Análisis de secuencias.
Patrones, perfiles y dominios.

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

Federico Abascal
Museo Nacional de Ciencias Naturales

Recordatorio

Queremos comparar secuencias porque creemos que nos pueden hablar de la historia evolutiva de las proteínas, donde quizás podamos encontrar huellas de sus características funcionales y estructurales.

Para poder hacer esta comparación lo mejor posible: debemos encontrar el alineamiento que con mayor probabilidad (nunca sabremos si es el real) refleje qué cambios se han producido.

Limitación del alineamiento entre pares de secuencias

Problema: las mismas proteínas alinean de forma distinta según la matriz de sustitución y las penalizaciones por gaps utilizadas.

¿Cómo podemos saber cuál es el mejor alineamiento?

Observación: cuantas más secuencias, mayor cantidad de información, menor incertidumbre.

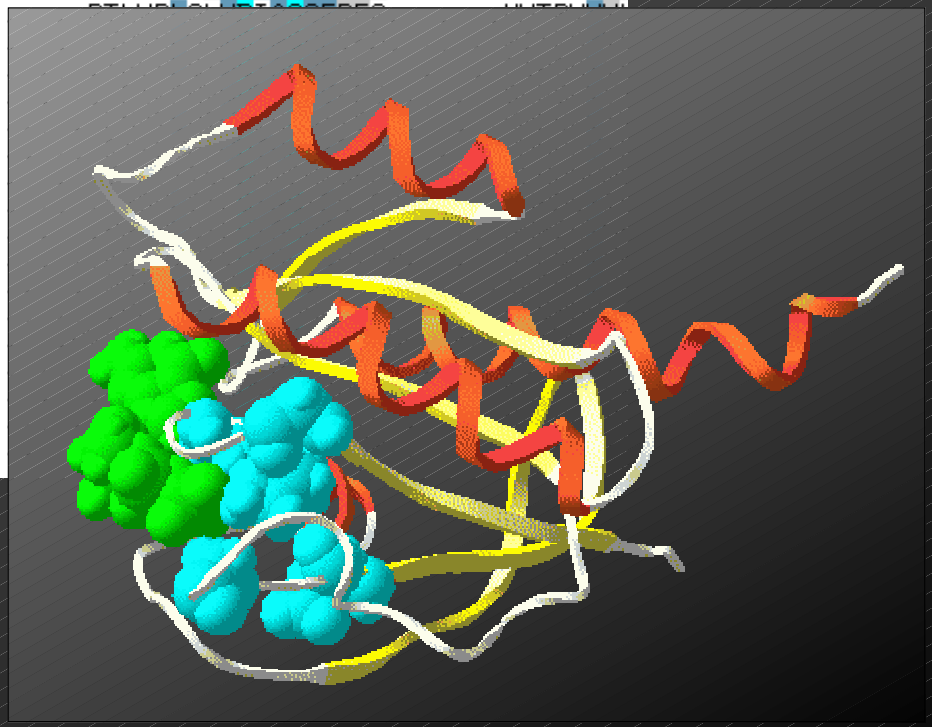
¿Cómo utilizar la información de muchas secuencias?

Construyendo un alineamiento múltiple.

```
# Matrix: BLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
DGHFVPNITLGQP (prot 1)
| |||.|.::...
D-HFVDNTVFQGE (prot 2)
  # Score: 296.0
```

```
# Matrix: BLOSUM45
# Gap_penalty: 10.0
# Extend_penalty: 0.5
DGHFVPN-ITLGQP (prot 1)
| |||.| :..|:.
D-HFVDNTVFQGEH (prot 2)
  # Score: 130.5
```

Alineamiento múltiple



Otra limitación de las comparaciones entre pares

Problema: si dos homólogos han divergido mucho (parecido < 20-25%), BLAST no es capaz de distinguir ese parecido del azar.

BLAST no es capaz de encontrar homólogos remotos

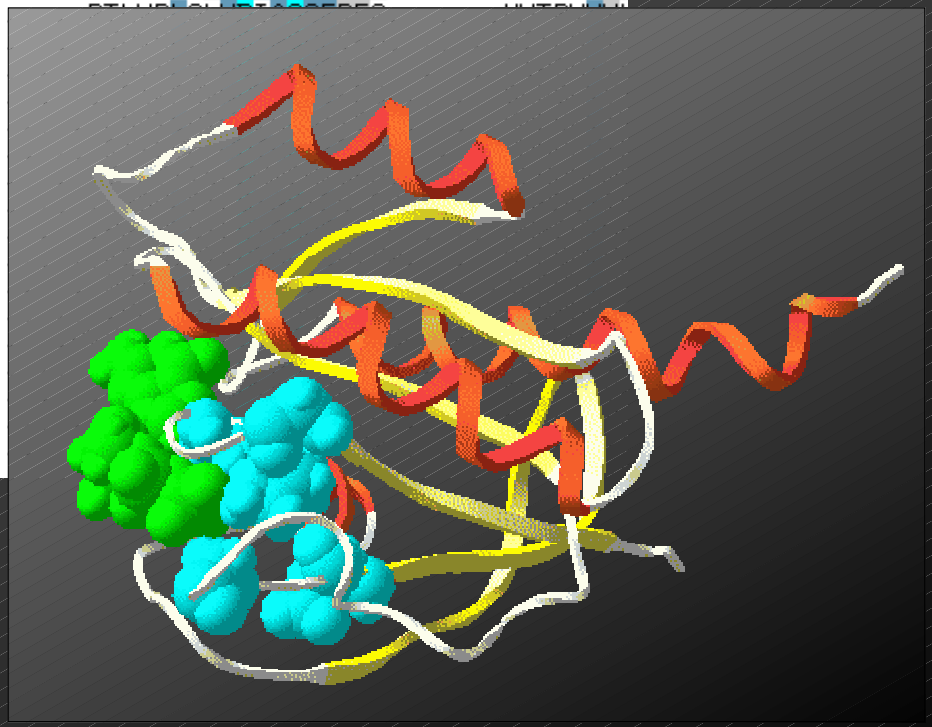
Observación: cuando hacemos un alineam. múltiple vemos qué posiciones son más importantes.

