Alineamiento de secuencias. Búsqueda de homólogos. Alineamientos múltiples. Patrones y perfiles.

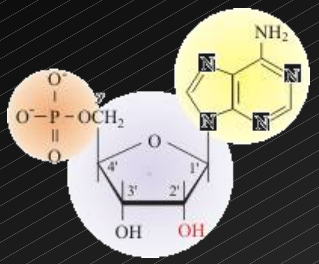
Curso de verano de *Bioinformática* de la UCM

Madrid 2007

Federico Abascal

Centro Nacional de Biotecnología

¿Qué es una secuencia?



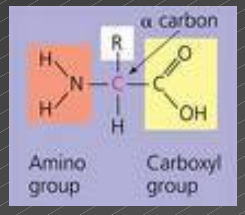
Nucleótido
A: adenina
C: citosina
T: timina
G: guanina

GTGATAATCACTCGTTGACTATTCTCAACCAACCAC
AAAGATATTGGTACCCTATACATGATTTTCGGGGC
CTGAGCTGGAATAGTTGGAACCGCTCTAAGCCTAC
TTATTCGAGCCGAACTCAGCCAACCTGGAGCTCTC
CTA

Manual de instrucciones

Traducción del mensaje
(previa transcripción a ARN)

Código genético:
AGG = R (Arg)
Codón = amino ácido

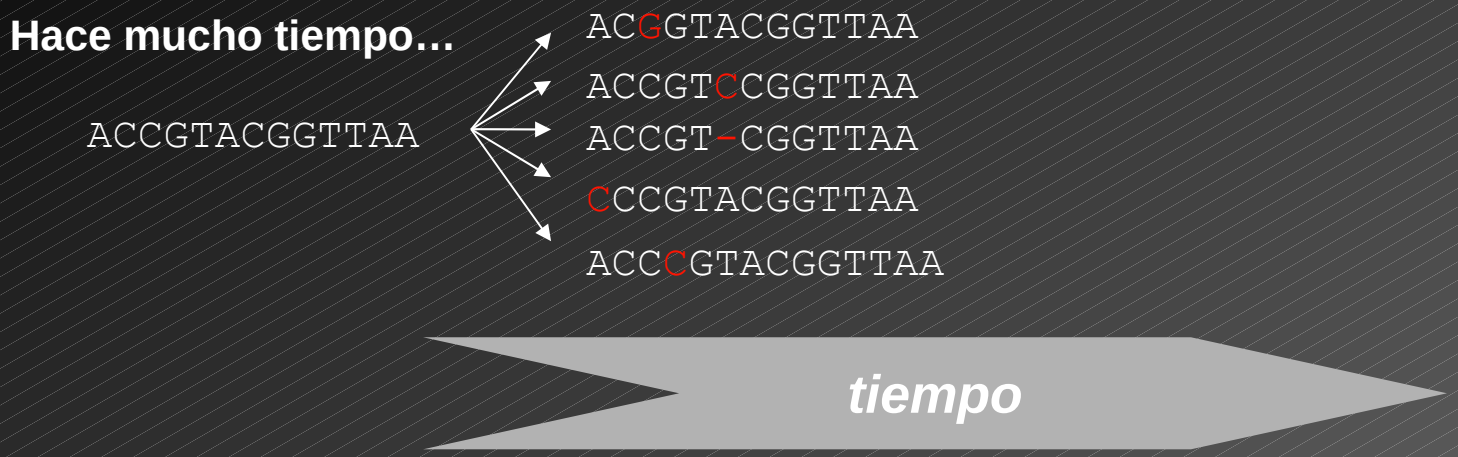


MMITRWLFSTNHKDIGTLYMIFGAWAGMVGTALSLLIRAEL
SQPGALLGDDQIYNVIV

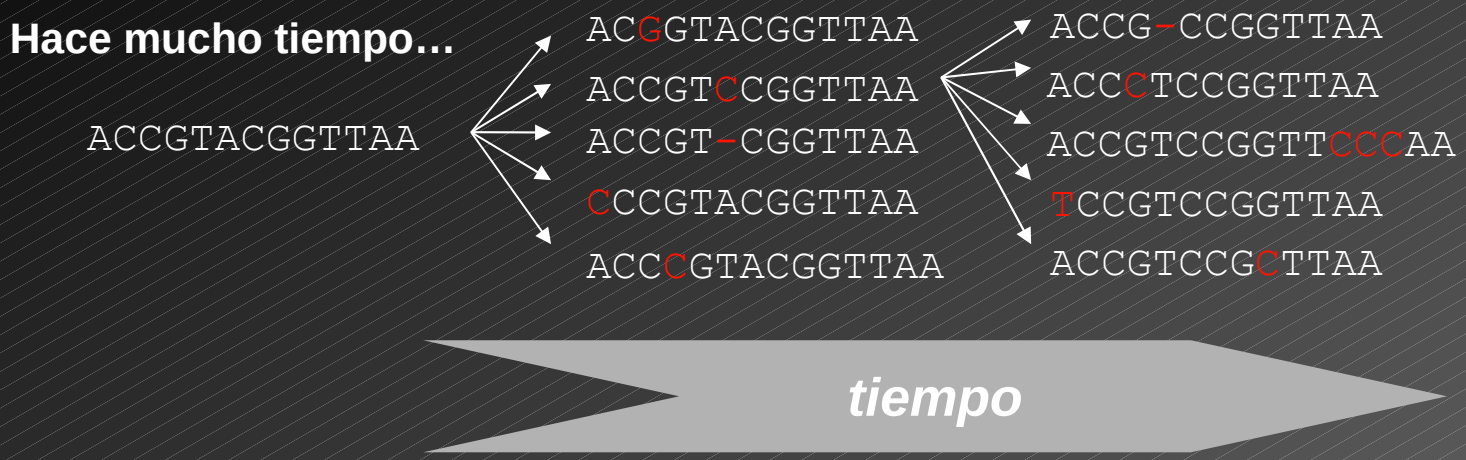
Amino ácido
ACDEFGHIKLMNPQRSTVWY

"Actores" en la célula

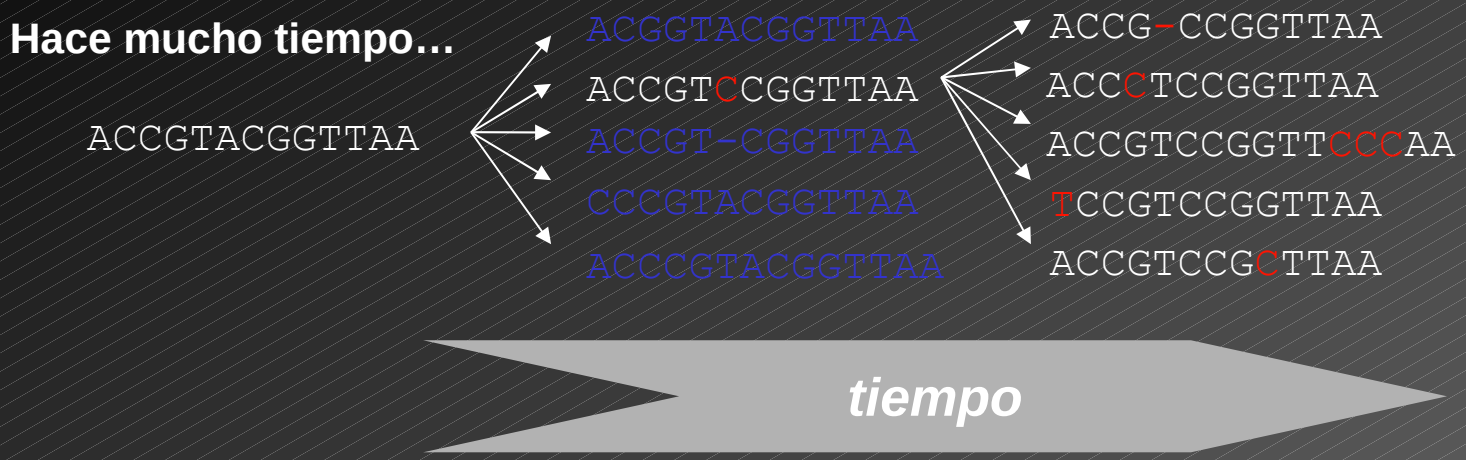
Modelo evolutivo: cambio al azar + selección natural



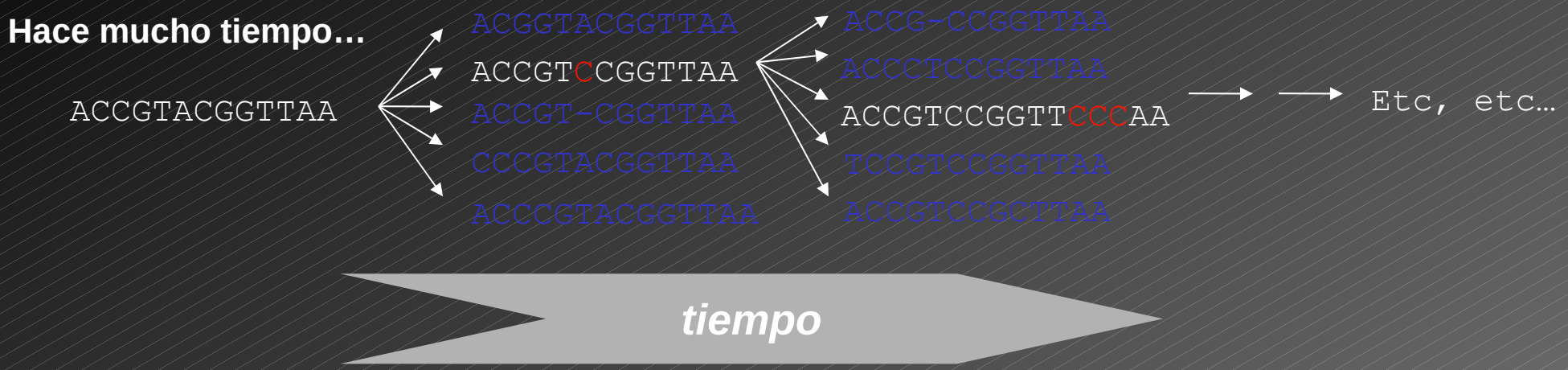
Modelo evolutivo: cambio al azar + selección natural



