Análisis de secuencias.
Familias de proteínas.

Curso de verano de
Bioinformática y Biología Computacional
de la UCM, Madrid 2005

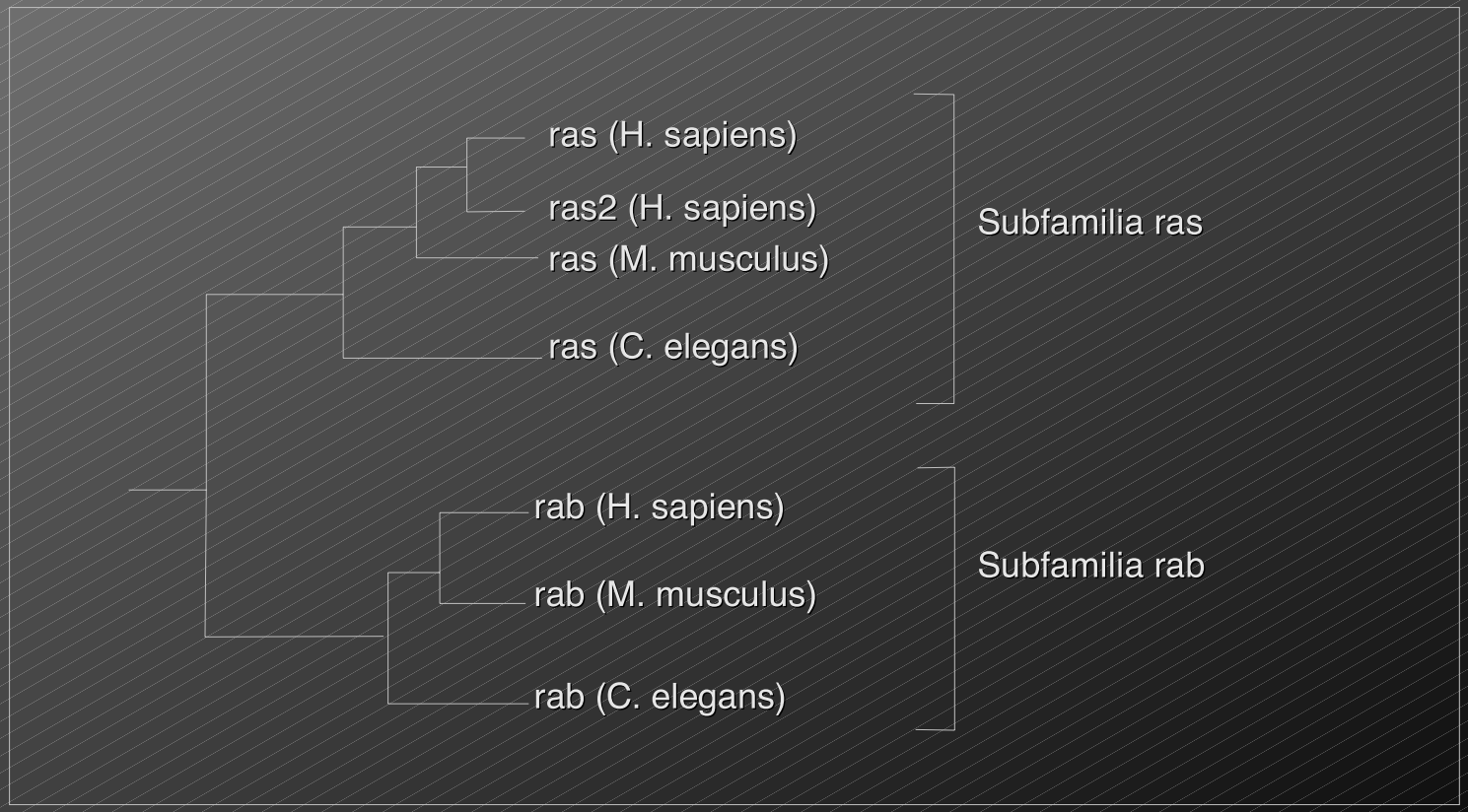
Federico Abascal
Museo Nacional de Ciencias Naturales

Lo que encontramos en las bases de datos

Observación: las proteínas homólogas pueden tener funciones distintas.

Hipótesis: duplicación génica, barajado de dominios y divergencia dan lugar a nuevas familias de proteínas con nuevas funciones.

Observación (concordante con la hipótesis): las proteínas con una misma función (misma familia) están más cercanas evolutivamente entre sí.



Guión de la charla. Familias de proteínas.

-Las proteínas homólogas pueden tener funciones distintas.

  -domain-shuffling

  -ortólogos y parálogos

  -superfamilias, familias y subfamilias

-¿Por qué analizar la organización en familias de las proteínas?

-Algunas aproximaciones y bases de datos para la clasificación de proteínas

  -PFam y Prosite

  -InterPro

  -Protomap

  -COGs

Guión de la charla. Familias de proteínas.

-Las proteínas homólogas pueden tener funciones distintas.

- -domain-shuffling
- -ortólogos y parálogos
- -superfamilias, familias y subfamilias

-¿Por qué analizar la organización en familias de las proteínas?

-Algunas aproximaciones y bases de datos para la clasificación de proteínas

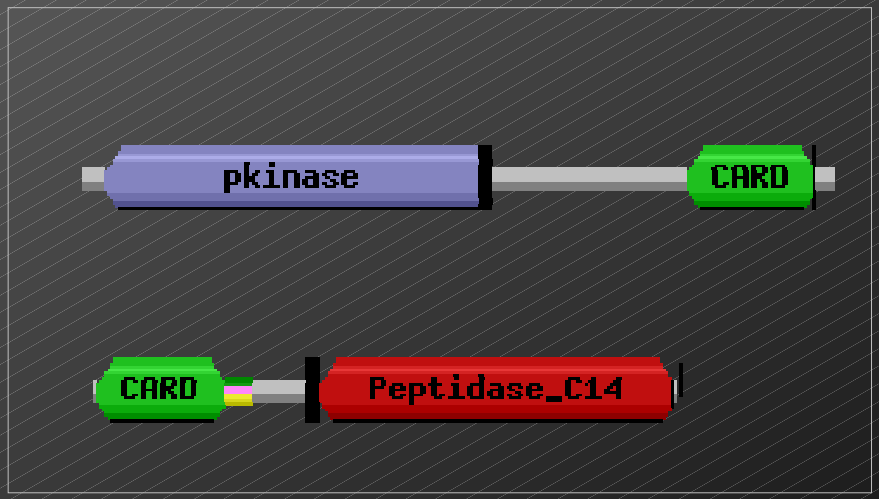
- -PFam y Prosite
- -InterPro
- -Protomap
- -COGs

Barajado de dominios (domain-shuffling)

Observación: las proteínas homólogas pueden tener diferente organización de dominios.

El dominio, y no el gen, es la unidad evolutiva básica.