Idea: si las coincidencias en el alineamiento entre dos secuencias se producen en los sitios más importantes, la confianza en que sean homólogas ha de aumentar

Objetivo: utilizar la información de los alineam. múltiples para hacer búsquedas de homólogos más sensibles.

¿Cómo aprovechar la información del alineamiento múltiple?

Alineamiento múltiple



Guión de la charla. Patrones, perfiles y dominios.

-cómo utilizar la información de los alineamientos múltiples

- -secuencias consenso y expresiones regulares
- -perfiles y perfiles-hmm

-algunas bases de datos de patrones y perfiles:

- -Prosite
- -Pfam

-búsquedas en bases de datos:

- -PSI-BLAST
- -HMMer
- -búsqueda con secuencias intermedias

Guión de la charla. Patrones, perfiles y dominios.

-cómo utilizar la información de los alineamientos múltiples

- -secuencias consenso y expresiones regulares
- -perfiles y perfiles-hmm

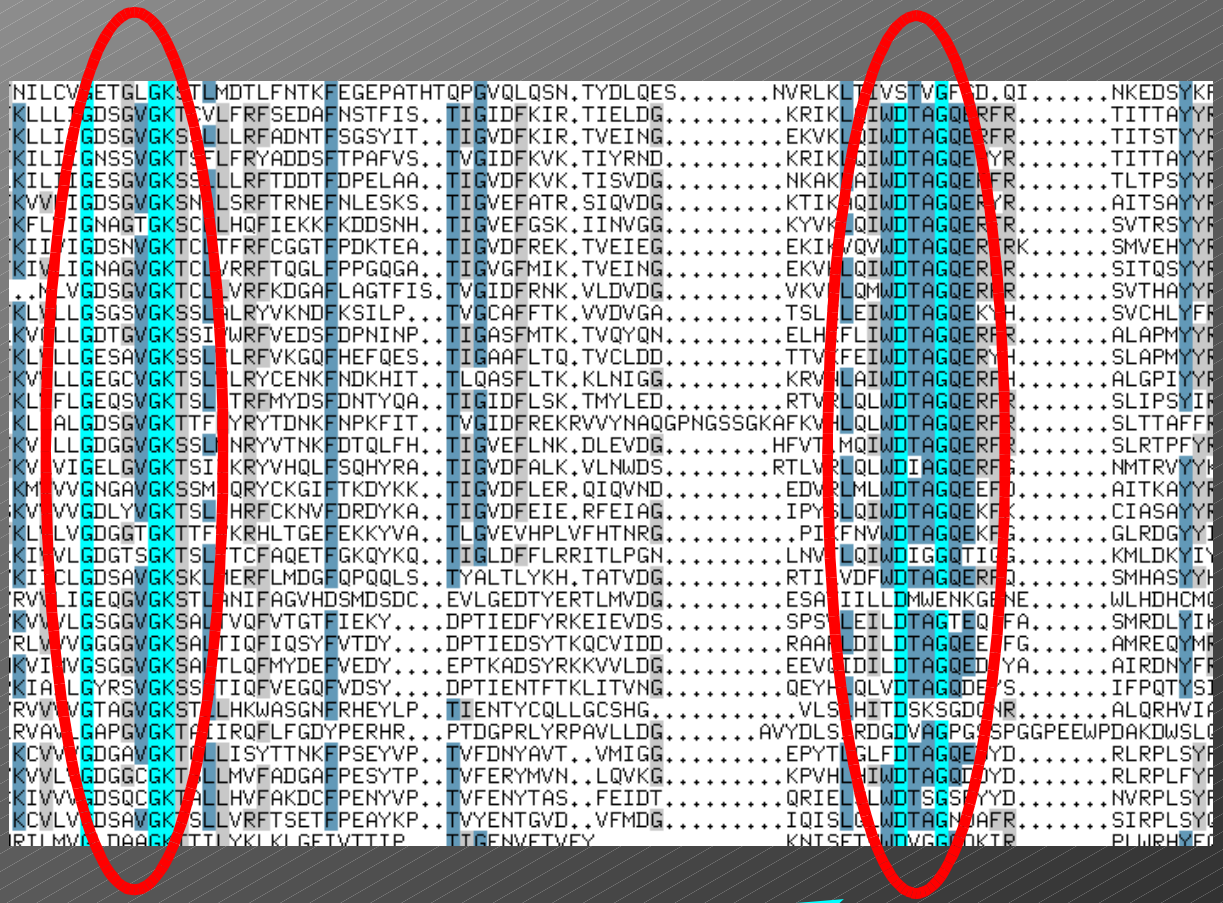
-algunas bases de datos de patrones y perfiles:

- -Prosite
- -Pfam

-búsquedas en bases de datos:

- -PSI-BLAST
- -HMMer
- -búsqueda con secuencias intermedias

Definición de motivo



Motivos

Son pequeñas zonas conservadas.