Modelo evolutivo: cambio al azar + selección natural



Modelo evolutivo: cambio al azar + selección natural



x

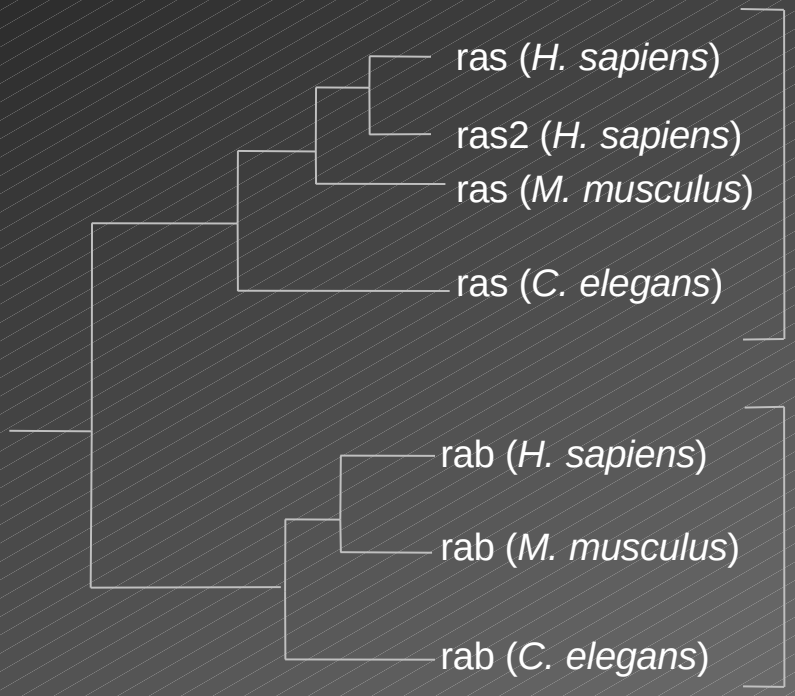
***n* especies**

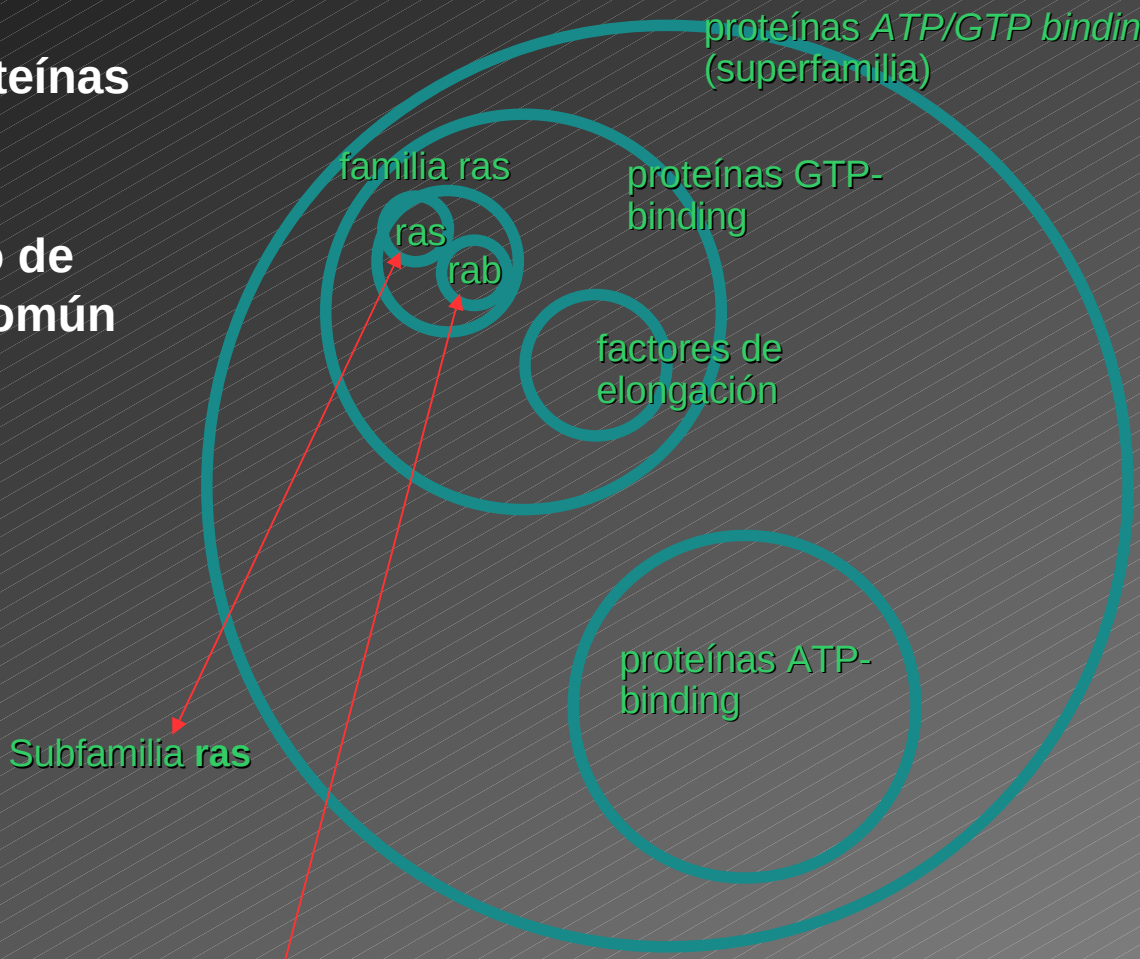
ACCGTCCGGTTGA
ACCGTCGTAA
ACCTCTAGTTAA
GGAGTACGGTTAA
ACCGTTCCGAA
CCGTCCCGTTAA
ACCGTACGGTTATA
AACCGTACGGTTAA
ACCTGCAATTA
GCCGTACCGTGGTCCA
ACCGTACCCCGGTTAA

Cambio al azar + selección natural + duplicaciones génicas

Superfamilia: grupo de proteínas con un origen común.

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





Dos formas de representarlo

Homólogos: ortólogos y parálogos.

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

Los mismos genes en distintas especies.

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

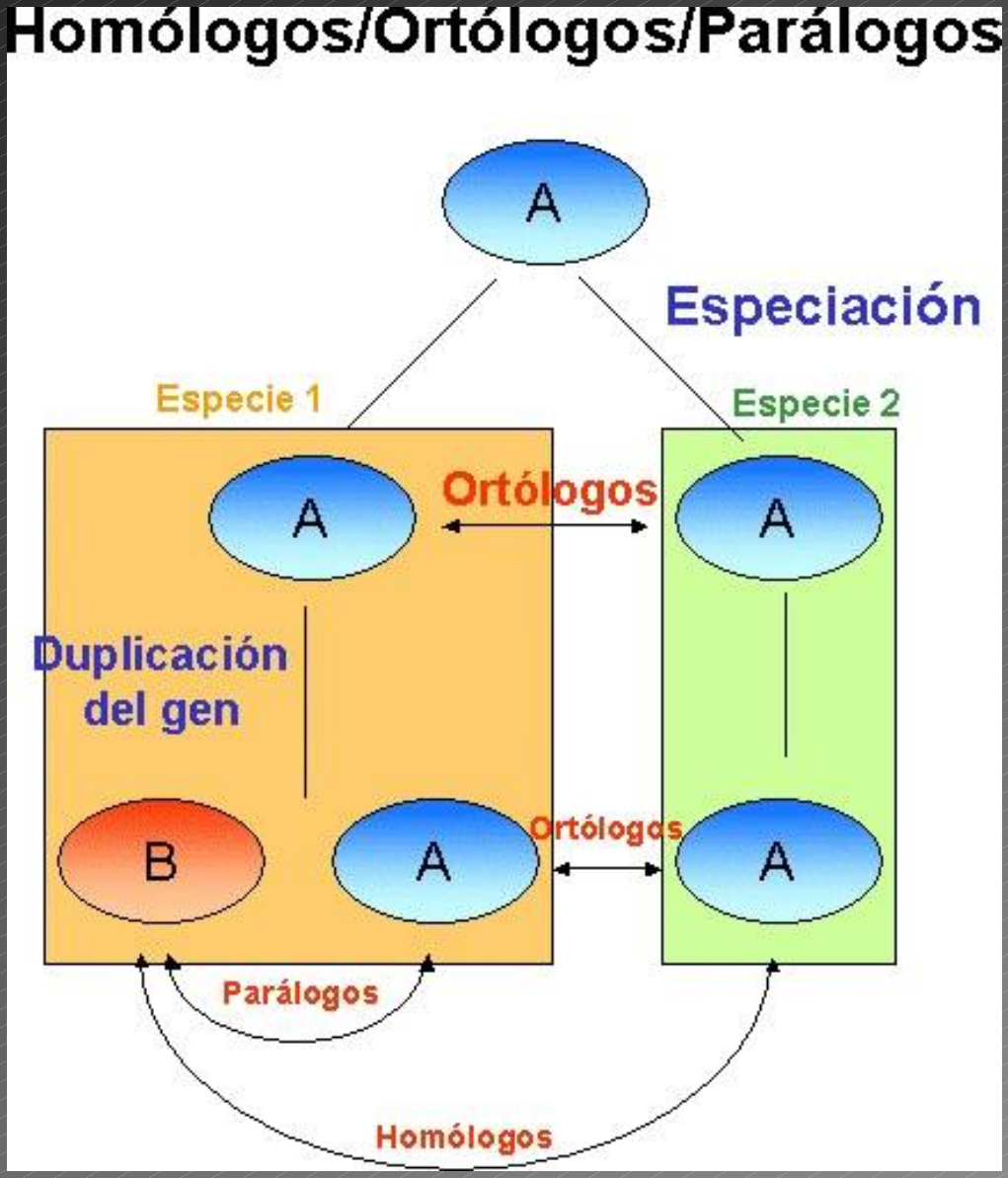
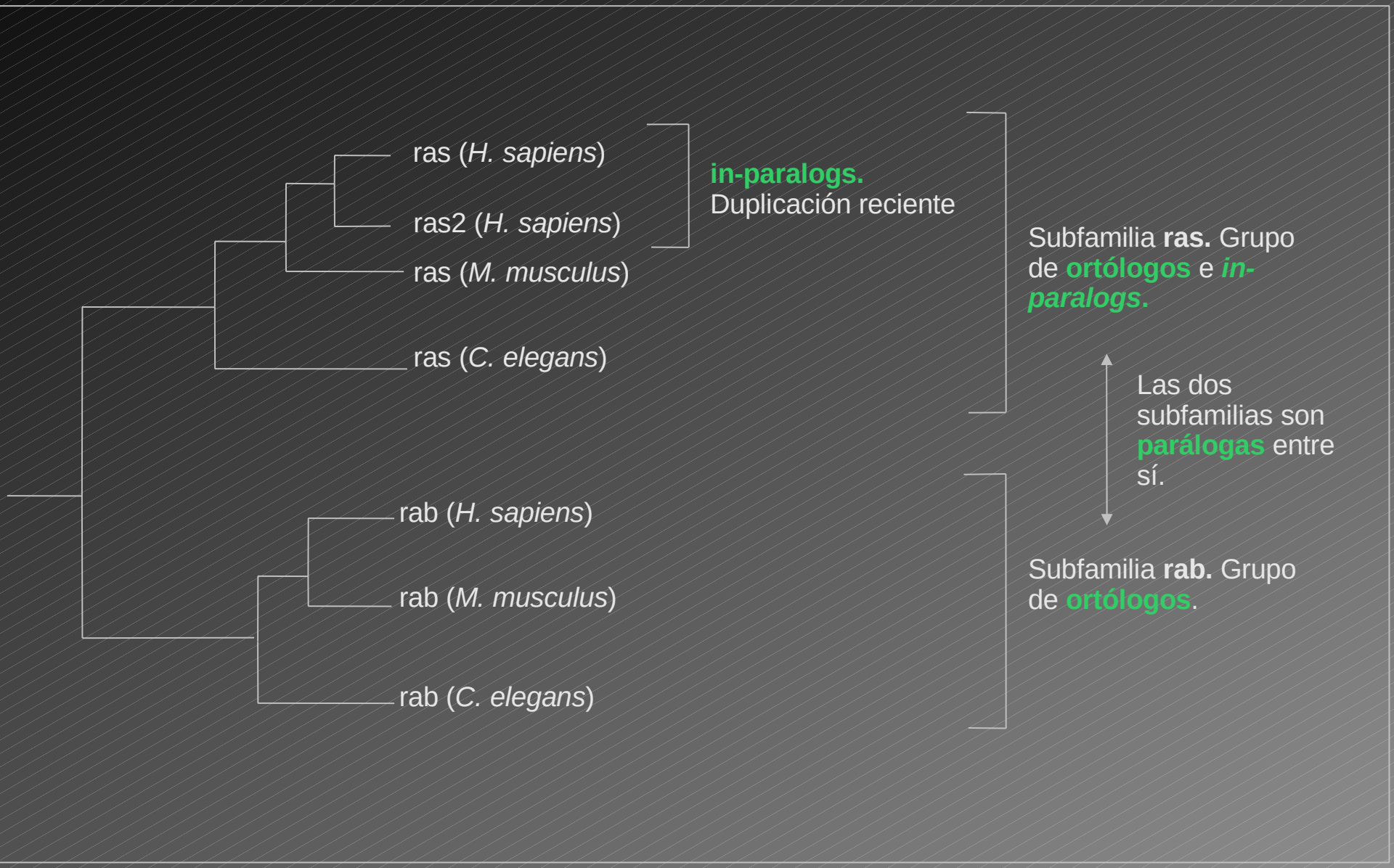


Imagen tomada de una presentación de Manuel José Gómez (CAB)

Homólogos: ortólogos y parálogos.

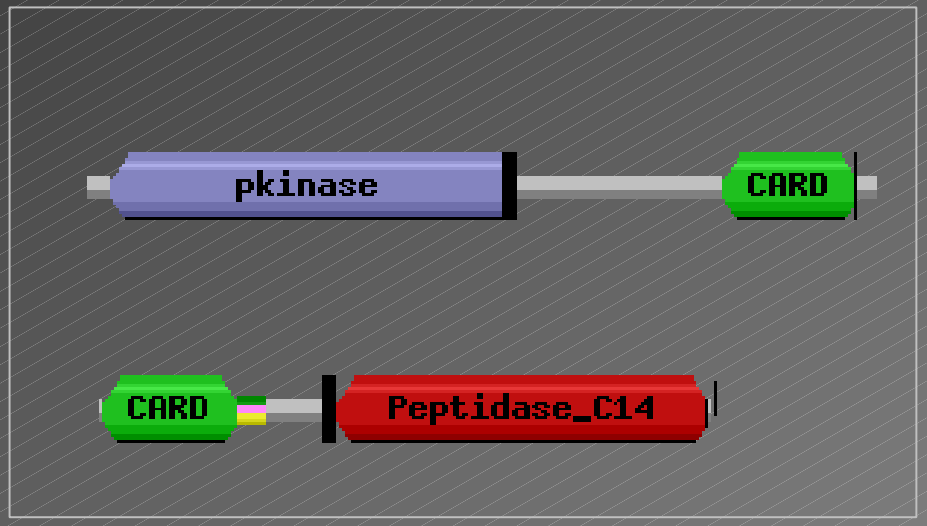


Cambio + selección + duplicaciones + barajado de dominos

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

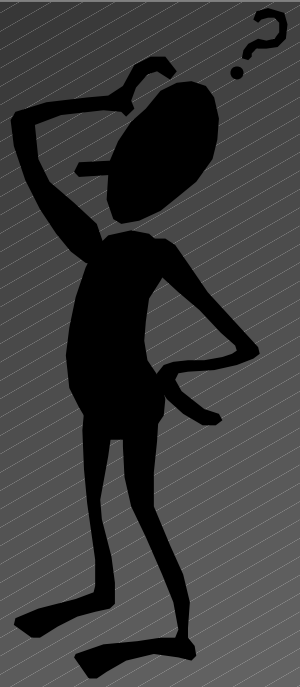
El dominio, y no el gen, es la unidad evolutiva básica.