- La función de una proteína es el resultado de las funciones de sus dominios.
- Las propiedades de las proteínas pueden ser explicadas, pero no deducidas, a partir de sus dominios.



Homólogos: ortólogos y parálogos.

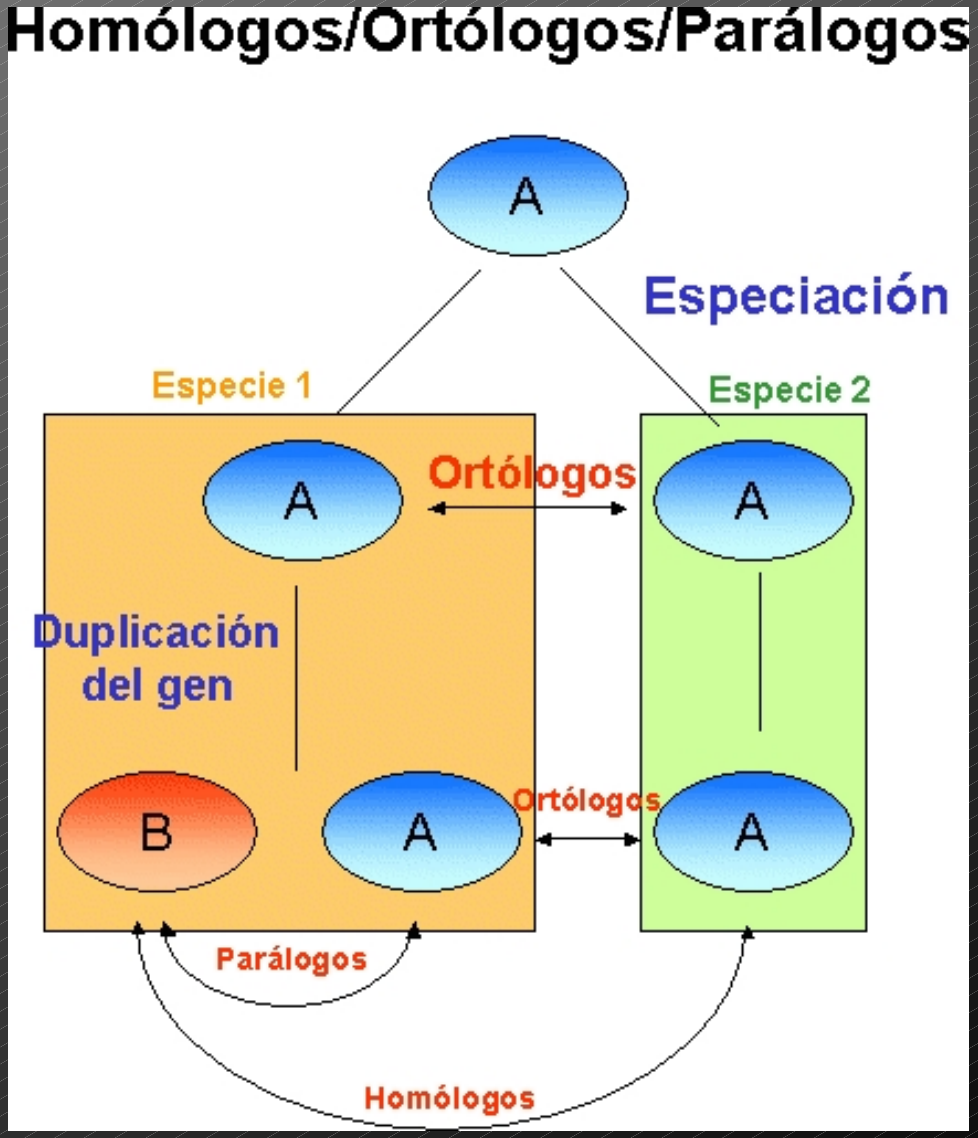
Ortólogos: genes que comparten el último ancestro común y cuya divergencia se debe a la especiación.

Ejemplo: isomerasa de glucosa-6P de Bacillus subtilis y de Escherichia coli.

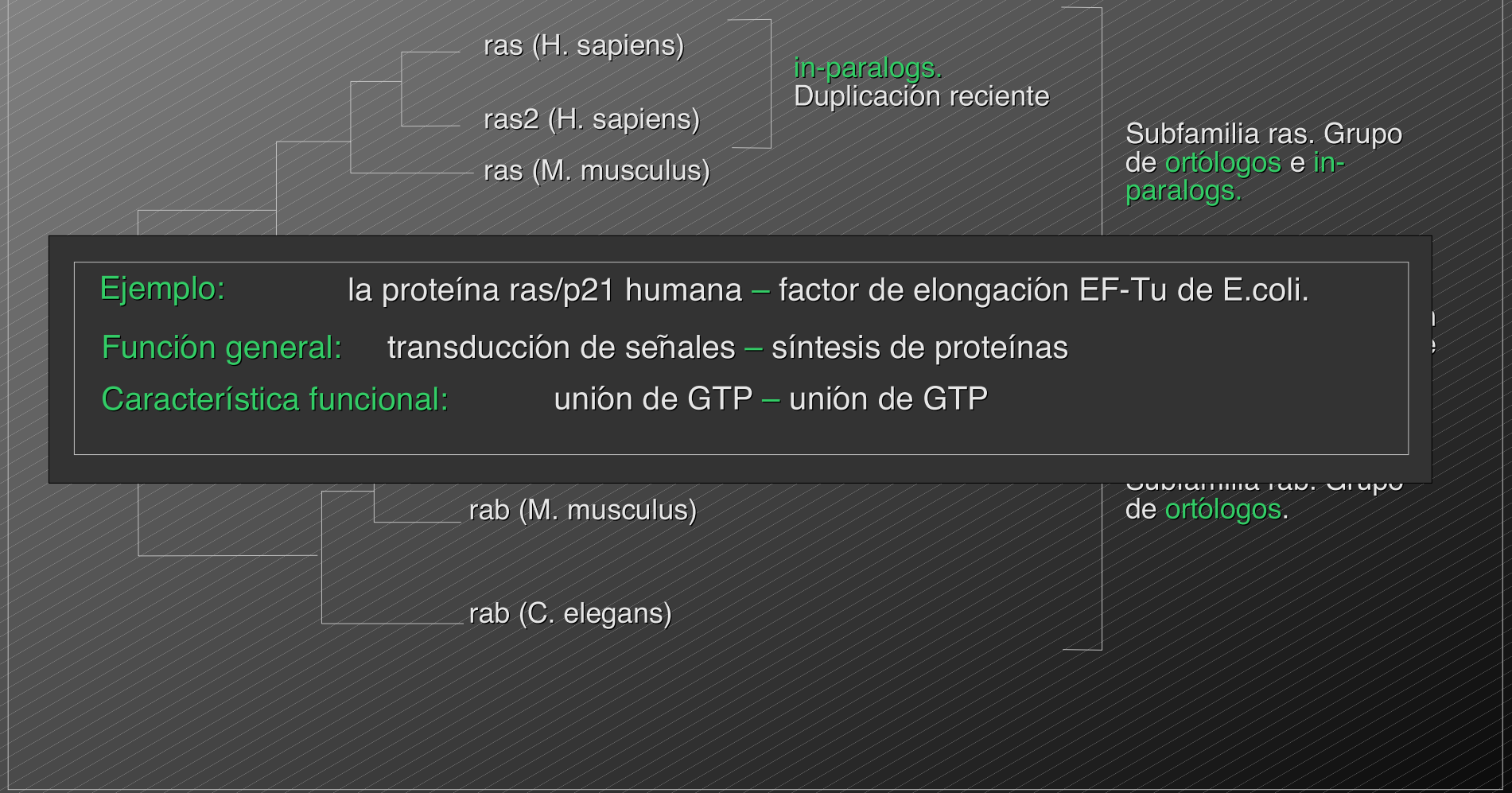
Los mismos genes en distintas especies.