Se suelen corresponder con características funcionales de las proteínas:

-centros activos

-sitios de unión de ligandos

-etc

Secuencias consenso y patrones

¿Cómo aprovechar la información del alineamiento múltiple?

-Secuencias consenso:

```
AGTVATVSC
AGTSATHAC
IGRCARGSC
IGEMARLAC
IGDYARWSC
.........
IGTVARVSC <= Ejemplo de secuencia consenso
```

-Patrones o expresiones regulares:

(para caracterizar motivos)

ALRDF**ATHD**DF
SMTAE**ATHD**SI
ECDQA**ATHE**AS

A-T-H-[DE]

Patrones (expresiones regulares)

¿Cómo expresarse regularmente?

- Cualquier aminoácido: x
- Ambigüedad:
  - [A,B] A, o B...
  - {A,B..} cualquiera menos A y B.
- Repetición: A(2,4) significa A-A o A-A-A o A-A-A-A
- N terminal: <, C-terminal: >

Ejemplo: [AC]-x-V-x(4)-{E,D}.

[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

Patrones: un ejemplo

Ejemplo:

AGTVATVSC
AGTSATHAC
IGRCARGSC
IGEMARLAC
IGDYARWSC
....
IGTVARVSC ← Ejemplo de secuencia consenso

[AI]-G-X-X-A-[RT]-[SA]-C ← Ejemplo de patrón

Construcción de un patrón:

-Más o menos subjetivo. Ensayo y error.
-Objetivo: alta sensibilidad, alta especificidad

Debemos construirlos en torno a motivos conservados.

Suficientemente cortos (sensibilidad), suficientemente largos (especificidad)

Patrones (III)

Ventajas y desventajas de los patrones

-Su construcción es bastante laboriosa, pero...

existen algunos métodos automáticos (PRATT: http://www.ebi.ac.uk/pratt/)

existen bases de datos donde expertos hacen ese trabajo por nosotros

(Prosite: http://www.expasy.org/prosite)

-Muy estrictos.

Básicamente distingue posiciones importantes y no importantes (con 'X'), pero en la Naturaleza hay una mayor gradación.

Si una proteína nueva se sale de la regla general, no será detectada con el patrón.

Guión de la charla. Patrones, perfiles y dominios.

-cómo utilizar la información de los alineamientos múltiples

- -secuencias consenso y expresiones regulares
- -perfiles y perfiles-hmm

-algunas bases de datos de patrones y perfiles:

- -Prosite
- -Pfam

-búsquedas en bases de datos:

- -PSI-BLAST
- -HMMer
- -búsqueda con secuencias intermedias

Dominios

"Unidad estructural independiente", en otras áreas se le da un sentido diferente (en estudios genéticos de deleción a veces se utiliza como sinónimo de la parte mínima de la secuencia capaz de realizar la función).

Se muestra el antígeno de Histocompatibilidad de clase I: dominios alpha1,2 y 3 y proteína beta-2-microglobulina.

¿dos dominios o uno?