- 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.



¿Qué nos dicen las secuencias?

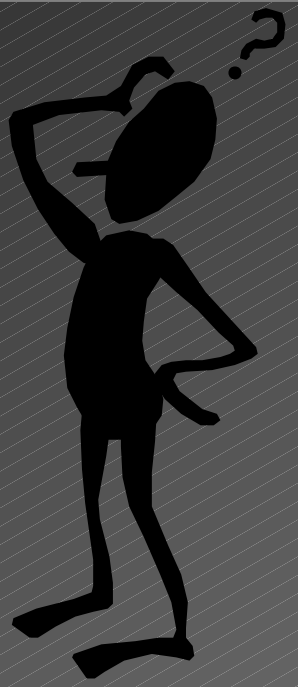
Una secuencia: **ADGHLSCETRDLWYALDSOPRL**



¿Qué nos dicen las secuencias?

Una secuencia: **ADGHLSCETRDLWYALDSOPRL**

Dos secuencias: **ADGHLSCETRDLWYALDSOPRL**
EGHICECSSELWPILDTOPPPDL



¿Qué nos dicen las secuencias?

Una secuencia: ADGHLSCETRDLWYALDSOPRL

Dos secuencias:
ADGHLSCETRDLWYALDSOPRL
EGHICECSSELWPILDTOPPPDL

Dos secuencias alineadas:
ADGHLSCETR-DLWYALDSOP--RL
-EGHI-CECSSELWPILDTOPPPDL

¿Qué nos dicen las secuencias?

Una secuencia:

ADGHLSCETRDLWYALDSOPRL

Dos secuencias:

ADGHLSCETRDLWYALDSOPRL
EGHICECSSELWPILDTOPPPDL

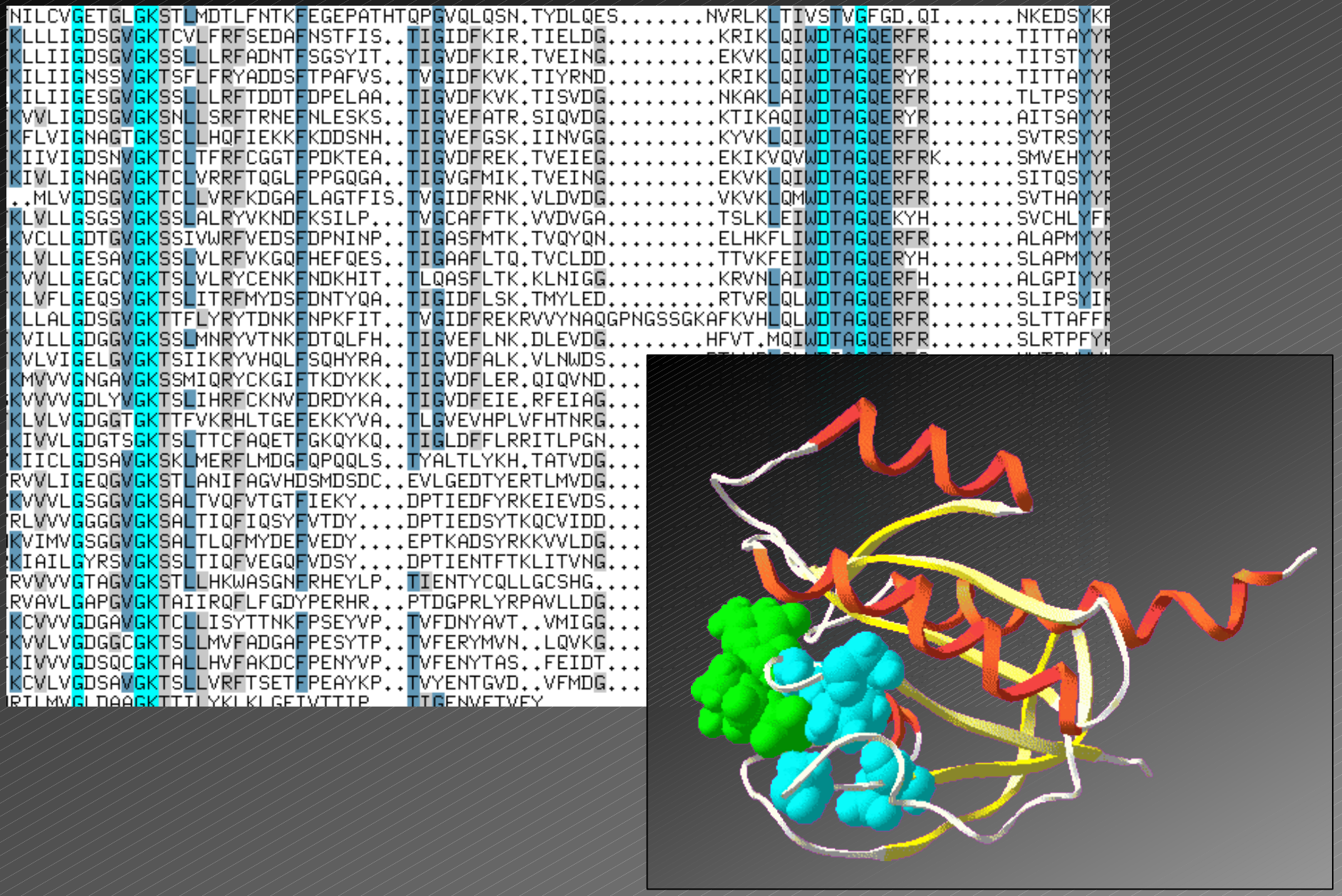
Dos secuencias alineadas:

ADGHLSCETR-DLWYALDSOP--RL
-EGHI-CECSSELWPILDTOPPPDL

Muchas secuencias alineadas:

ADGHLSCETR-DLWYALDSOP--RL
-EGHISCECSSELWPILDTORPPDL
AESHLTDECDSELWPILETOPPPDL
ADGHL-CETSSELNPALDAOP--EL
-E-HI-MECYSELIPILETORP-RL
AESHLTDECDTELMKILDTOLPPDL
ADGHL-CETSSELWPALDSOP--D-
-E-HI-MECYSEL-KILDTOPP-DL

¿Por qué comparar secuencias ... de proteínas?



¿Por qué comparar secuencias...

... de proteínas?

-para conocer la función de las proteínas:
  -función general.
  -residuos importantes: p.e. centros activos.
-para predecir la estructura 3D de las proteínas.
-para determinar en qué especies está una proteína.
-...

... de ADN?

-para buscar genes:
  -ESTs.
  -ADN genómico.
-para estudios de genética poblacional (SNPs).
-para comparar secuencias no codificantes.

¿Por qué comparar secuencias...

... de proteínas?

-para conocer la función de las proteínas:
  -función general.
  -residuos importantes: p.e. centros activos.
-para predecir la estructura 3D de las proteínas.
-para determinar en qué especies está una proteína.
-...

... de ADN?

-para buscar genes:
  -ESTs.
  -ADN genómico.
-para estudios de genética poblacional (SNPs).
-para comparar secuencias no codificantes.

¿Cúal es el objetivo de la comparación?

El objetivo es encontrar el alineamiento que con mayor probabilidad (*nunca sabremos si es el real*) refleje qué cambios se han producido.

```
RPE_YEAST          6 IAPSIL----ASDFANLGCECHKVINAGADWLHIDVMDGHFVPNITLGQP     51
                     ||.|:|     ..|...|    .:.:..|...:|.|||| |||.|.::...
RPE_MYCPN         10 IAFSLLPLLHQFDRKLL----EQFFADGLRLIHYDVMD-HFVDNTVFQGE     54
```

¿Cómo comparar las secuencias?

-por pares

-alineamiento de dos secuencias

-búsqueda en bases de datos con BLAST.

-muchas a la vez

-alineamiento múltiple con Clustalw.

-con patrones, perfiles y hmm's

-búsqueda en bases de datos con PSI-BLAST.

-bases de datos de interés:

· PROSITE

· PFam

· InterPro

¿Cómo comparar las secuencias?

-por pares

- -alineamiento de dos secuencias
- -búsqueda en bases de datos con BLAST.

-muchas a la vez

- -alineamiento múltiple con Clustalw.

-con patrones, perfiles y hmm's

- -búsqueda en bases de datos con PSI-BLAST.
- -bases de datos de interés:
  - · PROSITE
  - · PFam
  - · InterPro

Alineamiento de pares de secuencias

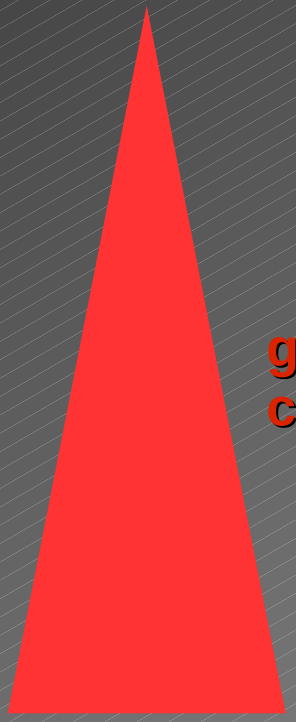
¿Cómo encontrar el alineamiento que refleja con mayor probabilidad la historia evolutiva? (i.e. el *mejor* alineamiento)

-comparación por identidades

-comparación por semejanza

- ***matrices de sustitución (BLOSUM, PAM)***

-comparación incluyendo INDELs.



grado de complejidad

Alineamiento de pares de secuencias

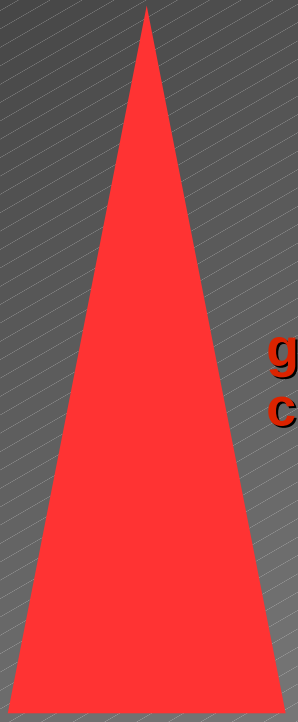
¿Cómo encontrar el alineamiento que refleja con mayor probabilidad la historia evolutiva? (i.e. el *mejor* alineamiento)

-comparación por identidades

-comparación por semejanza

- ***matrices de sustitución (BLOSUM, PAM)***

-comparación incluyendo INDELs.



grado de complejidad

Alineamiento de pares de secuencias

Comparación por identidades

```
    RWDG    0
VKDG
```

```
RWDG       2
VKDG
```

```
   RWDG     0
VKDG
```

```
RWDG       0
 VKDG
```

```
  RWDG      0
VKDG
```

```
RWDG       0
   VKDG
```

```
RWDG       0
    VKDG
```

Objetivo: encontrar el “alineamiento” con mayor número de coincidencias.

Alineamiento de pares de secuencias

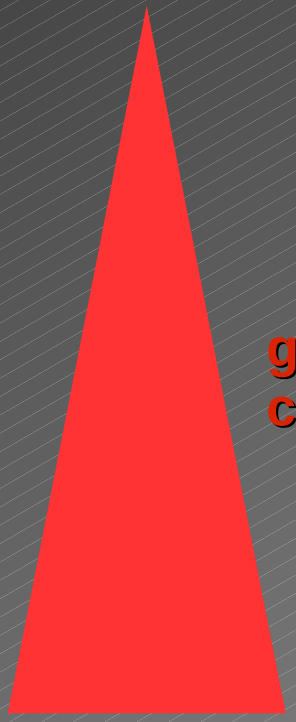
¿Cómo encontrar el alineamiento que refleja con mayor probabilidad la historia evolutiva? (i.e. el *mejor* alineamiento)

-comparación por identidades

-comparación por semejanza

- ***matrices de sustitución (BLOSUM, PAM)***

-comparación incluyendo INDELs.



grado de complejidad

Alineamiento de pares de secuencias

Comparación por semejanza

Observación: hay aa's con propiedades físico-químicas similares:

- -aa's ácidos: D, E.
- -aa's básicos: K, R, H, ...
- -aa's hidrofóbicos: L, I, W, ...
- -aa's con estr. similar: Y -P, I -L, D -N, E -Q,...
- -etc.

Objetivo: utilizar esa información para mejorar el alineamiento.