Parálogos: genes que debido a una duplicación, ya no comparten el último ancestro. Frecuentemente tienen funciones distintas.

Ejemplo: tripsina, quimiotripsina, elastasa y trombina.



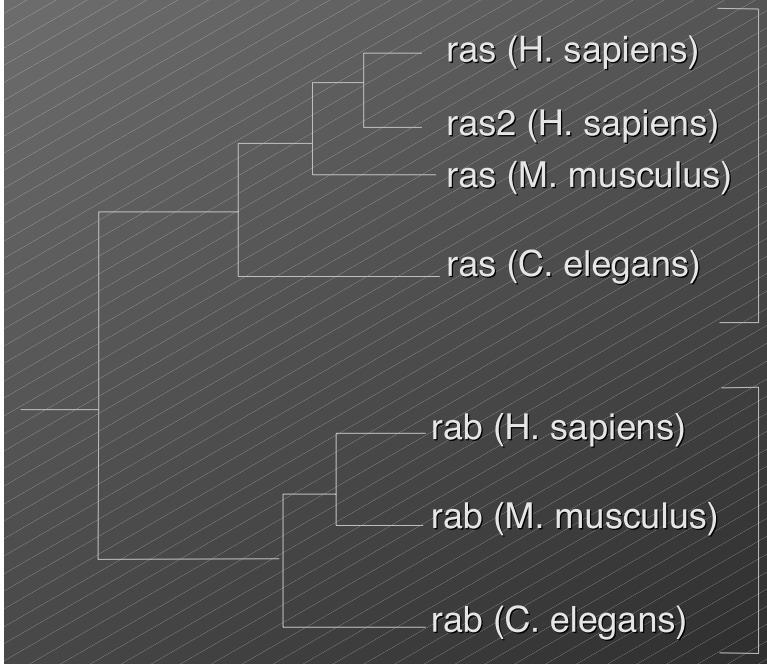
Homólogos: ortólogos y parálogos.

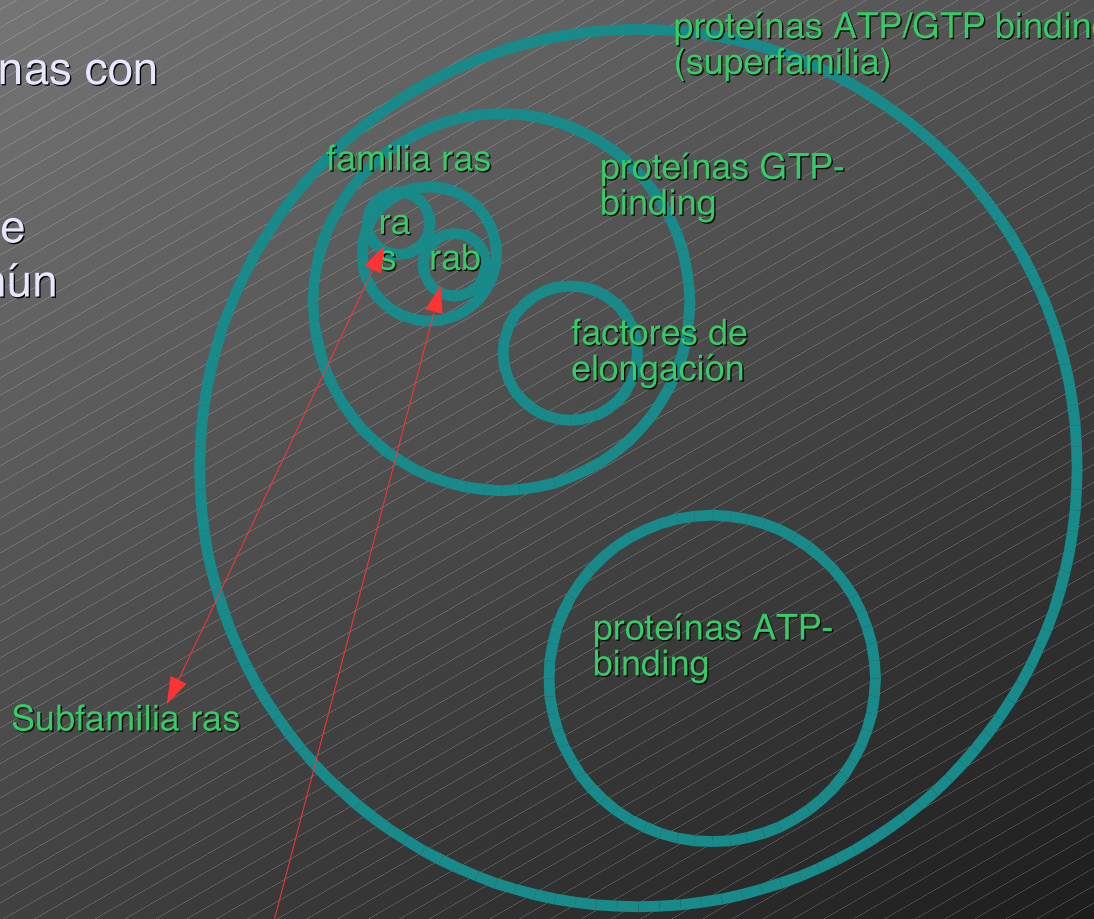


Homólogos: superfamilias, familias y subfamilias.

Superfamilia: grupo de proteínas con un origen común.

Familia / **Subfamilia**: grupo de proteínas con una función común (jerarquía subjetiva).





Dos formas de representarlo

Guión de la charla. Familias de proteínas.

-Las proteínas homólogas pueden tener funciones distintas.