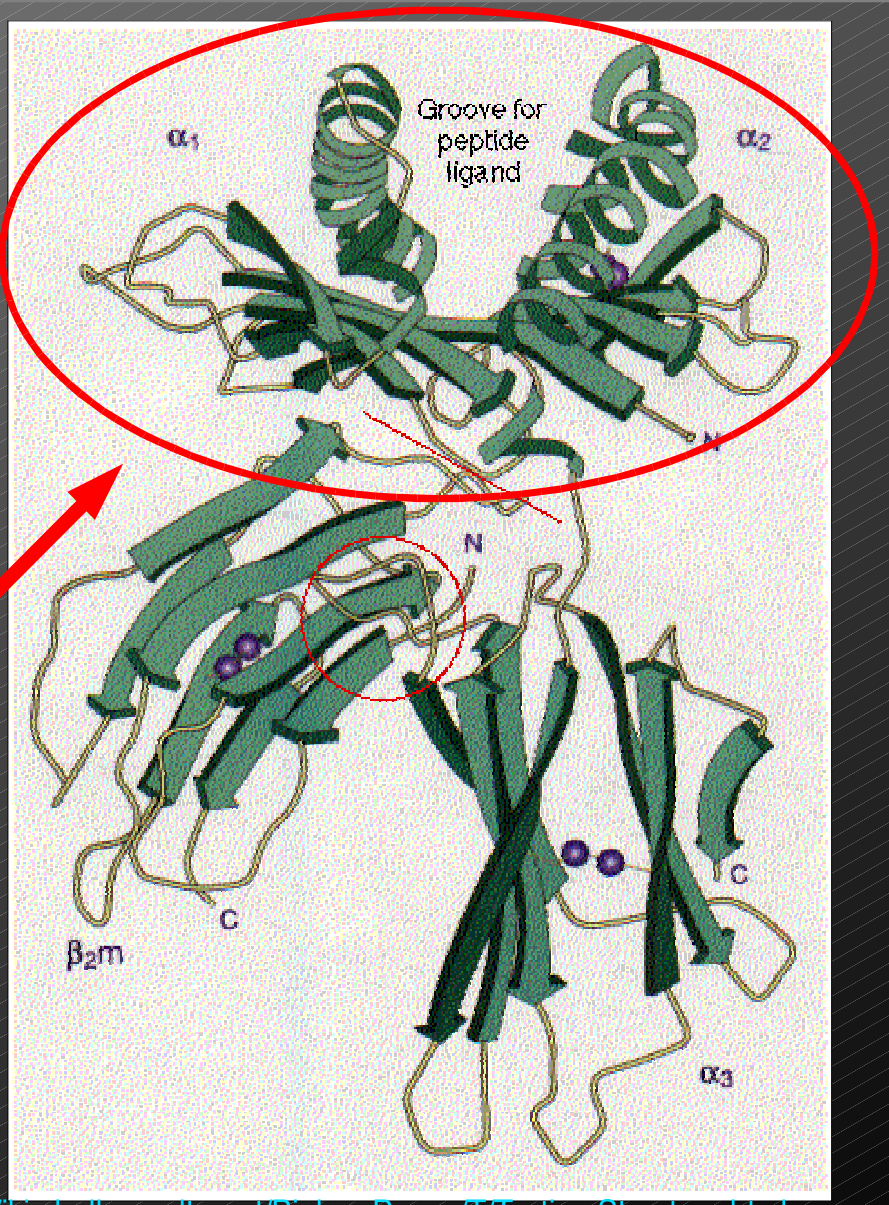


Imagen tomada de: http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/T/TertiaryStructure.html, que a su vez la tomó de P. J. Bjorkman from Nature 329:506, 1987

Perfiles

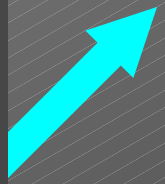
Si queremos alinear la secuencia FKTLGCCLLV al perfil, el mejor alineamiento será:

```
F K L L S H C L L V
F K A F G Q T M F Q
Y P I V G Q E L L G
F P V V K E A I L K
F K V L A A V I A D
L E F I S E C I I Q
F K L L G N V L V C
F K T L G C C L L V
```

Y la puntuación:

```
60 25 -6 27 28 -26 22 33 26 -16
```

Lo cual suma en total: 173.



	F	K	L	L	S	H	C	L	L	V
	F	K	A	F	G	Q	T	M	F	Q
	Y	P	I	V	G	Q	E	L	L	G
	F	P	V	V	K	E	A	I	L	K
	F	K	V	L	A	A	V	I	A	D
	L	E	F	I	S	E	C	I	I	Q
	F	K	L	L	G	N	V	L	V	C
A	-18	-10	-1	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

Perfiles (II)

¿Cómo utilizar un perfil para buscar homólogos?

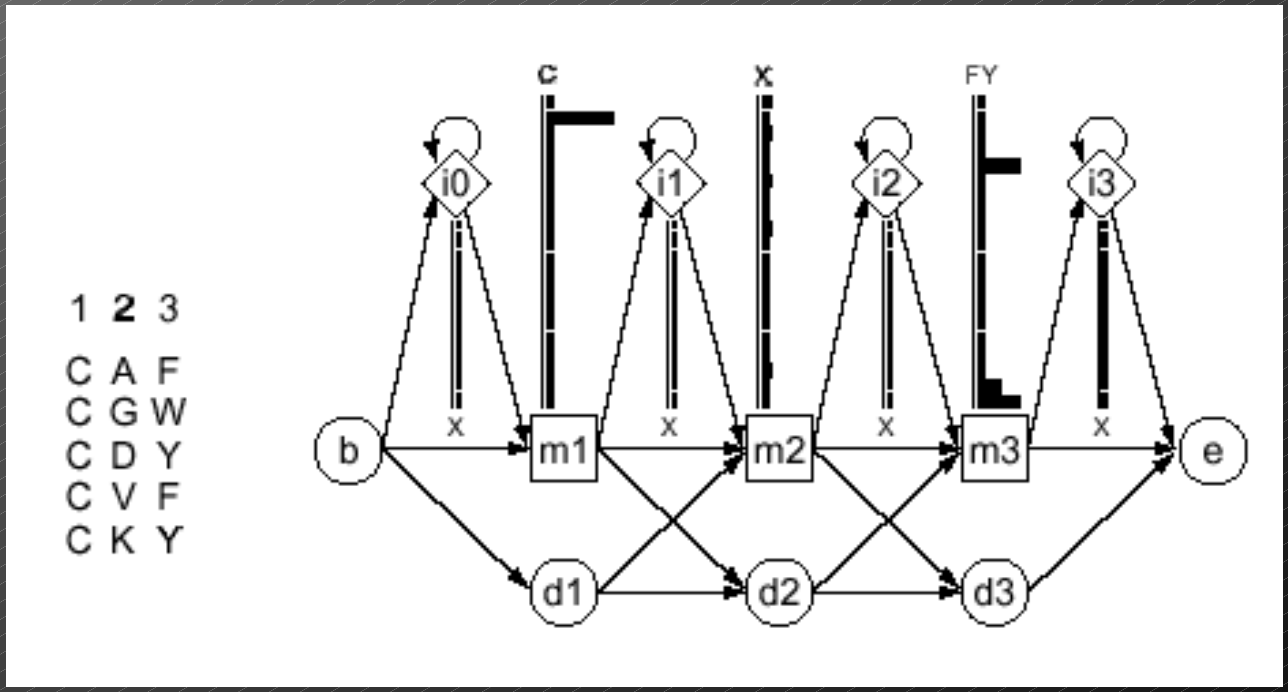
El mismo algoritmo usado para alinear dos secuencias (Smith & Waterman) sirve para alinear una secuencia y un perfil.