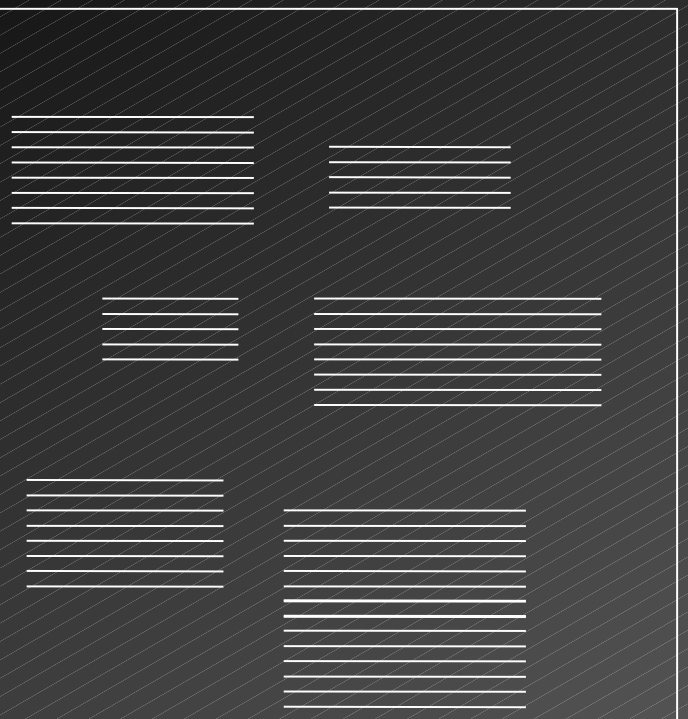
¿Cómo pasar del conocimiento general qué aa's se parecen a una estimación más precisa, cuantificada?

¿Qué sustituciones se toleran más en la Naturaleza?

Matrices de sustitución (ejs: PAM, BLOSUM)

Alineamiento de pares de secuencias

Construcción de las matrices de sustitución tipo Blosum



Análisis de miles de alineamientos múltiples

Modelo aleatorio:
A partir de las frecuencias de los aa se calculan las frecuencias esperables de cambio:

Fexp(A<->S) = F(A)*F(S)

Modelo observado:
A partir de los alineamientos se calculan las frecuencias observadas de cada posible cambio:

Fobs(A<->S) = nº cambios A<->S/nº cambios totales

Cálculo de log-odds:

Score (A-S) = log(Fobs/Fexp)

Alineamiento de pares de secuencias

Matrices de sustitución: **se construyen analizando miles de alineamientos.**

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

Alineamiento de pares de secuencias

Comparación por semejanza: alineamiento de RWDG y VKDG

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

```
RWDG
VKDG
```

Según Blosum62: -3+(-3)+6+6 = 6

```
RWDG
  VKDG
```

Según Blosum62: (-3)+(-1)+(-1) = -5
etc.

Alineamiento de pares de secuencias

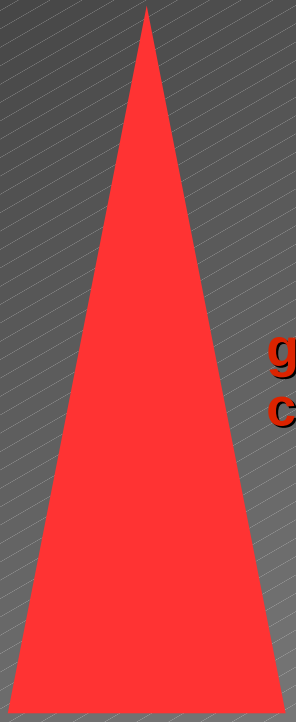
¿Cómo encontrar el alineamiento que refleja con mayor probabilidad la historia evolutiva? (i.e. el *mejor* alineamiento)

-comparación por identidades

-comparación por semejanza

- ***matrices de sustitución (BLOSUM, PAM)***

-comparación incluyendo INDELs.



grado de complejidad

Alineamiento de pares de secuencias

Comparación incluyendo INDELs (inserciones y deleciones)

```
RWDG-
V-KDG

RWDG--
V--KDG

RWDG---
V---KDG

R-WDG
VKDG-

R--WDG
VKDG--

R---WDG
VKDG---
```

```
RW-DG
V-KDG

R-WDG
VK-DG

RW-DG
VKD-G

-RWDG
VKD-G

R--WDG
-VKD-G
```

Etc, etc, etc

Alineamiento de pares de secuencias

Comparación incluyendo INDELs (inserciones y deleciones)

Observación: además de sustituciones pueden ocurrir inserciones y deleciones.

Objetivo: utilizar esa información para mejorar el alineamiento.

Problemas a resolver:

· ¿Cómo penalizar los INDELs (*los gaps*)?

Apertura y extensión de un gap.

· Las formas de alinear dos secuencias incluyendo gaps son enormes => problema computacional.

Programación dinámica.

(Needlemann & Wunsch, Smith & Waterman)

Alineamiento de pares de secuencias

Comparación incluyendo INDELs (inserciones y deleciones)

· **¿Cómo penalizar los INDELs (*los gaps*)?**

Apertura y extensión de un gap.

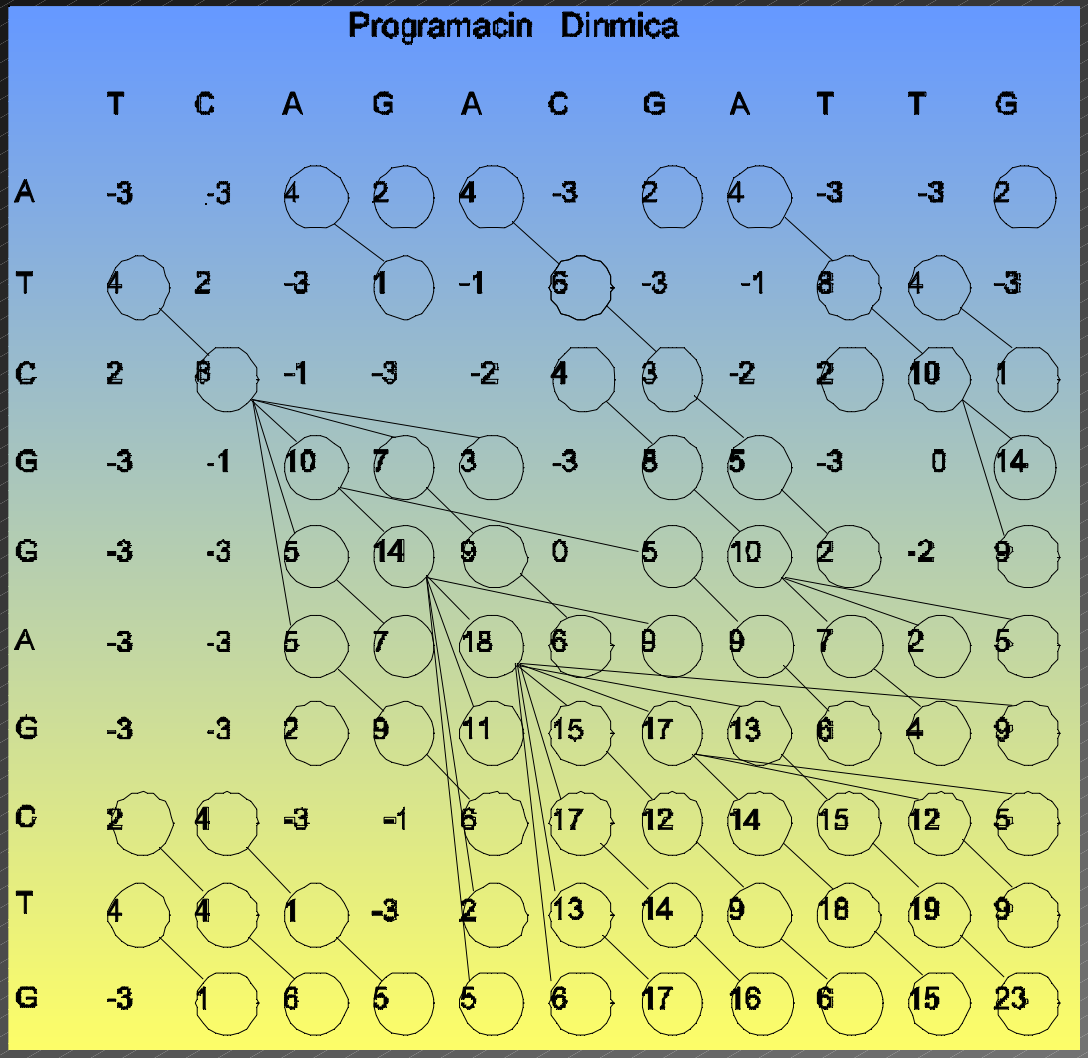
La idea es que cinco *gaps* separados son menos probables que un solo *gap* de extensión 5.

```
Caso 1:
ATGA-GATG-AT-GATACCG-ATG
ATGATGATGTATAGATTACGGATG

Caso 2:
ATGAGATG----ATGATACCGATG
ATGATGATGTATAGATTACGGATG
```

Alineamiento de pares de secuencias

Comparación incluyendo INDELs: Programáción dinámica.



Esquema de Pesos

[4] residuos iguales

[2] residuos del mismo tipo

[-3] Resto.

iGap: -5

eGap: -2

Mejor alineamiento:

```
TCAGACGATTG
||.||  ..||
ATCGGA--GCTG
```

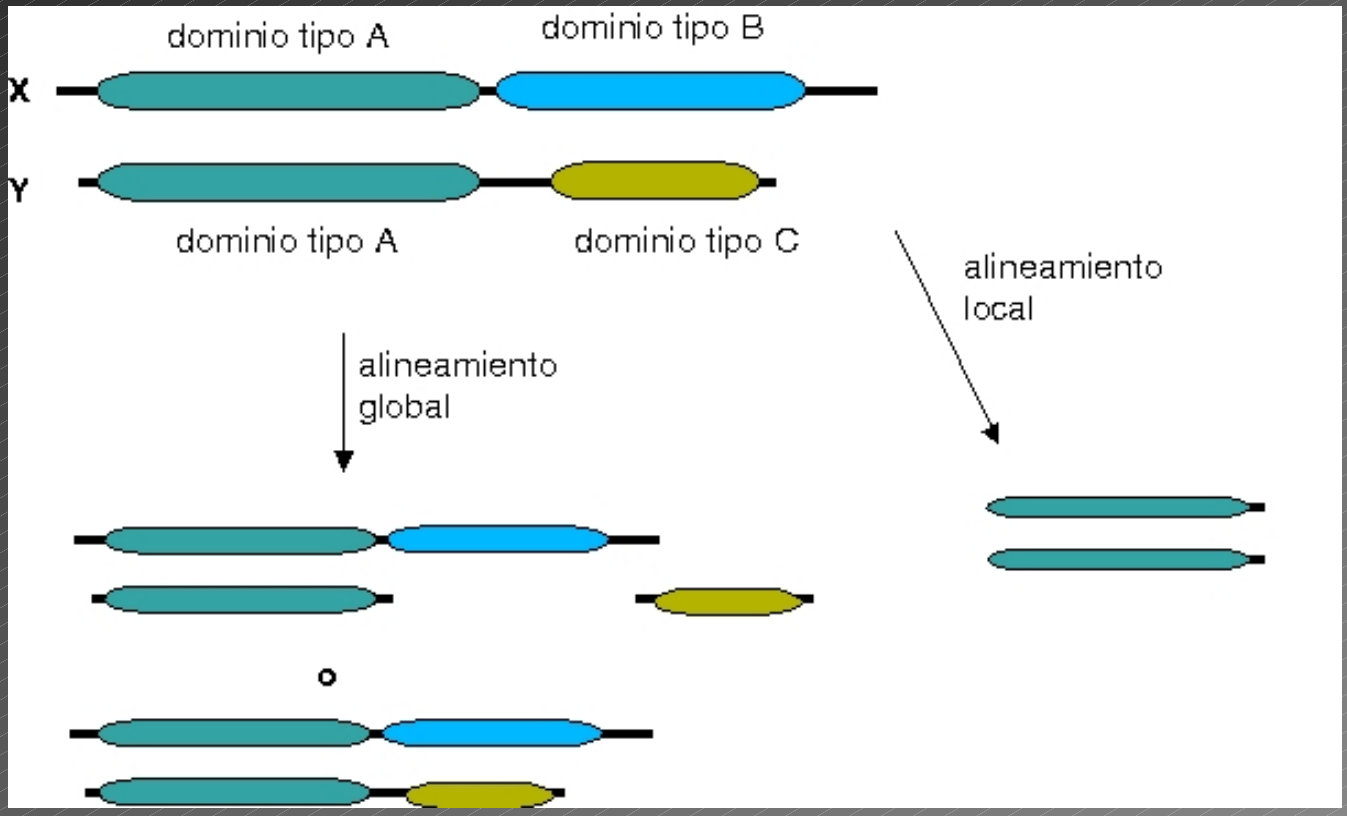
Alineamiento de pares de secuencias

Alineamiento global *versus* alineamiento local

Halla aquéllos trozos de las secuencias que superpuestos resultan en una puntuación máxima.

Trata de obtener el mejor alineamiento superponiendo las secuencias completas.