- -domain-shuffling
- -ortólogos y parálogos
- -superfamilias, familias y subfamilias

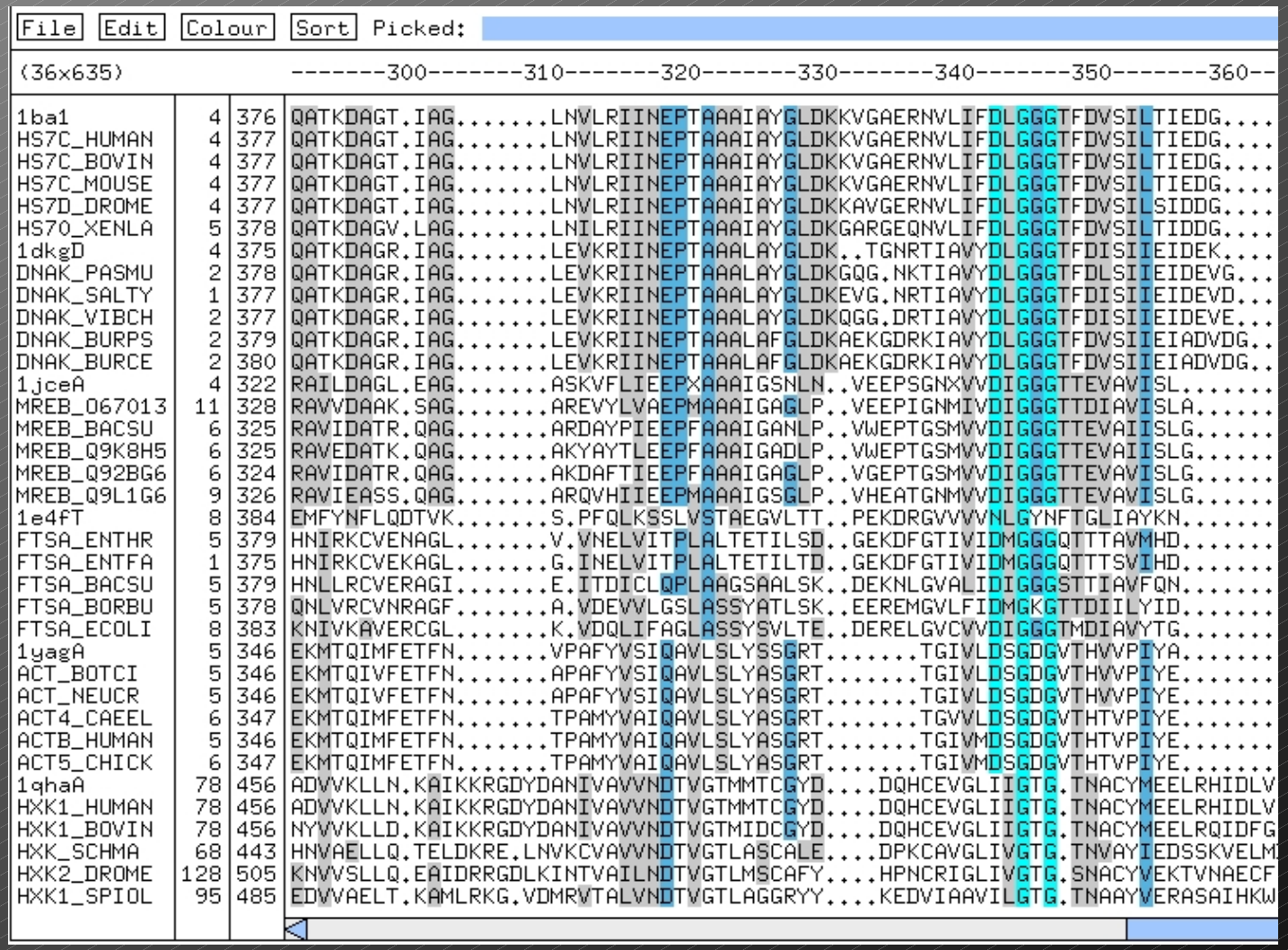
-¿Por qué analizar la organización en familias de las proteínas?

-Algunas aproximaciones y bases de datos para la clasificación de proteínas

- -PFam y Prosite
- -InterPro
- -Protomap
- -COGs

Interés de analizar la organización en familias de las proteínas

Predicción de función.



chaperones (dnak), proteínas implicadas en la formación del septo bacteriano (ftsA, mreB), hexokinasas (hxk), actina (act)...

Cómo analizar la organización en familias de las proteínas

Árboles filogenéticos: lo más fiable, pero es laborioso y hay que hacerlo manualmente

(lo veréis el próximo día)

Bases de datos construidas por expertos:

- Pfam
- Prosite
- InterPro
- ...

Métodos automáticos:

- ProtoMap
- COGs
- ...

Guión de la charla. Familias de proteínas.

-Las proteínas homólogas pueden tener funciones distintas.

  -domain-shuffling

  -ortólogos y parálogos

  -superfamilias, familias y subfamilias

-¿Por qué analizar la organización en familias de las proteínas?

-Algunas aproximaciones y bases de datos para la clasificación de proteínas

  -PFam y Prosite

  -InterPro

  -Protomap

  -COGs

Prosite

PROSITE:
http://us.expasy.org/prosite/

-caracterizan motivos conocidos con expresiones regulares y/o perfiles.

-gran cantidad de información para cada familia de proteínas.

-baja cobertura: sólo 1.245 familias

```
ID   MOLYBDOPTERIN_EUK; PATTERN.
AC   PS00559;
DT   DEC-1991 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE   Eukaryotic molybdopterin oxidoreductases signature.
PA   [GA]-x(3)-[KRNQHT]-x(11,14)-[LIVMFYWS]-x(8)-[LIVMF]-x-C-x(2)-[DEN]-R-
PA   x(2)-[DE].
NR   /RELEASE=38,80000;
NR   /TOTAL=50(50); /POSITIVE=45(45); /UNKNOWN=0(0); /FALSE_POS=5(5);
NR   /FALSE_NEG=2; /PARTIAL=5;
CC   /TAXO-RANGE=??E??; /MAX-REPEAT=1;
DR   P48034, ADO_BOVIN , T; Q06278, ADO_HUMAN , T; P11832, NIA1_ARATH, T;
DR   P39867, NIA1_BRANA, T; P27967, NIA1_HORVU, T; P16081, NIA1_ORYSA, T;
DR   P39865, NIA1_PHAVU, T; P54233, NIA1_SOYBN, T; P11605, NIA1_TOBAC, T;
DR   P11035, NIA2_ARATH, T; P39868, NIA2_BRANA, T; P27969, NIA2_HORVU, T;
DR   P39866, NIA2_PHAVU, T; P39870, NIA2_SOYBN, T; P08509, NIA2_TOBAC, T;
DR   P49102, NIA3_MAIZE, T; P27968, NIA7_HORVU, T; P36858, NIA_ASPNG , T;
DR   P43100, NIA_BEABA , T; P27783, NIA_BETVE , T; P43101, NIA_CICIN , T;
DR   P17569, NIA_CUCMA , T; P22945, NIA_EMENI , T; P39863, NIA_FUSOX , T;
DR   P36842, NIA_LEPMC , T; P39869, NIA_LOTJA , T; P17570, NIA_LYCES , T;
DR   P08619, NIA_NEUCR , T; P36859, NIA_PETHY , T; P49050, NIA_PICAN , T;
DR   P23312, NIA_SPIOL , T; Q05531, NIA_USTMA , T; P36841, NIA_VOLCA , T;
DR   P07850, SUOX_CHICK, T; P51687, SUOX_HUMAN, T; Q07116, SUOX_RAT  , T;
DR   P80457, XDH_BOVIN , T; P08793, XDH_CALVI , T; P47990, XDH_CHICK , T;
DR   P10351, XDH_DROME , T; P22811, XDH_DROPS , T; P91711, XDH_DROSU , T;
DR   P47989, XDH_HUMAN , T; Q00519, XDH_MOUSE , T; P22985, XDH_RAT   , T;
DR   P80456, ADO_RABIT , P; P17571, NIA1_MAIZE, P; P39871, NIA2_MAIZE, P;
DR   Q01170, NIA_CHLVU , P; P39882, NIA_LOTTE , P;
DR   P39864, NIA_PHYIN , N; Q12553, XDH_EMENI , N;
DR   P27034, BGLS_AGRTU, F; P03598, COAT_TOBSV, F; P19235, EPOR_HUMAN, F;
DR   P20054, PYR1_DICDI, F; Q23316, YHC6_CAEEL, F;
3D   1SOX;
DO   PDOC00484;
//
```

