarlos. A continuación presentamos algunos de ellos:

- Tasas diferentes de transversiones y transiciones. Si trabajamos con secuencias de aminoácidos solamente se tratan explícitamente en modelos basados en codones.

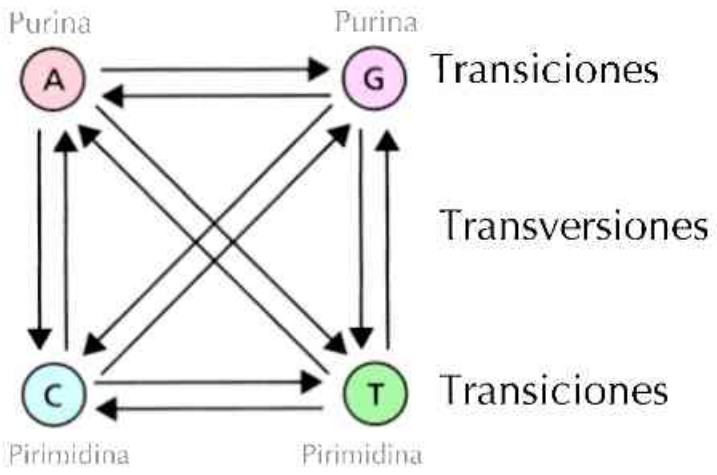

Figura 19: Transiciones y transversiones.

- Frecuencias de aminoácidos no equiprobables, representadas por el parámetro F.

- Tasas de sustitución (de matrices como BLOSUM) no homogéneas a lo largo de la secuencia. Esto se incorpora a los modelos evolutivos por medio de dos parámetros:

    - la proporción de sitios invariantes (I): la fracción de columnas conservadas
    - la distribución de tasas de sustitución entre sitios (G), que se representa por la forma de una función gamma.

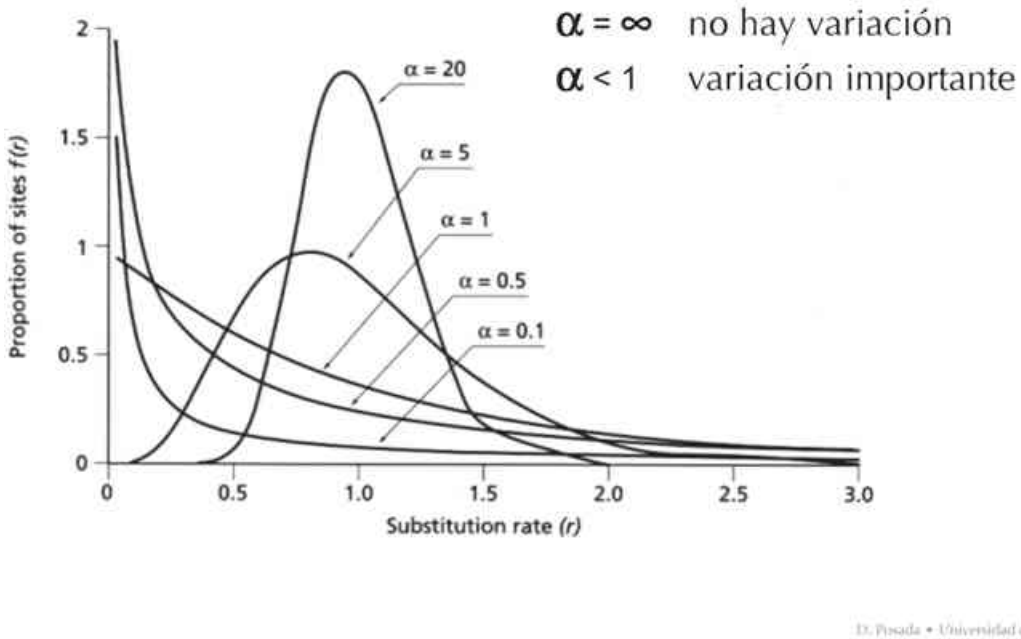

Figura 20: Distribución gamma de tasas de variación y su forma con distintos valores de alfa.