Sólo se debe utilizar cuando las proteínas son homólogas en toda su extensión (tienen los mismos dominios)



Ejemplos de Global vs. Local

Human alpha-1 **hemoglobin** and plant **Leghemoglobin**

Global alignment: Score: 17

```
  1 MGAFSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFSFLSNGVDPNNPKLK     60
    |   |   :: ||::|     :   : |     :     |  :  |       :   : ::|
  1 M-VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFD--LSHGSAQVK     57

 61 AHAEKVFKMTVDSAVQLRAKGEVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKTFKEAV    119
    | :||     ::   :    ::  |   |  :|  |  :|| :| ::   || |    :
 58 GHGKKVADALTNAVAHV---DDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHL    114

120 GDKWNDELGNAWEVAYDELAAAIKKAMGS--A   149
        |   |   : |:  |::   : |
115 ----PAEFTPAVHASLDKFLASVSTVLTSKYR   142
```

Local alignment: Score: 42

```
  5 SEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFSFLSNGVDPNNPKLKAHAE     64
    |   :: ||::|     :   : |     :     |  :  |       :   : ::| | :
  4 SPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFD--LSHGSAQVKGHGK     61

 65 KVFKMTVDSAVQLRAKGEVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT   114
    ||     ::   :    ::  |   |  :|  |  :|| :| ::   || |
 62 KVADALTNAVAHV---DDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVT   109
```

From G. Lunter

¿Cómo comparar las secuencias?

-por pares

- -alineamiento de dos secuencias
- -búsqueda en bases de datos con BLAST <= artículo más citado en los 90

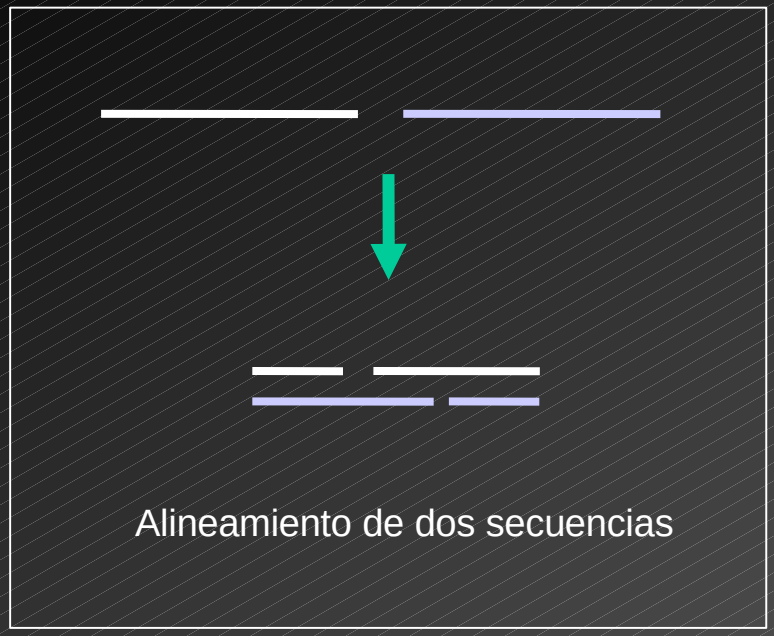
-muchas a la vez

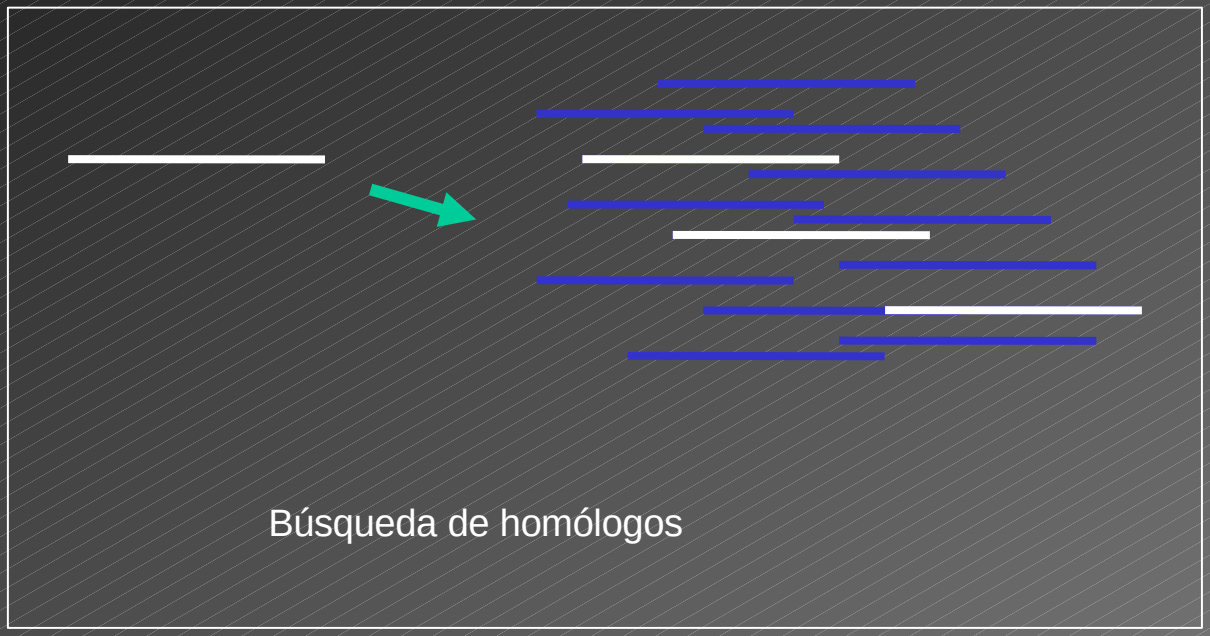
- -alineamiento múltiple con Clustalw.

-con patrones, perfiles y hmm's

- -búsqueda en bases de datos con PSI-BLAST.
- -bases de datos de interés:
  - · PROSITE
  - · PFam
  - · InterPro

Búsqueda en bases de datos con BLAST





e

Búsqueda en bases de datos con BLAST

Observaciones:

· Complejidad algorítmica de la programación dinámica: **NxM**

(N y M son las longitudes de las dos secuencias a alinear)

· Conocemos la secuencia de 1,5 millones de proteínas y la de unos 22 millones de ADN (28.000 millones de pdb).

Problema: la programación dinámica es demasiado lenta para buscar homólogos en las bases de datos.

Solución: aplicar heurísticas (*truquillos*) para aumentar la velocidad:

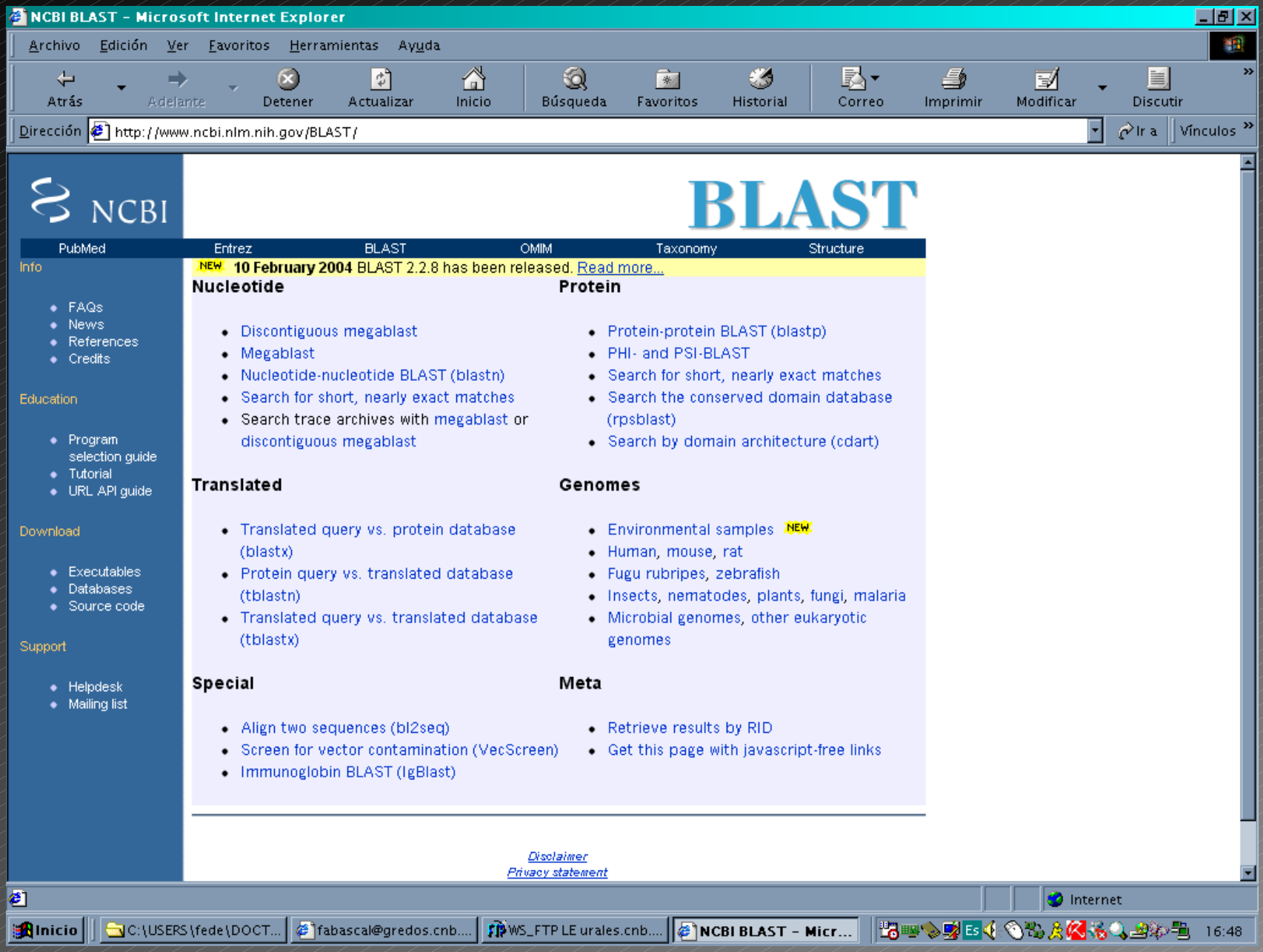
· tablas de dispersión.

· *k*-tuplas.

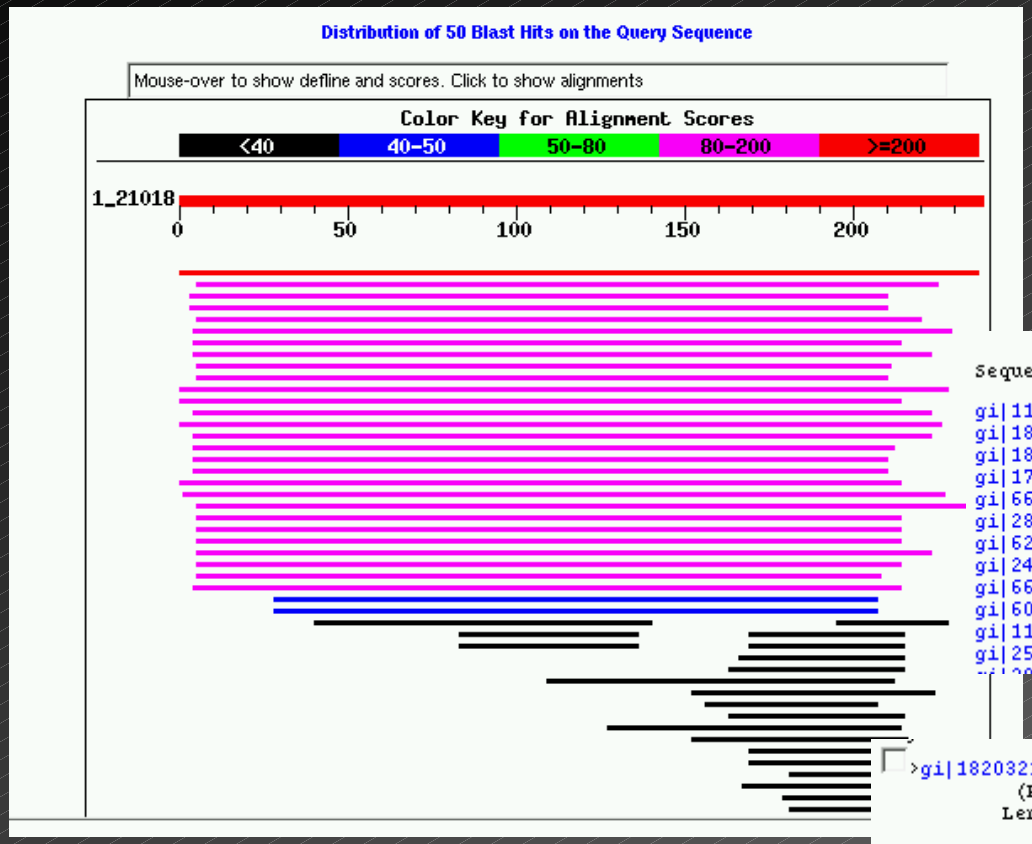
· búsqueda en las diagonales más probables.

Heurística: truquillo que, aunque no garantiza la solución óptima, en la mayoría de los casos funciona.