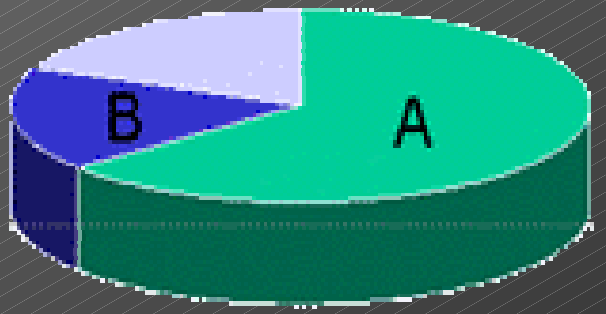
Pfam

Pfam: http://www.sanger.ac.uk/Pfam/

-caracterizan dominios de proteínas con perfiles HMM.

-gran cantidad de información.

-alta cobertura (7.316 familias, 73% swiss-prot y TrEMBL)



Rick:

Caspasa 9:

-Clasifican dominios y no proteínas completas (el dominio es la unidad evolutiva básica)

-Interfaz web muy útil:

- -alineamientos
- -distribución filogenética
- -organización de dominios
- -búsqueda usando perfiles-hmm
- -etc.

InterPro (I)

Interpro:
http://www.ebi.ac.uk/interpro/

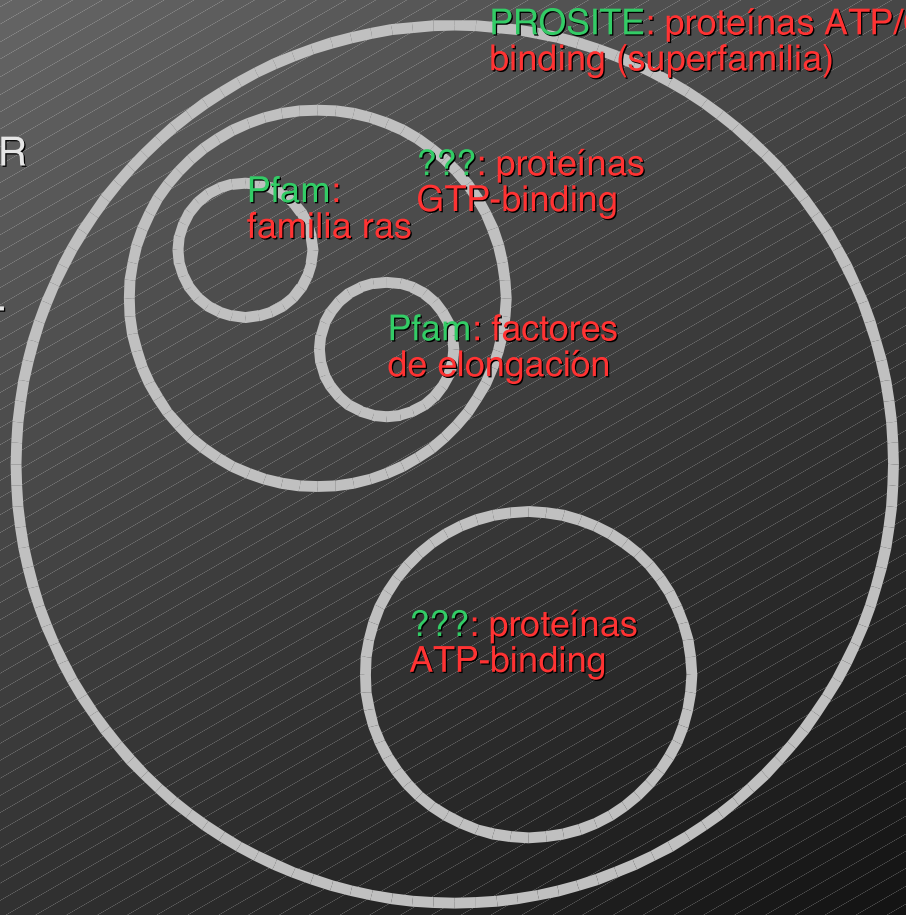
-para poner un poco de orden en el maremagnum de las bases de datos:
PROSITE, Pfam, Prints, PRODOM, Smart, PIR

-distingue entre dominios, familias, repeticiones, sitios de modificación post-transduccional...

-introduce jerarquía

-gran cantidad de información.