Perfiles de tipo HMM

Perfiles de tipo HMM (hidden markov model)

La base probabilística de los perfiles simples es pobre, especialmente en cuanto a la penalización de gaps.

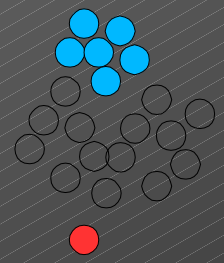
Los HMM son más sólidos (y complejos)



Perfiles (III)

Posible problema de los perfiles: sesgo en la representación de las secuencias.

Solución: asignar distintos “pesos” a las secuencias en función de cuánta información extra aporten al perfil.



ATCILYC
ATCILYC
MTCIRYC
ATCDLYC
ATCILYC
ATCILYC
SRCDRMC

Guión de la charla. Patrones, perfiles y dominios.

-cómo utilizar la información de los alineamientos múltiples

- -secuencias consenso y expresiones regulares
- -perfiles y perfiles-hmm

-algunas bases de datos de patrones y perfiles:

- -Prosite
- -Pfam

-búsquedas en bases de datos:

- -PSI-BLAST
- -HMMer
- -búsqueda con secuencias intermedias

Prosite (I)

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;
//
```

Prosite (II)

Lo que podemos hacer con Prosite:

-buscar con una secuencia para ver si se parece a alguno de los patrones o perfiles descritos en la base de datos.

-encontrar información de una familia determinada y ver qué proteínas pertenecen a ella.

-buscar con uno de los patrones o perfiles descritos contra una base de datos de secuencias.

-etcétera

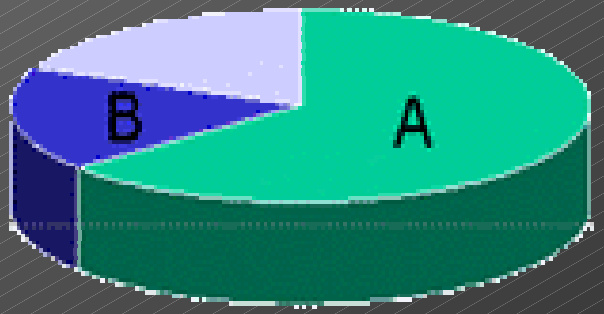
Pfam (I)

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:

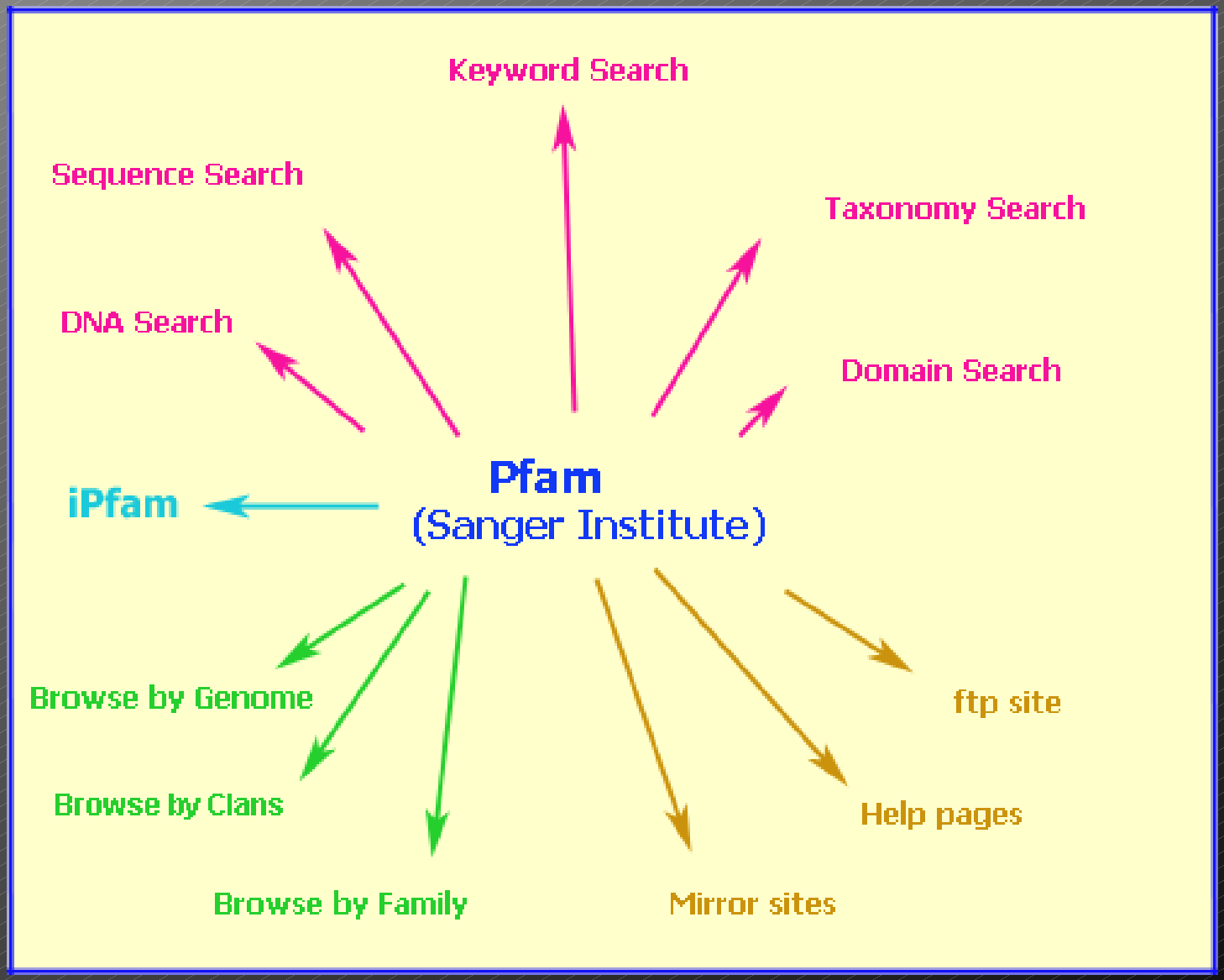
Caspasa 9:

-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.

Lo que podemos hacer con Pfam



Guión de la charla. Patrones, perfiles y dominios.

-cómo utilizar la información de los alineamientos múltiples

- -secuencias consenso y expresiones regulares
- -perfiles y perfiles-hmm

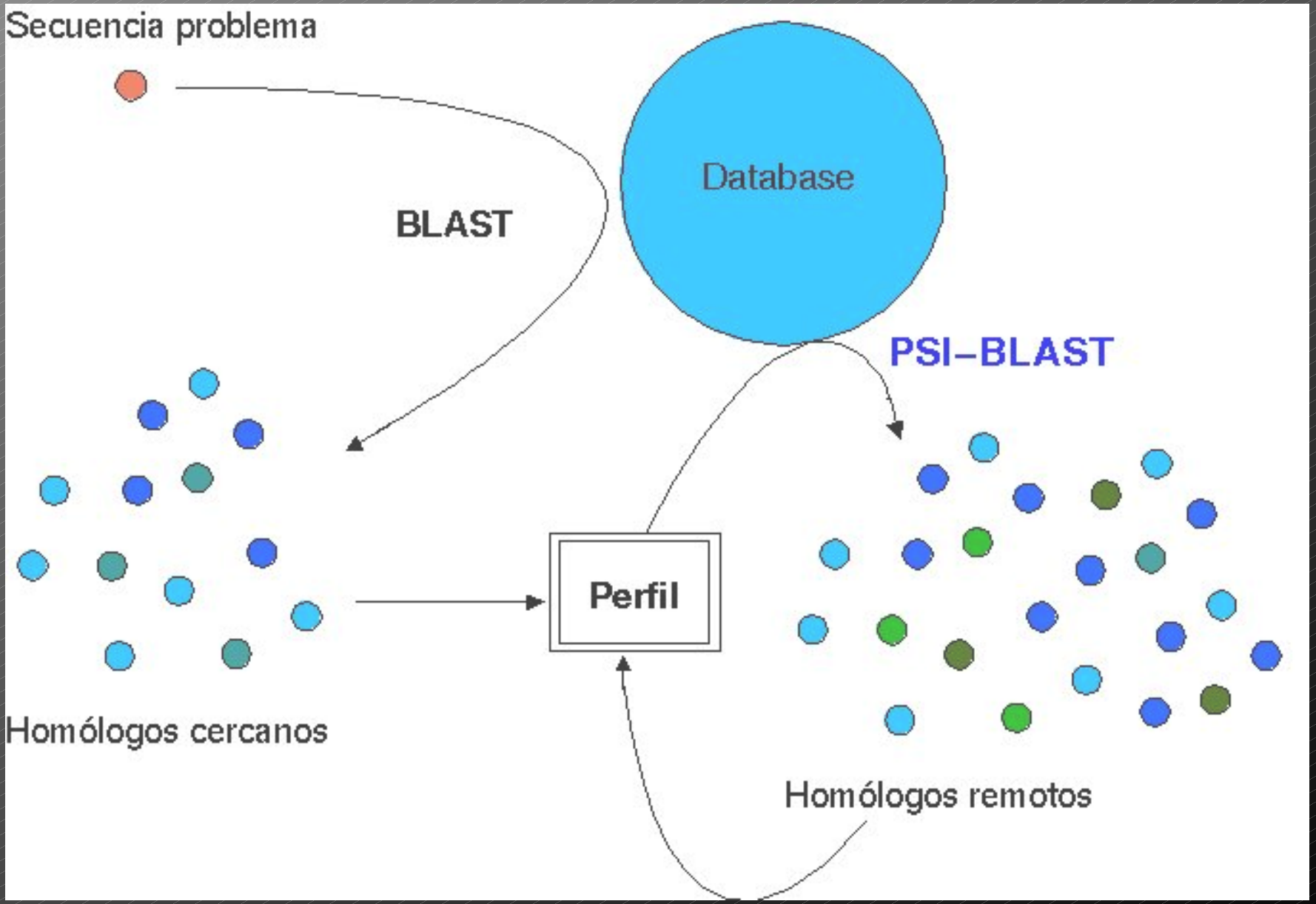
-algunas bases de datos de patrones y perfiles:

- -Prosite
- -Pfam

-búsquedas en bases de datos:

- -PSI-BLAST
- -HMMer
- -búsqueda con secuencias intermedias

Búsqueda de homólogos con PSI-BLAST



Demostración del funcionamiento de PSI-BLAST.

Página de PSI-BLAST:

http://www.ncbi.nlm.nih.gov/BLAST/

Secuencia de:

```
>gi|2501594|sp|Q57997|Y577_METJA PROTEIN MJ0577
MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVAGLNKSVEEFE
NELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEGVDIIIMGSHGKTNLKEILLG
SVTENVIKKSNKPVLVVKRKNS
```

(es el ejemplo que se sigue en el tutorial del NCBI:
http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/psi1.html)

Búsqueda de homólogos con HMMer

http://hmmer.wustl.edu/

Es el método más sensible, pero es muy lento.

Requiere demasiados recursos, por lo que no se puede utilizar a través de la web.

Strategia básica:

1.- obtener homólogos (p.e. con BLAST) y construir un alineamiento múltiple (p.e. con Clustalw).

2.- transformar el alineam. múltiple en un perfil HMM:

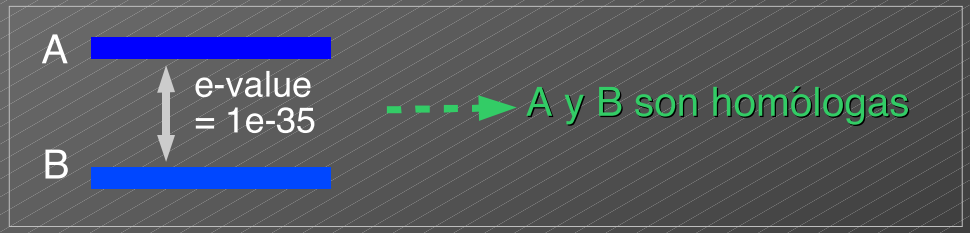
1º: hmmbuild, 2º hmmcalibrate.

3.- búsqueda con el perfil HMM en una base de datos de secuencias: hmmsearch.

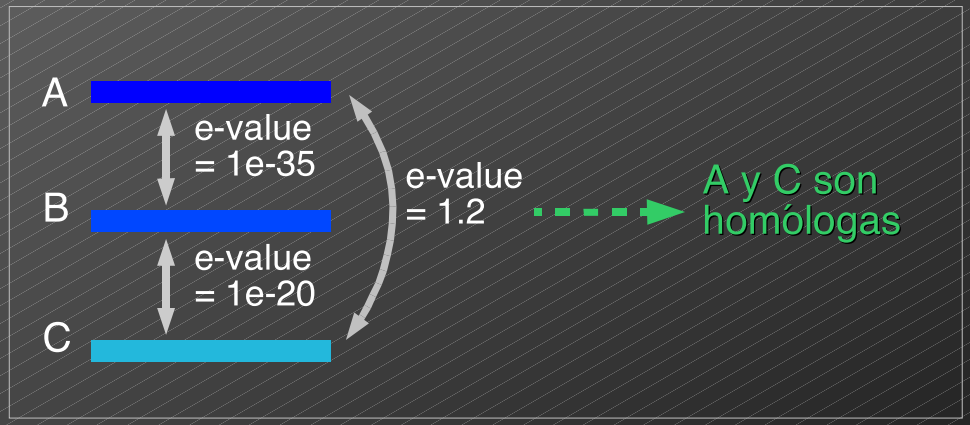
4.- con los nuevos homólogos que encontremos podemos volver al paso "2".

Búsqueda con secuencias intermedias

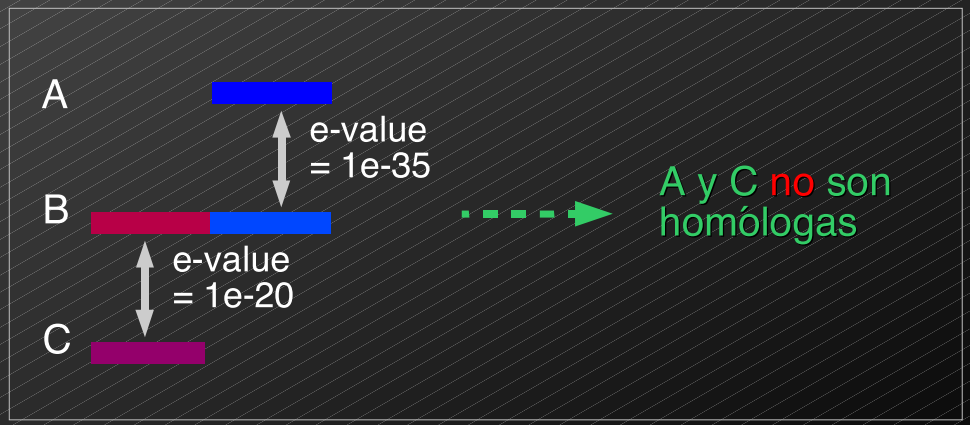
No utiliza información del alineamiento múltiple, pero puede superar las limitaciones de métodos sencillos como BLAST.



Propiedad transitiva de la homología:
si dos proteínas A y B son homólogas, y a su vez B y C son homólogas, entonces A y C también son homólogas, aunque no se parezcan entre sí.



La propiedad transitiva sólo es aplicable cuando los dominios de las proteínas se corresponden unos con otros ("la unidad evolutiva son los dominios").