Búsqueda en bases de datos con BLAST



Búsqueda en bases de datos con BLAST



Sequences producing significant alignments:		Score (bits)	E Value
gi|1173139|sp|P46969|RPE_YEAST	RIBULOSE-PHOSPHATE 3-EPIMERA...	478	e-135
gi|18203210|sp|Q9L0Z5|RPE_STRC0	Ribulose-phosphate 3-epimer...	194	1e-49
gi|18269961|sp|P71676|RPE_MYCTU	Ribulose-phosphate 3-epimer...	184	1e-46
gi|17380276|sp|Q9CCP9|RPE_MYCLE	Ribulose-phosphate 3-epimer...	180	2e-45
gi|6647758|sp|O34557|RPE_BACSU	Ribulose-phosphate 3-epimera...	170	3e-42
gi|2829613|sp|P74061|RPE_SYNY3	Ribulose-phosphate 3-epimera...	162	7e-40
gi|6226024|sp|O67098|RPE_AQUAE	Ribulose-phosphate 3-epimera...	158	1e-38
gi|2499728|sp|Q43843|RPE_SOLTU	Ribulose-phosphate 3-epimera...	155	8e-38
gi|6647759|sp|O66107|RPE_TREPA	Ribulose-phosphate 3-epimera...	155	9e-38
gi|6094119|sp|P51012|RPE_RHOCA	Ribulose-phosphate 3-epimera...	154	2e-37
gi|1169387|sp|P32661|RPE_ECOLI	Ribulose-phosphate 3-epimera...	152	5e-37
gi|25091192|sp|Q8K940|RPE_BUCAP	Ribulose-phosphate 3-epimer...	151	9e-37

```
>gi|18203210|sp|Q9L0Z5|RPE_STRC0   Ribulose-phosphate 3-epimerase (Pentose-5-phosphate
          (PPE) (R5P3E)
         Length = 228

 Score =  194 bits (492), Expect = 1e-49
 Identities = 101/225 (44%), Positives = 142/225 (63%), Gaps = 19/225 (8%)

Query: 6   IAPSILASDFANLGCECHKVINAGADWLHIDVMDGHFVPNITLGQPIVTSLRRSVPRPGD 65
           I PSIL++DFA L  E   V   GADWLH+DVMD HFVPN+TLG P+V SL R+   P
Sbjct: 5   INPSILSADFARLADEAKAV--EGADWLHVDVMDNHFVPNLTLGVPVVESLARATDTP-- 60

Query: 66  ASNTEKKPTAFFDCHMMVENPEKWVDDFAKCGADQFTFHYEATQDPLHLVKLIKSKGIKA 125
                    DCH+M+E P++W   + + GA   TFH EA   P+ L + I++KG +A
Sbjct: 61  -----------LDCHLMIEAPDRWAPQYVEAGAGSVTFHAEAAAAPVRLAREIRAKGARA 109

Query: 126 ACAIKPGTSVDVLFELAPHLDMALVMTVEPGFGGQKFMEDMMPKV----ETLRAKFPHLN 181
           + A+KP T V+   +L P LDM L+MTVEPGFGGQ F++ M+PK+    E ++    L
Sbjct: 110 SMALKPATPVEPYEDLLPELDMLLIMTVEPGFGGQAFLDIMLPKIRRTRELIKKHGLELW 169

Query: 182 IQVDGGLGKETIPKAAKAGANVIVAGTSVFTAADPHDVISFMKEE 226
           +QVDGG+   TI + A AGA+V VAG++V+ A+DP + +  ++ +
Sbjct: 170 LQVDGGVSAATIERCADAGADVFVAGSAVYGASDPAEAVRALRTQ 214
```

Búsqueda en bases de datos con BLAST

Estimación de la confianza de una puntuación o *score.*

Problema: discriminar cuándo un parecido refleja una relación evolutiva de cuándo puede darse por azar.

Factores que afectan a la probabilidad de que por azar, tras una búsqueda, aparezcan alineamientos con una determinada puntuación:

- **la matriz de sustitución**
- **la longitud de las secuencias (el tamaño de la base de datos)**
- **la composición de aminoácidos de las secuencias alineadas**
- **características particulares de las secuencias (sesgos):**
  - ***coiled-coils* (filtro COILS)**
  - **secuencias de baja complejidad. (filtro SEG, filtro DUST)**

El e-value: ***dice cuántas veces esperamos que por azar (en las condiciones de una búsqueda) aparezca un alineamiento con una puntuación igual o mayor que un determinado score.***

$$E = Kmn\, e^{-\lambda S} \qquad (1)$$

Búsqueda en bases de datos con BLAST

E-value: algunos consejos prácticos

· Con bases de datos grandes....

Si e-value < 1e-05: muy-muy fiable
Si 1e-05 < e-value < 0.1: casi siempre son homólogos
Si e-value > 0.1: más arriesgado.

· Lo mejor: el propio criterio.

· La prueba *definitiva* de la homología: el alineamiento múltiple, buscar con métodos más sofisticados (p.e. PSI-BLAST), la estructura de las proteínas, etc.

· En cuanto a los filtros, lo mejor es probar con y sin filtrado y determinar si en el caso concreto resultan útiles.

¿Cómo comparar las secuencias?

-por pares

- -alineamiento de dos secuencias
- -búsqueda en bases de datos con BLAST.

-muchas a la vez

- -alineamiento múltiple.

-con patrones, perfiles y hmm's

- -búsqueda en bases de datos con PSI-BLAST.
- -bases de datos de interés:
  - · PROSITE
  - · PFam
  - · InterPro

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

Objetivo: alinear muchos homólogos al mismo tiempo.

Motivación:

· incluimos más información => alineamientos mejores.

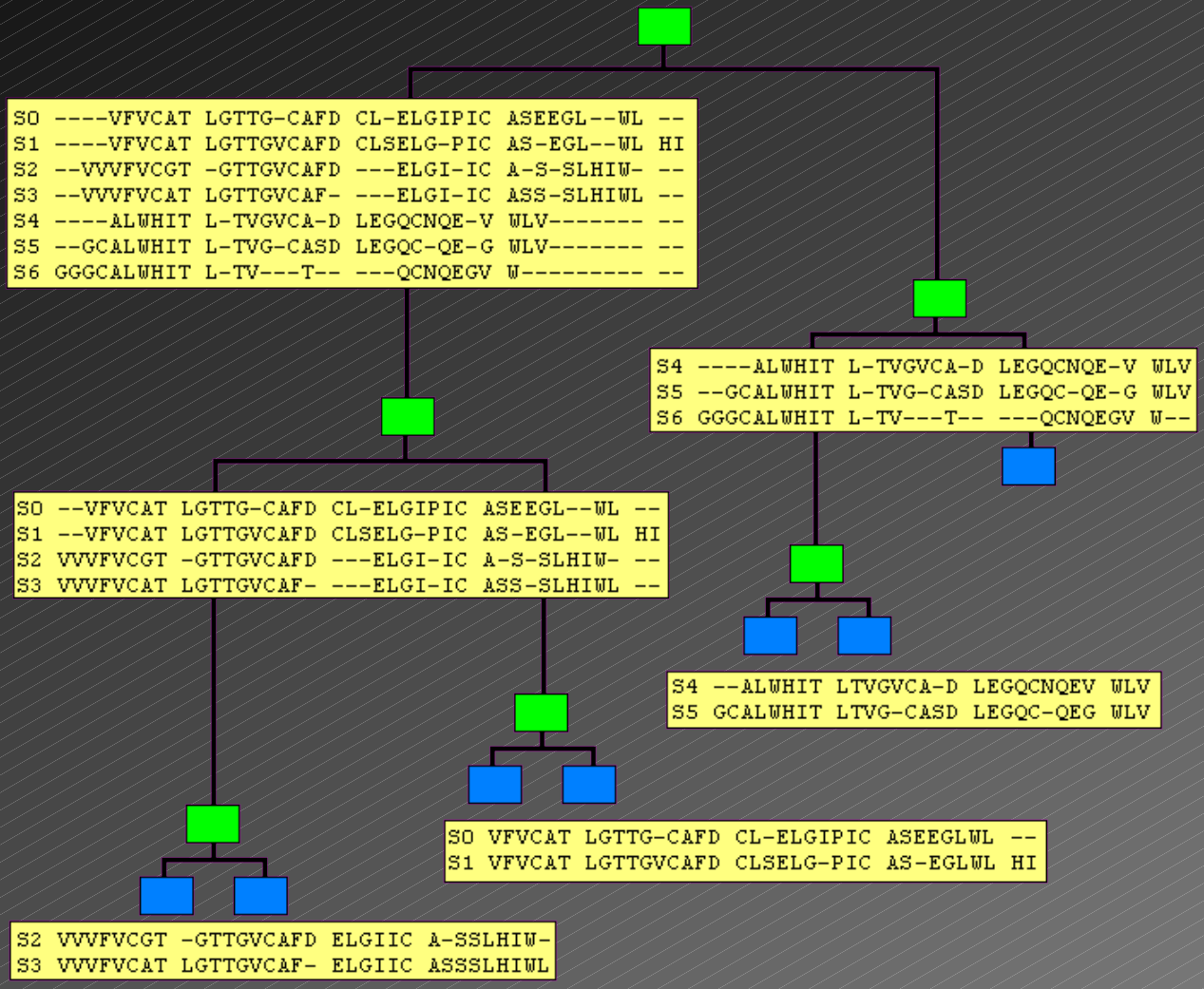
· el alineamiento múltiple nos indica **qué posiciones son más importantes**.

Problema:

· Si la complejidad comput. de alinear dos secuencias es NxM, la de alinear tres es: NxMxL. Si alinear dos sec. (de 300 aa) tardase 1 segundo, alinear tres tardaría 300... y alinear 10 tardaría 300^8 segundos (más que la edad del universo.

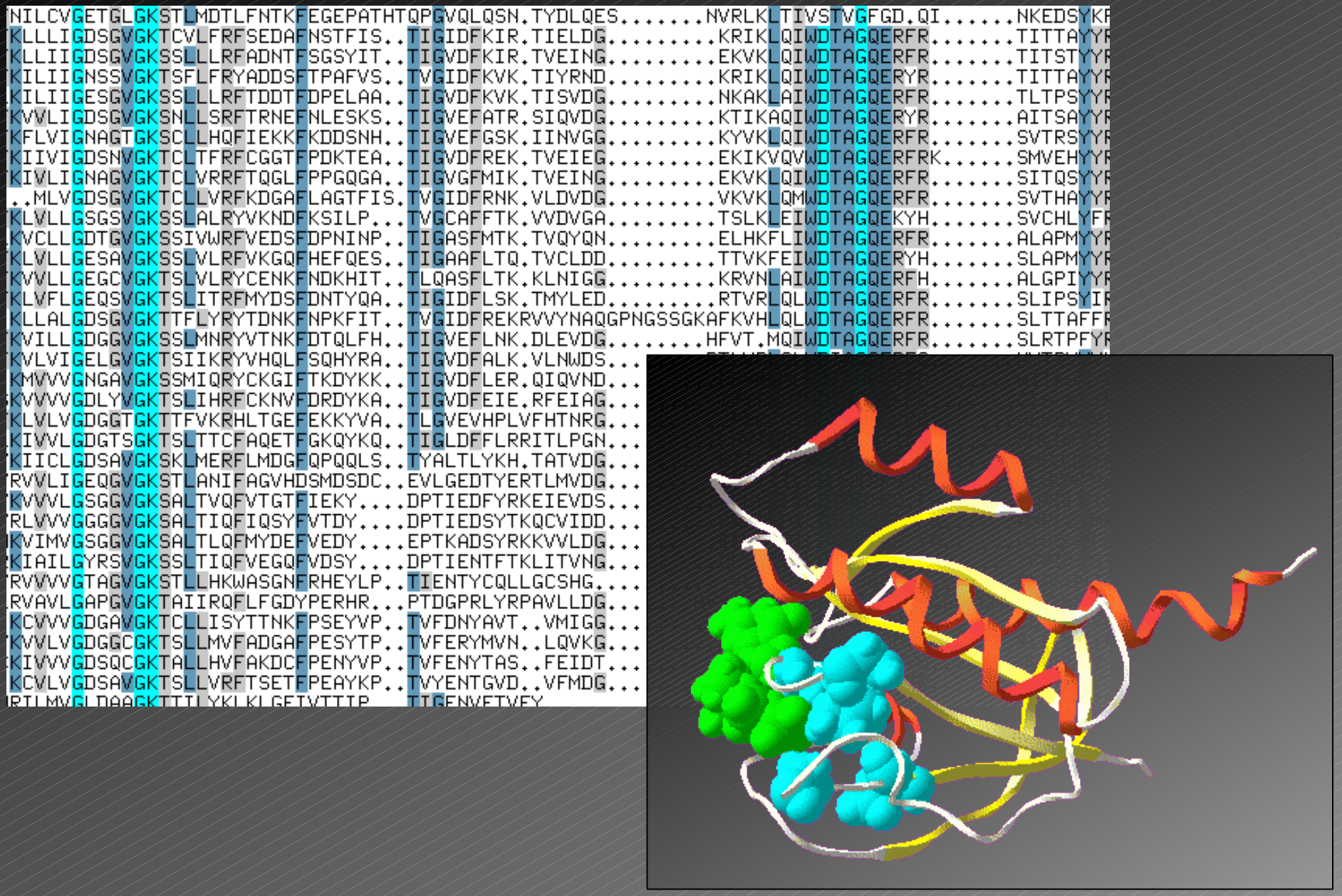
Solución: aplicar heurísticas. Ejemplos: ClustalW, Muscle, T-coffee.