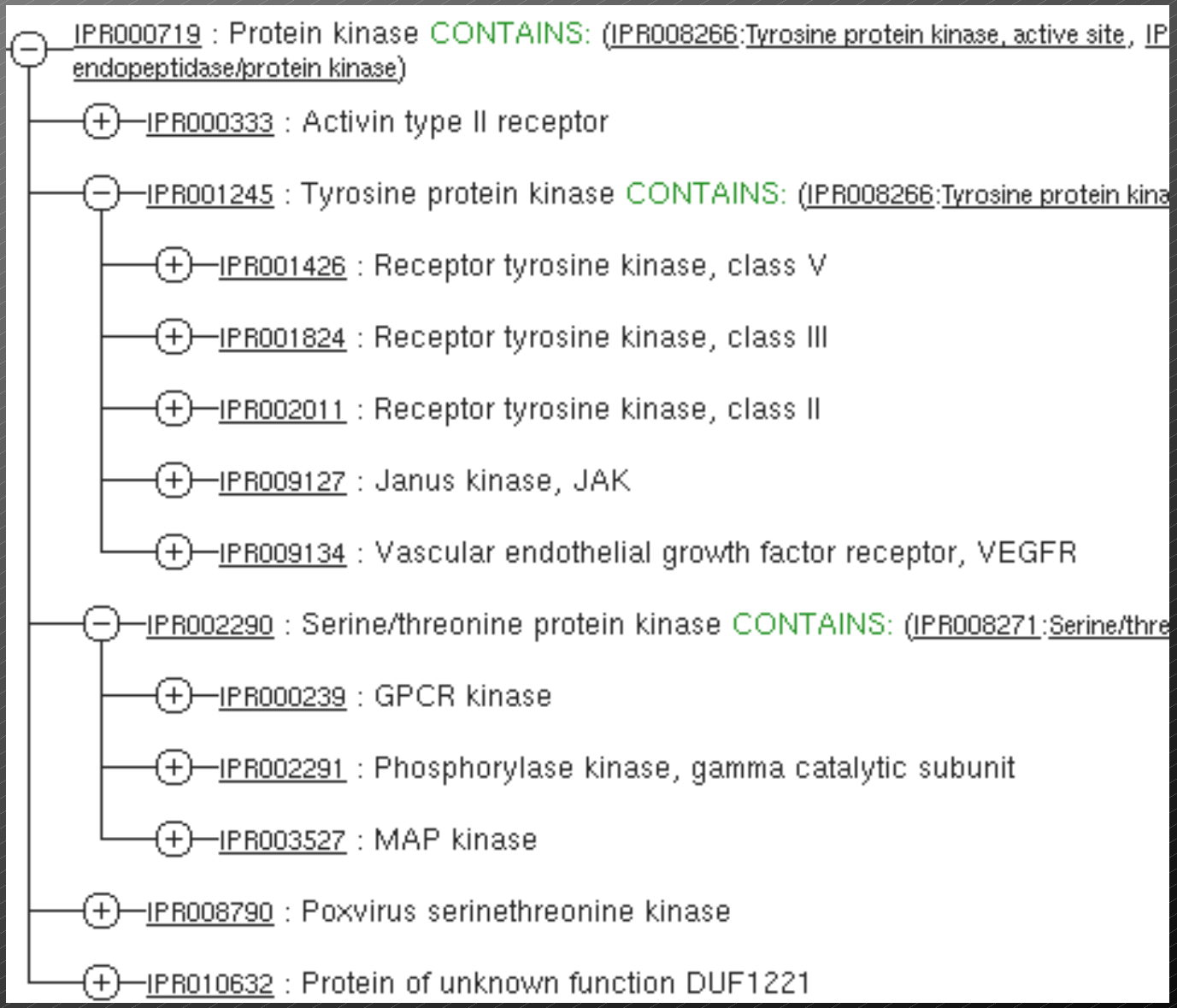
-alta cobertura.



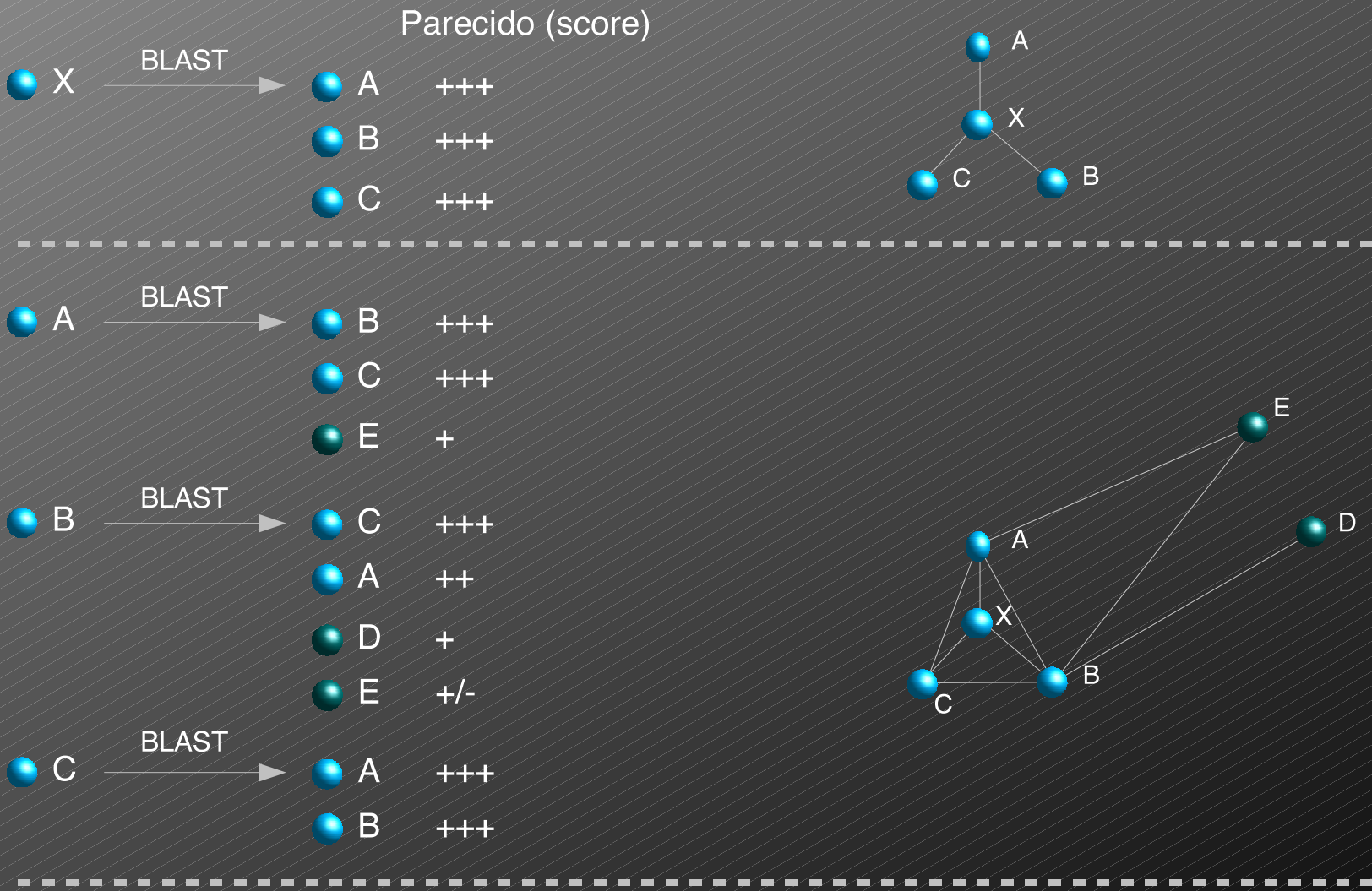
InterPro (II)

La jerarquía en InterPro:

ejemplo de las kinasas de proteínas.