Guión de la charla. Patrones, perfiles y dominios.

-cómo utilizar la información de los alineamientos múltiples

- -secuencias consenso y expresiones regulares
- -perfiles y perfiles-hmm

-algunas bases de datos de patrones y perfiles:

- -Prosite
- -Pfam

-búsquedas en bases de datos:

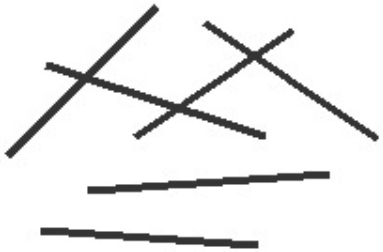
- -PSI-BLAST
- -HMMer
- -búsqueda con secuencias intermedias

Formas de comparar secuencias (I)

1 secuencia contra una base de datos de secuencias.
BLAST(web/local), FASTA(web/local), Smit & Waterman(web/local)

1 patrón contra una base de datos de secuencias.
ScanProsite (web), ps_scan(local)

[AI]–G–X–X–A–[RT]–[SA]–C *versus*



1 secuencia contra una base de datos de patrones.
ScanProsite(web), ps_scan(local), MotifScan(web)

[AI]–G–X–X–A–[RT]
[N]–RPG–A–[HT]>
P–E–P–X–T–[OA]
[ED](2)–G–[MN]–D–{RST}
L–S–X(3,10)–A–[DEHGGD]

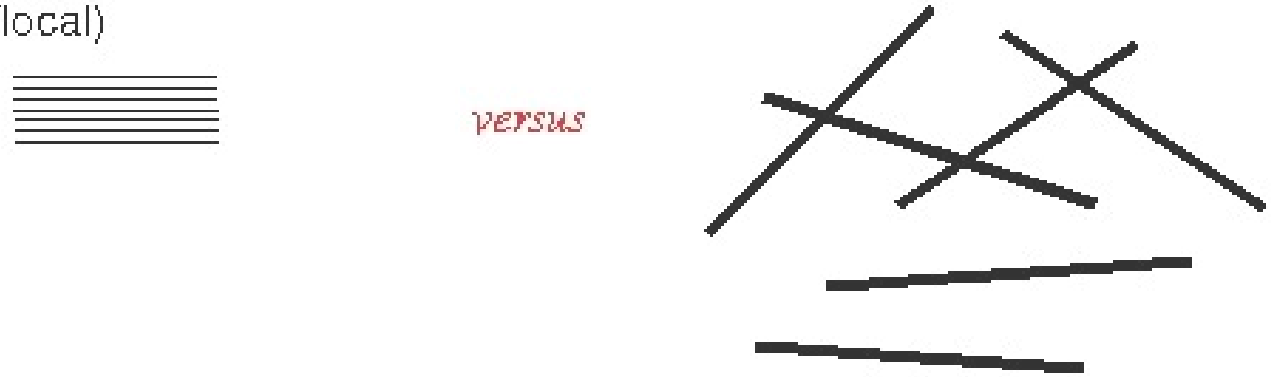
Formas de comparar secuencias (y II)

1 secuencia contra una base de datos de perfiles o HMMs.

ScanProsite(web), MotifScan(web), Pfam (hmmpfam)(web/local), Impala(web/local), Rps-blast(web/local)

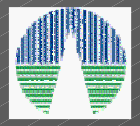
1 perfil o un perfil-HMM contra una base de datos de secuencias

Bioccelerator (profileSearch)(web), hmmsearch(local), PSI-BLAST(web/local)



Agradecimientos

Algunas figuras han sido tomadas de...

-Paulino Gómez Puertas  Centro de Astrobiología

-Oswaldo Trelles  Arquitectura de Computadores
Universidad de Málaga

-Joaquín Dopazo  Bioinformatics Unit
CNIO