El software ProtTest (Darriba *et al.*, 2011) es una herramienta adecuada para la selección de modelos evolutivos de secuencias de proteínas, y en su versión 3 soporta las siguientes matrices de sustitución, [WAG, Dayhoff, JTT, mtREV, MtMam, MtArt, VT, RtREV, CpREV, Blosum62, LG, DCMut, HIVw/HIVb, FLU], que combinados con los parámetros F, I y G en total generan 120 modelos evolutivos.

A parte de este tipo de modelos se han propuesto en la literatura modelos con base termodinámica o mecánico-estadística que en vez de centrarse en la historia evolutiva aproximan la influencia que tiene una sustitución sobre el plegamiento de la proteína.

Cómo se infiere una filogenia a partir de secuencias de proteínas

Si estamos trabajando con secuencias de especies que han divergido lo suficiente entre si podremos seguramente estimar filogenias a partir de secuencias de proteínas, en caso contrario es posible que las secuencias genómicas proporcionen mejor señal.

En cualquier caso, un protocolo genérico incluiría las siguientes etapas:

- Búsqueda de secuencias homólogas de una proteína problema de interés. Esta tarea puede completarse obteniendo una predicción de estructura secundaria y posibles regiones desordenadas a partir de los homólogos encontrados.

- Alineamiento múltiple de la familia de proteínas resultante.

- Determinación del modelo evolutivo más verosímil. Por qué es esto importante? Porque distintos modelos en ocasiones pueden producir topologías diferentes, como se ve en la siguiente figura:

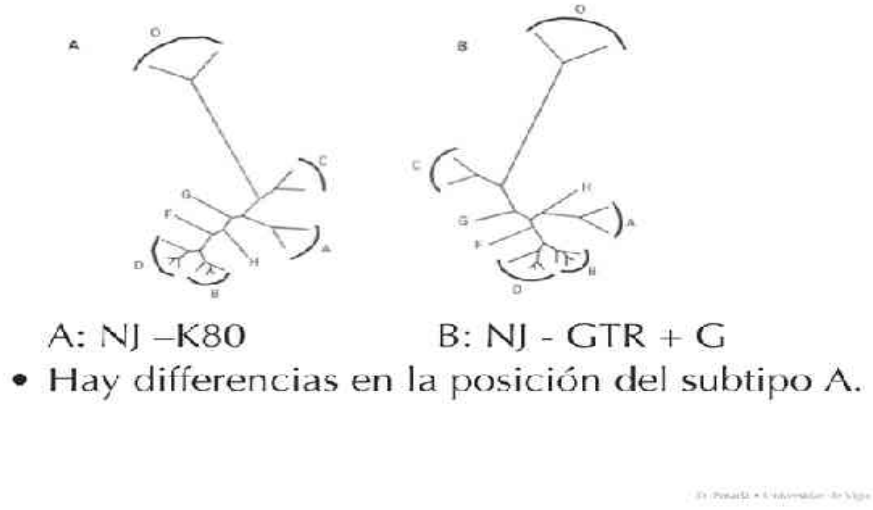

Figura 21: Relación entre dos modelos evolutivos alternativos y topologías obtenidas.

- Reconstrucción de una filogenia molecular en base a las secuencias alineadas y el modelo seleccionado. Hay una gran variedad de estrategias, que se explican en los enlaces al material del Dr. Pablo Vinuesa, de la UNAM:

    - Métodos basados en matrices de distancias.

    - Algoritmos basados en la búsqueda de árboles por parsimonia.

    - Métodos basados en la selección de modelos y topologías por máxima verosimilitud (ML).

    - Inferencia bayesiana de filogenias.