ProtoMap (I)

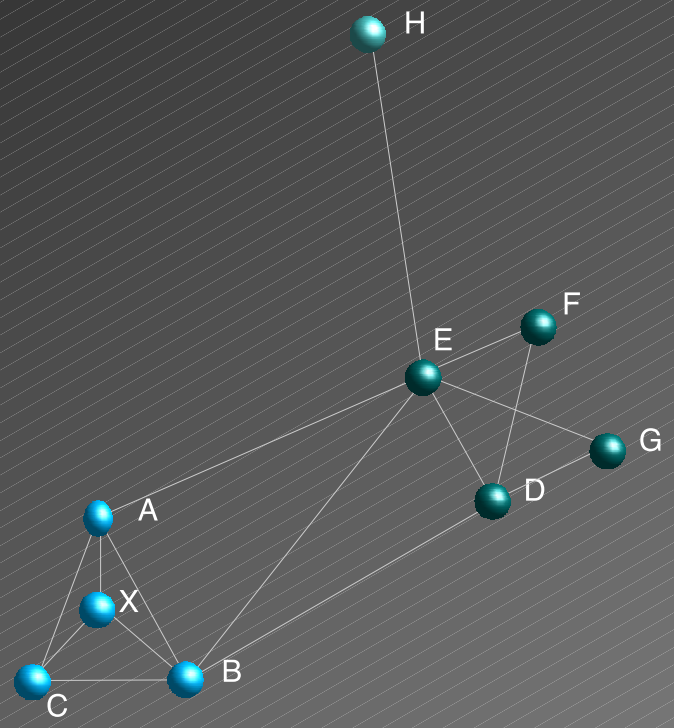

Parecido (score)
X
BLAST
A +++
B +++
C +++
A
X
C
B
A
BLAST
B +++
C +++
E +
B
BLAST
C +++
A ++
D +
E +/-
C
BLAST
A +++
B +++
E
D
A
X
B
C

ProtoMap (II)

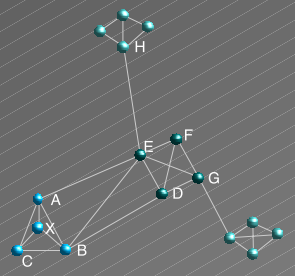
Parecido (score)

E BLAST →
- D +++
- F +++
- G +++
- B +
- A +
- H +

D BLAST →
- E +++
- F +++
- G +++
- B +/-



F, G, H, ... BLAST → etcétera

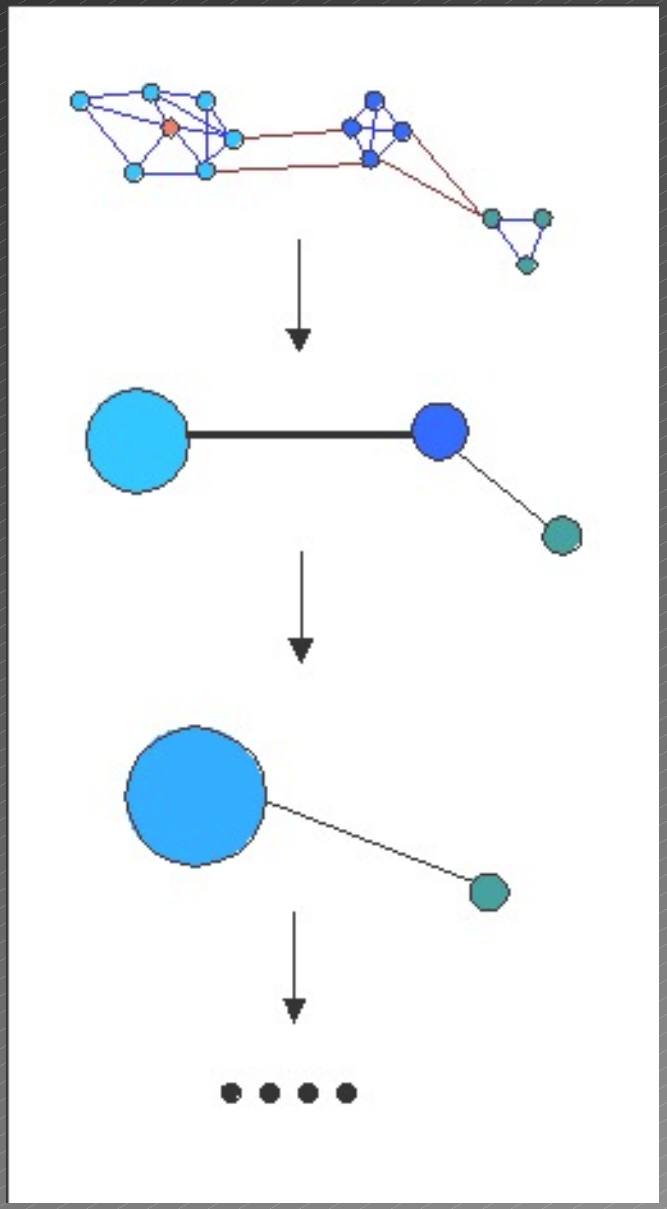


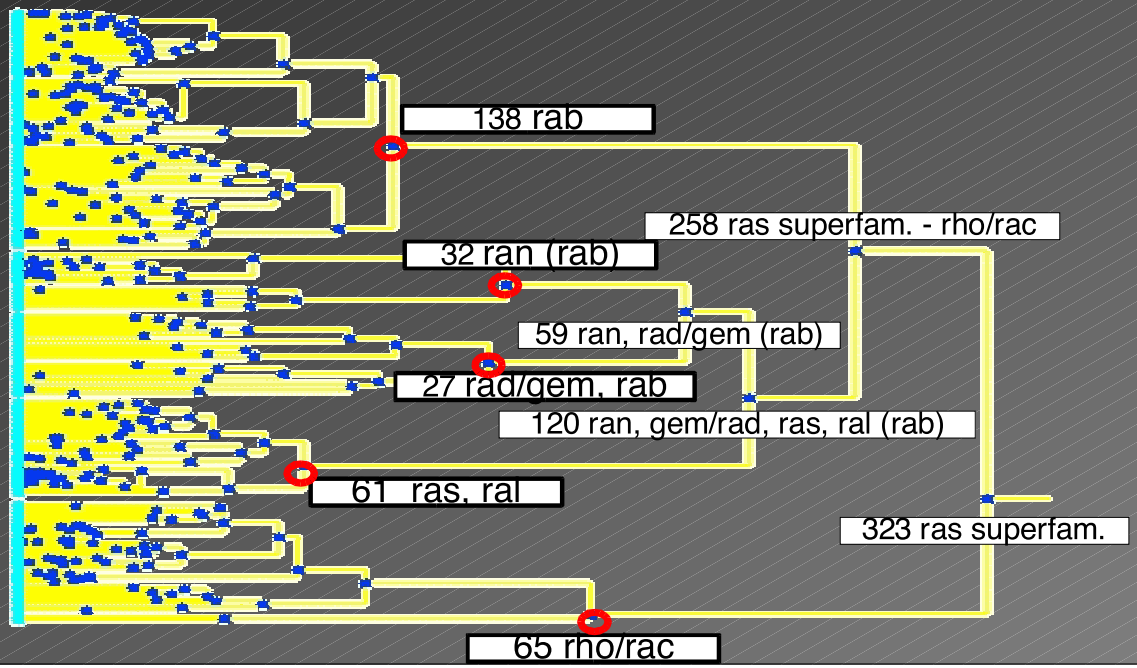
ProtoMap (III): el algoritmo en detalle

Recursivo.