Alineamiento múltiple



Tomado de una presentación de Alberto Pascual (CNB)

Alineamiento múltiple



De los homologos al alineamiento multiple y del alineamiento multiple a los homologos.

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 de alineamiento múltiple?

¿Cómo comparar las secuencias?

-por pares

- -alineamiento de dos secuencias
- -búsqueda en bases de datos con BLAST.

-muchas a la vez

- -alineamiento múltiple con Clustalw.

-con patrones, perfiles y hmm's

- -búsqueda en bases de datos con PSI-BLAST.
- -bases de datos de interés:
  - · PROSITE
  - · PFam
  - · InterPro

Métodos sofisticados de búsqueda de homólogos

¿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]

-Perfiles y perfiles hmm

Métodos sofisticados de búsqueda de homólogos

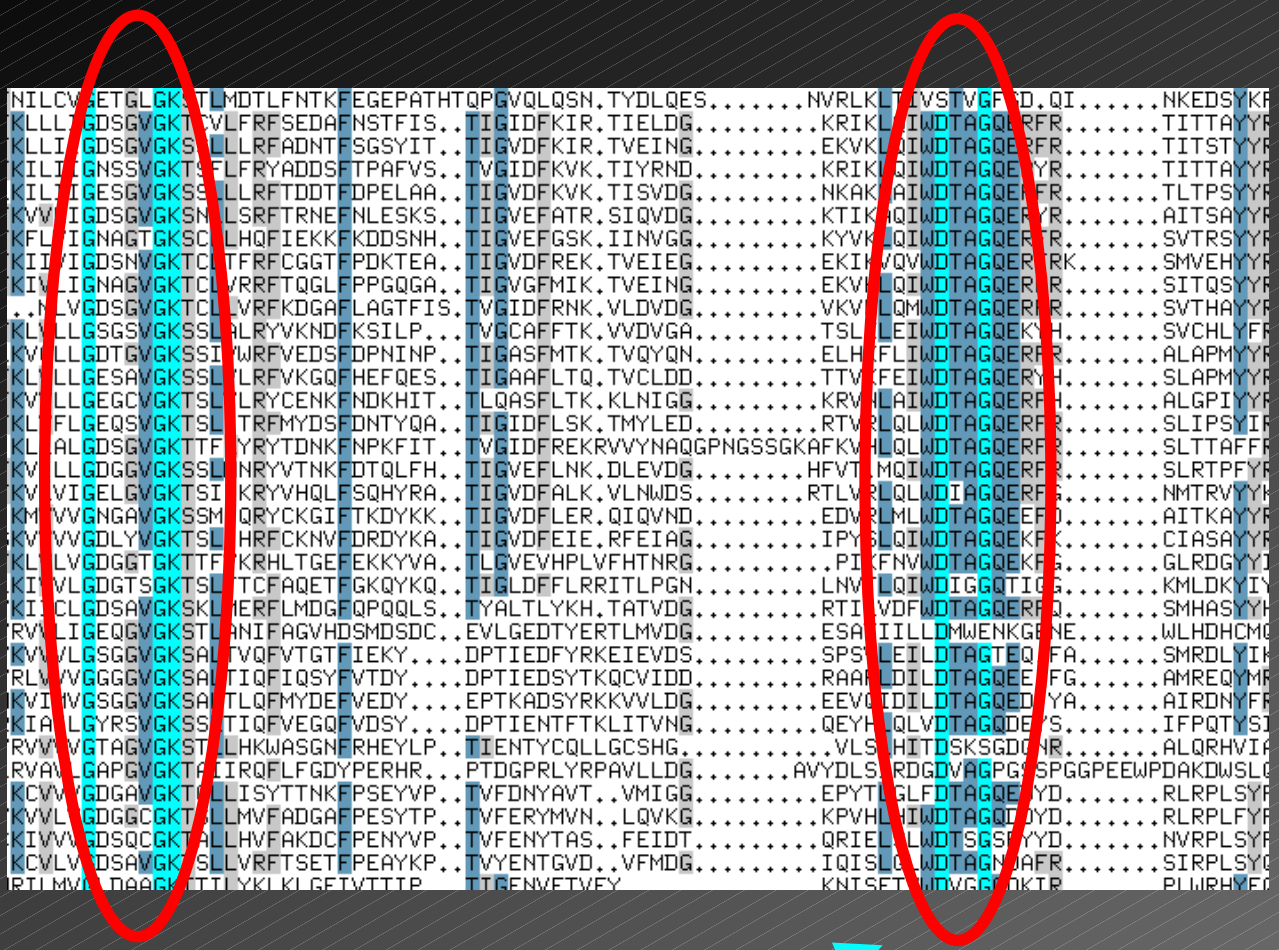
¿Cómo expresarse *regularmente*?

- 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}

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

Métodos sofisticados de búsqueda de homólogos

Perfiles (o PSSM): son matrices de sustitución (como BLOSUM) específicas de posición.

alin. múltiple

perfil

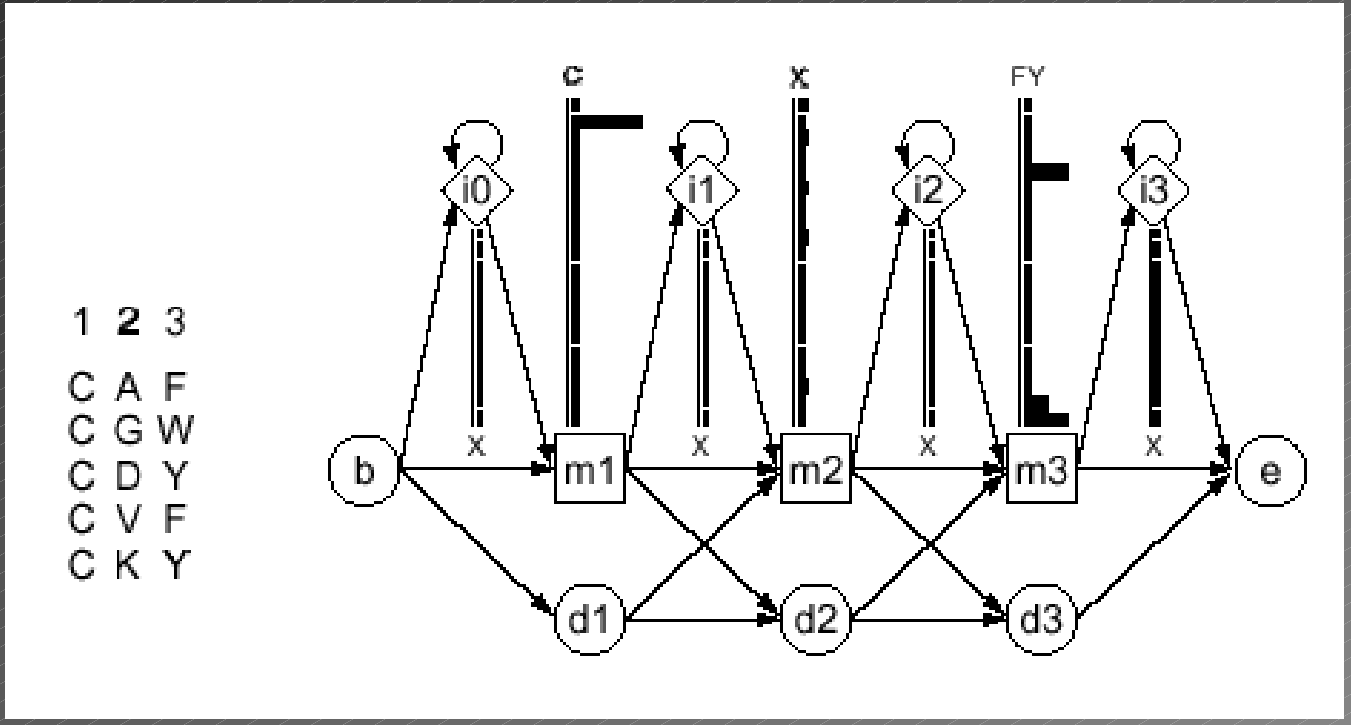
	F	K	L	L	S	H	C	L	L	V
	F	K	A	F	G	Q	T	M	F	Q
	Y	P	I	V	G	Q	E	L	L	G
	F	P	V	V	K	E	A	I	L	K
	F	K	V	L	A	A	V	I	A	D
	L	E	F	I	S	E	C	I	I	Q
	F	K	L	L	G	N	V	L	V	C
A	-18	-10	-1	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	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

Métodos sofisticados de búsqueda de homólogos

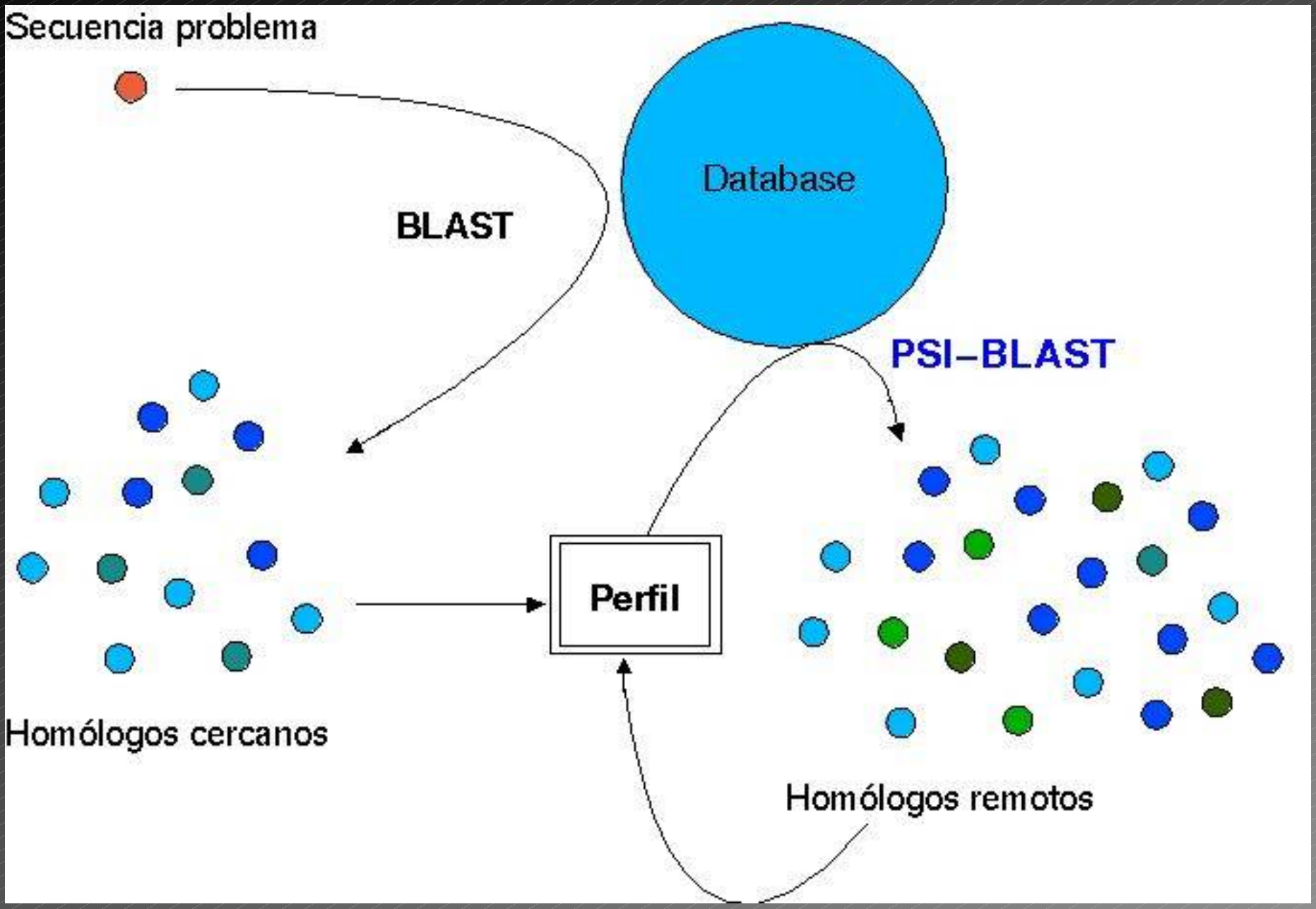
Perfiles de tipo HMM (*hidden markov model)*

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

Los HMM son más sólidos (y complejos)



Búsqueda de homólogos con PSI-BLAST



Búsqueda de homólogos con PSI-BLAST

Demostración del funcionamiento de PSI-BLAST.