Ejercicios

Para los siguientes ejercicios podemos trabajar con cualquier secuencia que se os ocurra o alguna de éstas:

```
>Homeobox protein hox-b1
MEPNTPTARTFDWMKVKRNPPKTAKVSEPGLGSPSGLRTNFTTRQLTELEKEFHFNKYLSRARRVEIAATLELNETQVKI
WFQNRRMKQKKREREGG

>sp|P46310|FAD3C_ARATH Omega-3 fatty acid desaturase, chloroplastic OS=Arabidopsis
thaliana GN=FAD7 PE=1 SV=1
MANLVLSECGIRPLPRIYTTPRSNFLSNNNKFRPSLSSSSYKTSSSPLSFGLNSRDGFTR
NWALNVSTPLTTPIFEESPLEEDNKQRFDPGAPPPFNLADIRAAIPKHCWVKNPWKSLSY
VVRDVAIVFALAAGAAYLNNWIVWPLYWLAQGTMFWALFVLGHDCGHGSFSNDPKLNSVV
GHLLHSSILVPYHGWRISHRTHHQNHGHVENDESWHPMSEKIYNTLDKPTRFFRFTLPLV
MLAYPFYLWARSPGKKGSHYHPDSDLFLPKERKDVLTSTACWTAMAALLVCLNFTIGPIQ
MLKLYGIPYWINVMWLDFVTYLHHHGHEDKLPWYRGKEWSYLRGGLTTLDRDYGLINNIH
HDIGTHVIHHLFPQIPHYHLVEATEAAKPVLGKYYREPDKSGPLPLHLLEILAKSIKEDH
YVSDEGEVVYYKADPNLYGEVKVRAD

>YY1
MEPRTIACPHKGCTKMFRDNSAMRKHLHTHGPRVHVCAECGKAFVESSKLKRHQLVHTGEKPFQCTFEGCGKRFSLDFNL
RTHVRIHTGDRPYVCPFDGCNKKFAQSTNLKSHILTHAKAKNNQ
```

Resolveremos las siguientes tareas con ayuda de herramientas en la Web, pero todas ellas están disponibles como aplicaciones que podemos instalar en ordenadores, preferiblemente bajo sistemas Linux.

Búsqueda de secuencias homólogas de una proteína problema

Esta tarea consiste en buscar secuencias homólogas a una proteína problema con el fin de obtener, en primer lugar, la materia prima para hacer un alineamiento múltiple, y en segundo lugar, para obtener anotaciones funcionales e información útil para el alineamiento como la estructura secundaria o la composición de dominios.

1. En principio podemos explorar la base de datos de secuencias de proteína de alta calidad UniProt o, más generalmente, el conjunto no redundante (nr) del NCBI. A menudo obtendrás muchas secuencias redundantes y posiblemente tendrías que filtrarlas para los pasos siguientes. Para eso es muy útil el software HHblits, que permite hacerlo de manera objetiva y automática.

2. Conocer la estructura secundaria de nuestra secuencia puede ayudarnos a alinearla correctamente, por ejemplo haciendo que las alfahélices coincidan. Podemos predecirla en 3 estados (H,E y C) con una precisión a nivel de residuo del orden del 75-80% con programas como PSIPRED. Por esta razón repetiremos nuestra siguiente búsqueda de homólogos con el programa HHblits, porque nos permite explorar las mismas bases de datos mencionadas y también obtener su estructura secundaria aproximada.
    - Prueba a visualizar con Jalview el alineamiento completo o el representativo. Hay que descartar secuencias incompletas? Valora los elementos de estructura secundaria en el contexto del alineamiento.
    - Exporta las secuencias encontradas, completas, a un fichero en formato FASTA.
    - Prueba a cambiar el modo de alineamiento a global. Qué diferencias observas?

3. Para conocer de manera sencilla de que dominios se compone la secuencia podemos usar Pfam. Qué dominios encuentras en tu secuencia?

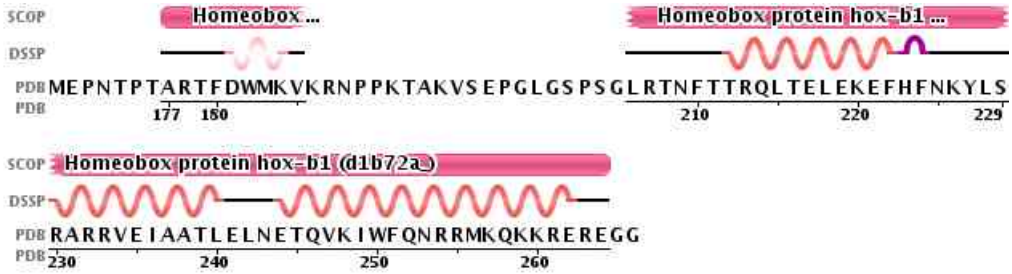

Figura 22: Estructura secundaria de hox-b1, tomada del Protein Data Bank.