0º.- Obtención de distancias entre secuencias => grafo
1º.- Agrupamiento de secuencias claramente relacionadas (e-value < 1e-100)
2º.- Inicialización de T = 1e-95.
3º.- Cálculo de distancias entre los distintos grupos o clusters:
-Se halla la media geométrica de los e-values entre cada par de clusters. En los casos en que no hay arcos, se asigna un e-value de 1.

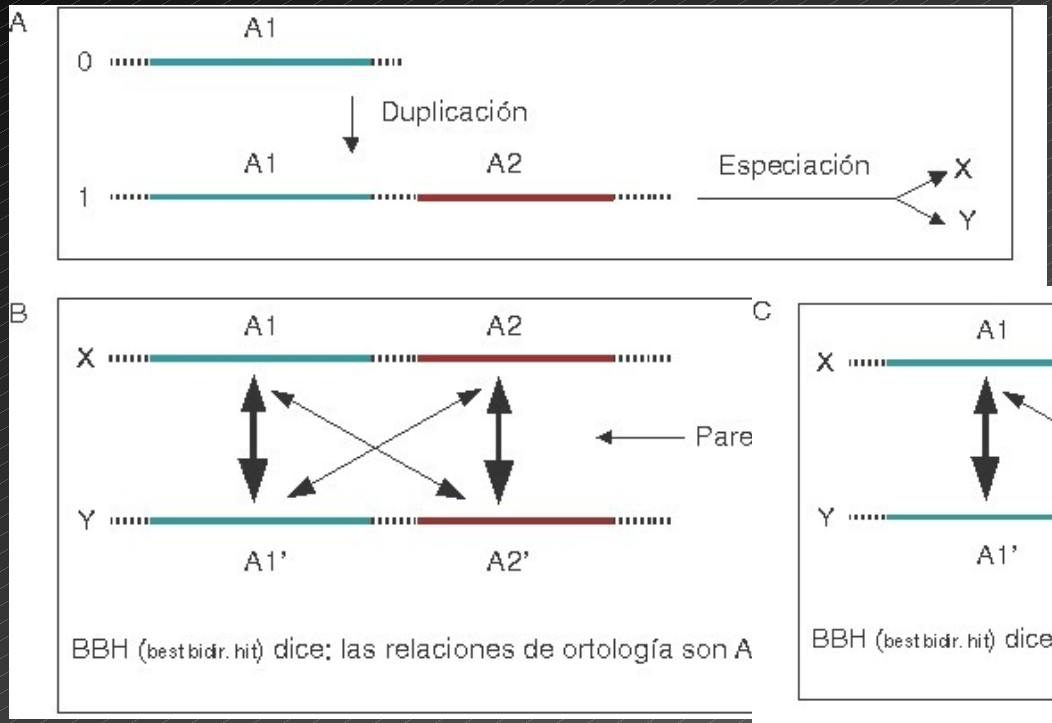
4º.- Unión de grupos vecinos: si la media geométrica de los e-values es menor que la raíz cuadrada de T, se unen los clusters.
5º.- Se relaja el umbral T: T = T*1e+05.
6º.- Si T > 1 => FIN.
Si no => se vuelve al punto 3º.





COGs: clasificación en grupos de ortólogos

Identificación de ortólogos basada en “Best Bidirectional Hits”

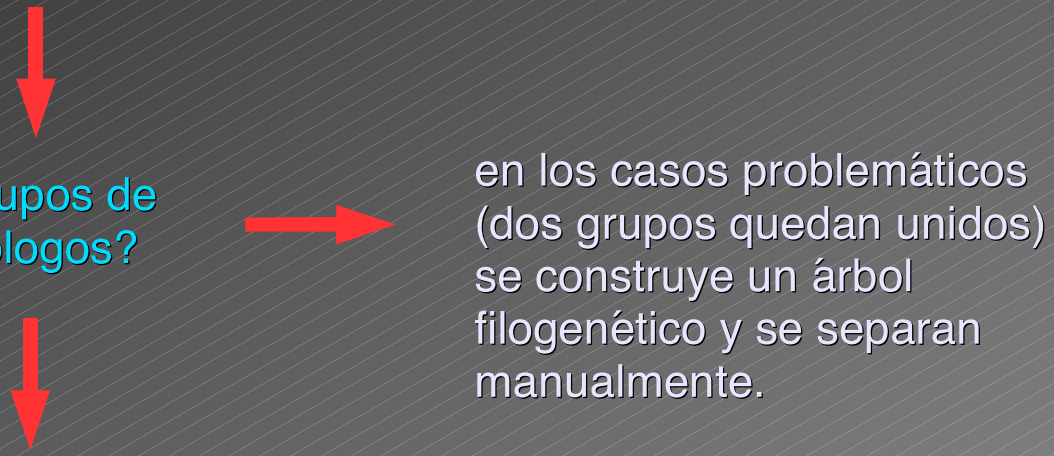


El BBH sólo es aplicable con genomas completos.

COGs: clasificación en grupos de ortólogos

Objetivo: clasificar las proteínas de microorganismos de los que se conoce el genoma completo.

Método(semiautomático):
1.- Identificación de BBH entre los genes de las distintas especies.
2.- Fusión de duplicaciones recientes (in-paralogs).
3.- Con las relaciones de BBH se construye un grafo.
4.- Identificación de triángulos en el grafo formados por especies de tres linajes distintos.
5.- Fusión de triángulos que comparten un lado.



Anotación funcional: función bioquímica, función general, rutas metabólicas...

COGs: clasificación en grupos de ortólogos

¿Qué se puede hacer con COGs?

- comparar genomas.
- buscar genes con un mismo patrón filogenético.
- estudiar el contexto genómico de un gen en distintas especies.
- buscar con una secuencia propia.
- etc, etc.

Versión previa de COGs: 44 genomas de microorganismos

Actualmente: 66 genomas de microorganismos y 7 de eucariotas

Guión de la charla. Familias de proteínas.

-Las proteínas homólogas pueden tener funciones distintas.

  -domain-shuffling

  -ortólogos y parálogos

  -superfamilias, familias y subfamilias

-¿Por qué analizar la organización en familias de las proteínas?

-Algunas aproximaciones y bases de datos para la clasificación de proteínas

  -PFam y Prosite

  -InterPro

  -Protomap

  -COGs

Agradecimientos

Algunas figuras han sido tomadas de...

-Paulino Gómez Puertas  Centro de Astrobiología

-Manuel José Gómez  Centro de Astrobiología