Alineamiento múltiple de la familia de proteínas

Es posible que como resultado de las tareas anteriores tengas ya un conjunto de secuencias alineadas, pero a menudo será necesario calcular un alineamiento nuevo, por ejemplo, para mezclar secuencias obtenidas de diferentes fuentes o búsquedas.

1. Si es necesario realineamos con Clustal omega y exportamos el resultado.

2. Convertimos el alineamiento obtenido, o el que ya teníamos, a formato Phylip. Se puede hacer con Readseq.

Determinación del modelo evolutivo mas verosímil

Para elegir un modelo evolutivo adecuado sugiero usar ProtTest, para lo cual debemos dar como entrada el alineamiento en formato Phylip.

1. Guarda y examina los resultados obtenidos. Selecciona los mejores modelos evolutivos obtenidos, la tabla tiene este aspecto, ordenada por defecto por AIC:

```
******************************************************
Maximum Likelihood (-lnL) framework
******************************************************
Best model according to -lnL: WAG+I+G+F
******************************************************
Model        deltaAIC     AIC          -lnL*        AICw
------------------------------------------------------
RtREV+I+G+F  0.96         5908.73      -2894.36     0.31
RtREV+G+F    2.25         5910.02      -2896.01     0.16
WAG+I+G+F    13.48        5921.25      -2900.63     0.00
WAG+G+F      16.46        5924.23      -2903.12     0.00
WAG+I+G      0.00         5907.77      -2912.89     0.49
WAG+G        5.29         5913.06      -2916.53     0.04
Blosum62+I+G 9.84         5917.61      -2917.81     0.00
CpREV+I+G+F  51.66        5959.43      -2919.71     0.00
(rest of lines omitted)
```

Reconstrucción de una filogenia molecular

En este punto ya tenemos todo lo necesario para reconstruir una filogenia molecular en phylogeny.fr en modo "A la carte":

1. Desmarca la opción de alineamiento múltiple porque ese paso ya lo hemos hecho previamente. Aprovecha para revisar el resto de opciones, aunque en principio las dejaremos todas por defecto. Presiona "Create Workflow" para crear tu propia tubería de análisis a medida.

2. En las opciones de "Phylogeny: PhyML" puedes elegir si hacer test de *boostrap* más lentos que aLRT, pero sobre todo debes elegir el modelo evolutivo más parecido al obtenido en ProtTest, recurriendo a las casillas de "Advanced Settings".

3. Guarda alguna figura de la filogenia obtenida, con todos los parámetros empleados, como si fuera una figura para una publicación. Un software que recomendamos para la preparación de arboles para su publicación es FigTree.

Bibliografía

Branden, C.-I. & Tooze, J. (**1999**).
Introduction to protein structure.
Garland Pub., New York, 2nd edition.

Chothia, C. & Lesk, A. M. (**1986**).
The relation between the divergence of sequence and structure in proteins.
EMBO J., 5(4):823-826.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1166865/.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. (**2011**).
ProtTest 3: fast selection of best-fit models of protein evolution.
Bioinformatics, 27(8):1164-1165.
URL
http://bioinformatics.oxfordjournals.org/content/27/8/1164.long.

Fitch, W. M. (**2000**).
Homology a personal view on some of the problems.
Trends Genet., 16(5):227-231.
URL http://www.ncbi.nlm.nih.gov/pubmed/10782117.

Gabaldon, T. & Koonin, E. V. (**2013**).
Functional and evolutionary implications of gene orthology.
Nat. Rev. Genet., 14(5):360-366.
URL http://www.ncbi.nlm.nih.gov/pubmed/23552219.

Hou, J., Jun, S. R., Zhang, C. & Kim, S. H. (**2005**).
Global mapping of the protein structure space and application in structure-based inference of protein function.
Proc. Natl. Acad. Sci. U.S.A., 102:3651-3656.
URL http://www.pnas.org/content/102/10/3651.

Illergard, K., Ardell, D. & Elofsson, A. (**2009**).
Structure is three to ten times more conserved than sequence-a study of structural response in protein cores.
Proteins, 77(3):499-508.
URL http://onlinelibrary.wiley.com/doi/10.1002/prot.22458/abstract.

Isom, D. G., Castaneda, C. A., Cannon, B. R., Velu, P. D. & Garcia-Moreno E, B. (**2010**).
Charges in the hydrophobic interior of proteins.
Proc. Natl. Acad. Sci. U.S.A., 107:16096-16100.
URL http://www.pnas.org/content/107/37/16096.abstract.

King, N. P., Jacobitz, A. W., Sawaya, M. R., Goldschmidt, L. & Yeates, T. O. (**2010**).
Structure and folding of a designed knotted protein.
Proc. Natl. Acad. Sci. U.S.A., 107:20732-20737.

URL http://www.pnas.org/content/107/48/20732.abstract.

Kosloff, M. & Kolodny, R. (**2008**).
Sequence-similar, structure-dissimilar protein pairs in the PDB.
Proteins, 71:891-902.
URL http://onlinelibrary.wiley.com/doi/10.1002/prot.21770/abstract.

Needleman, S. B. & Wunsch, C. D. (**1970**).
A general method applicable to the search for similarities in the amino acid sequence of two proteins.
J.Mol.Biol., 48:443-453.
URL
http://www.sciencedirect.com/science/article/pii/0022283670900574.

Pascual-García, A., Abia, D., Ortiz, A. & Bastolla, U. (**2009**).
Cross-Over between Discrete and Continuous Protein Structure Space: Insights into Automatic Classification and Networks of Protein Structures.
PLoS Computational Biology, 5(3):e1000331.
URL http://dx.plos.org/10.1371/journal.pcbi.1000331.

Porter, L. L. & Rose, G. D. (**2012**).
A thermodynamic definition of protein domains.
Proc. Natl. Acad. Sci. U.S.A., 109(24):9420-9425.
URL http://www.pnas.org/content/109/24/9420.long.

Potestio, R., Micheletti, C. & Orland, H. (**2010**).
Knotted vs. unknotted proteins: evidence of knot-promoting loops.
PLoS Comput. Biol., 6:e1000864.
URL http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000864.

Schnabel, J. (**2010**).
Protein folding: The dark side of proteins.
Nature, 464:828-829.
URL http://www.nature.com/news/2010/100407/full/464828a.html.

Smith, T. & Waterman, M. (**1981**).
Identification of Common Molecular Subsequences.
J.Mol.Biol., 147:195-197.
URL
http://www.sciencedirect.com/science/article/pii/0022283681900875.

Taylor, W. R. (**2002**).
A 'periodic table' for protein structures.
Nature, 416:657-660.
URL http://www.ncbi.nlm.nih.gov/pubmed/11948354.

Wuerges, J., Garau, G., Geremia, S., Fedosov, S. N., Petersen, T. E. & Randaccio, L. (**2006**).
Structural basis for mammalian vitamin B12 transport by transcobalamin.
Proc. Natl. Acad. Sci. U.S.A., 103:4386-4391.
URL http://www.pnas.org/content/103/12/4386.long.

Zhen, R. G., Kim, E. J. & Rea, P. A. (**1997**).
Acidic residues necessary for pyrophosphate-energized pumping and inhibition of the vacuolar H+-pyrophosphatase by N,N'-dicyclohexylcarbodiimide.
J. Biol. Chem., 272:22340-22348.

URL http://www.jbc.org/content/272/35/22340.long.

Sobre este documento...

Alineamiento de secuencias de proteína y filogenias

This document was generated using the **LaTeX**2HTML translator Version 2008 (1.71)

The command line arguments were:
latex2html `alineafilog -split 0 -dir alineafilog1 -no_navigation`

The translation was initiated by Bruno Contreras Moreira on 2015-07-06

Bruno Contreras-Moreira e Inmaculada Yruela
http://www.eead.csic.es/compbio