Página de PSI-BLAST:

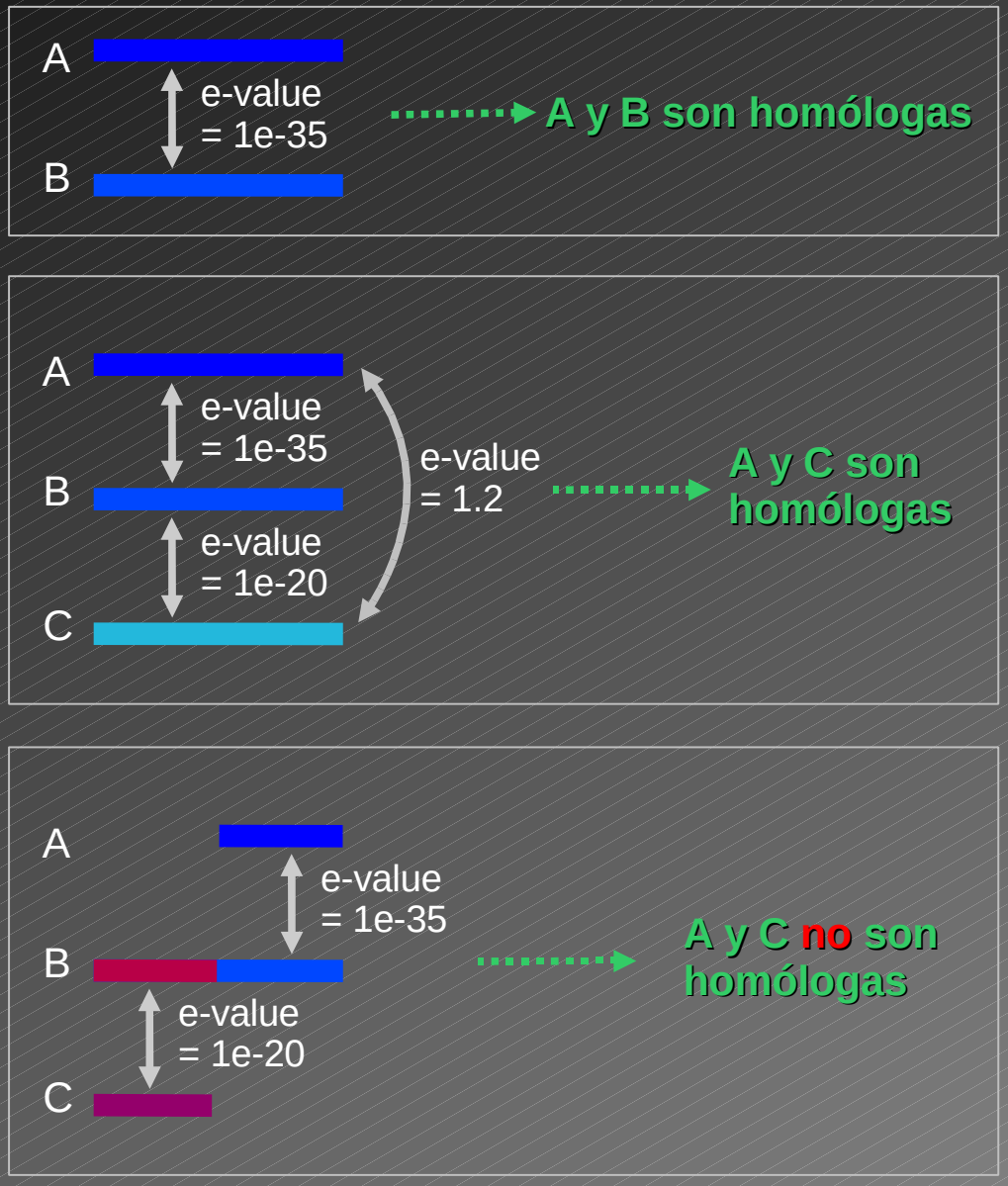
`http://www.ncbi.nlm.nih.gov/BLAST/`

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)

Busqueda con secuencias intermedias



¿Cómo comparar las secuencias?

-por pares

- -alineamiento de dos secuencias
- -búsqueda en bases de datos con BLAST.

-muchas a la vez

- -alineamiento múltiple con Clustalw.

-con patrones, perfiles y hmm's

- -búsqueda en bases de datos con PSI-BLAST.
- -bases de datos de interés:
  - · PROSITE
  - · PFam
  - · InterPro

Bases de datos de interés

Existen muchas bases de datos donde se utilizan patrones y/o perfiles para caracterizar (clasificar, diagnosticar...) familias de proteínas.

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;
//
```

Bases de datos de interés

Pfam:

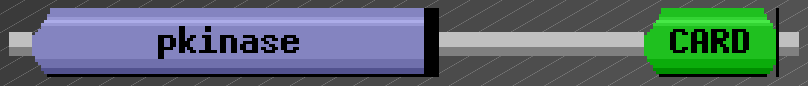
http://www.sanger.ac.uk/Pfam/

-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.

Bases de datos de interés

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

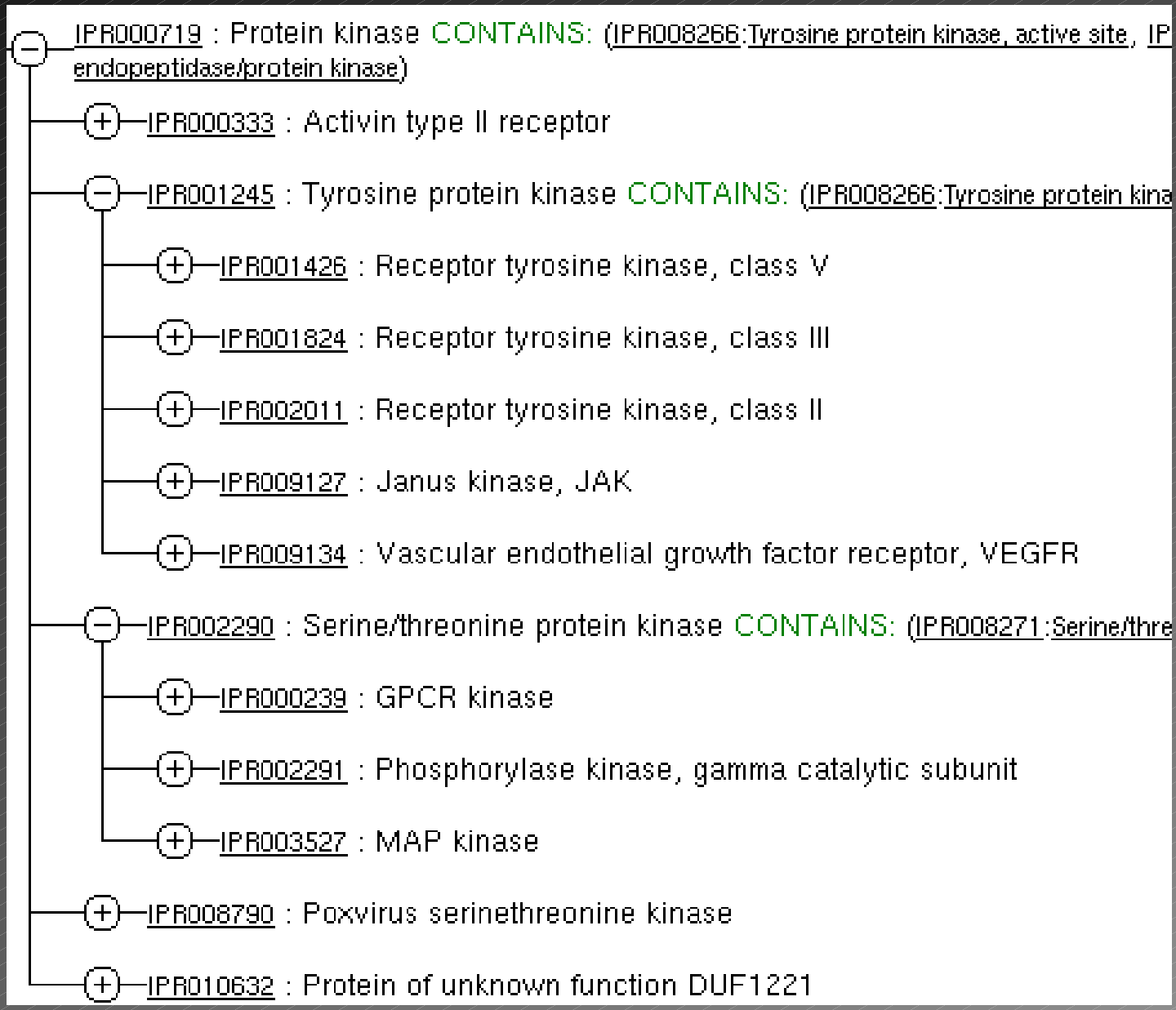
-gran cantidad de información.

-alta cobertura.

Bases de datos de interés

La jerarquía en InterPro:

ejemplo de las kinasas de proteínas.



Extracción de información evolutiva a partir de alineamientos múltiples de proteínas.

Ejemplo basado en el caso de las acetiltransferasas

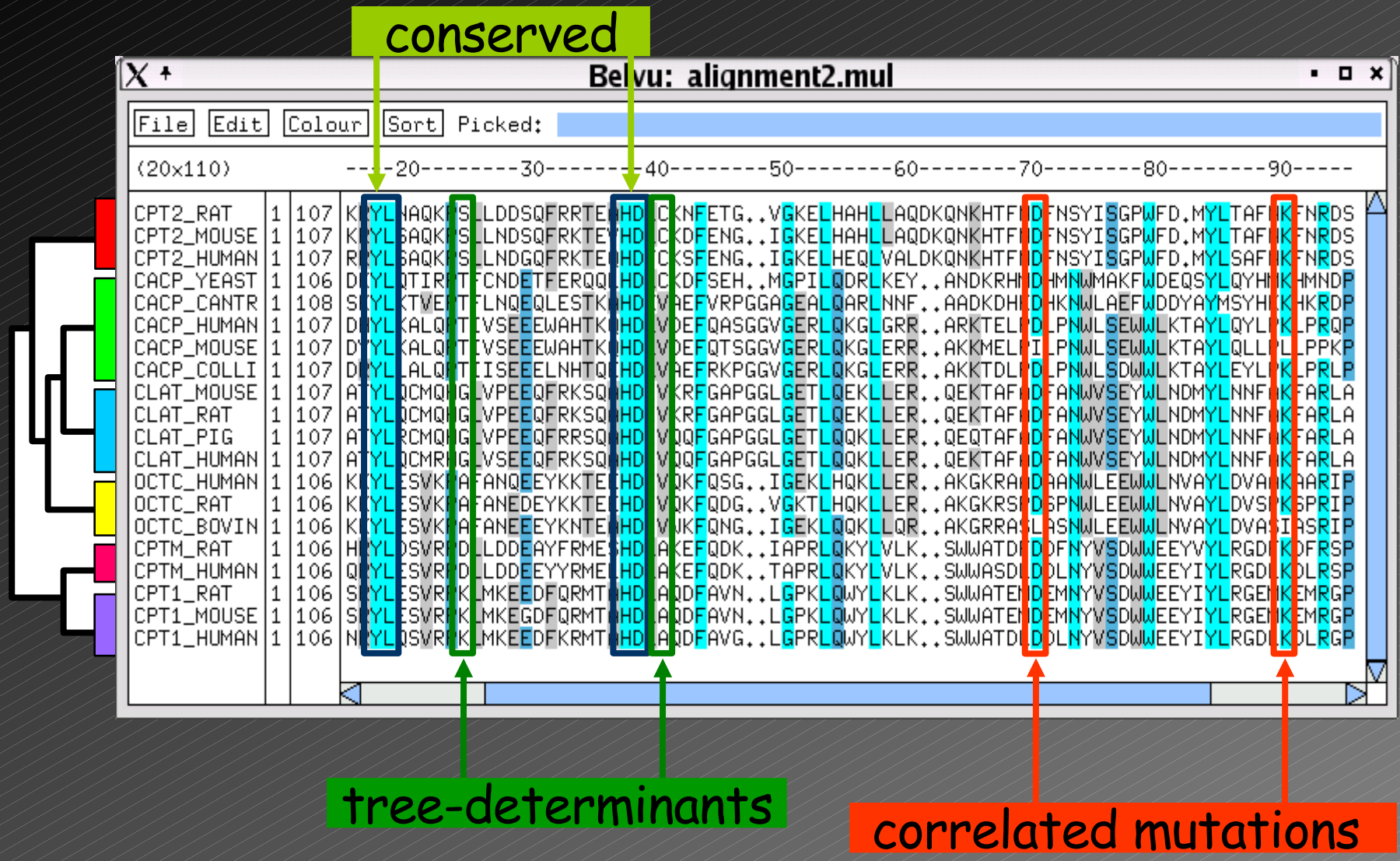
1: Cordente AG, Lopez-Vinas E, Vazquez MI, Gomez-Puertas P, Asins G, Serra D, Hegardt FG.

Mutagenesis of specific amino acids converts carnitine acetyltransferase into carnitine palmitoyltransferase.
Biochemistry. 2006 May 16;45(19):6133-41.
PMID: 16681386 [PubMed - indexed for MEDLINE]

2: Cordente AG, Lopez-Vinas E, Vazquez MI, Swiegers JH, Pretorius IS, Gomez-Puertas P, Hegardt FG, Asins G, Serra D.

Redesign of carnitine acetyltransferase specificity by protein engineering.
J Biol Chem. 2004 Aug 6;279(32):33899-908. Epub 2004 May 21.
PMID: 15155769 [PubMed - indexed for MEDLINE]

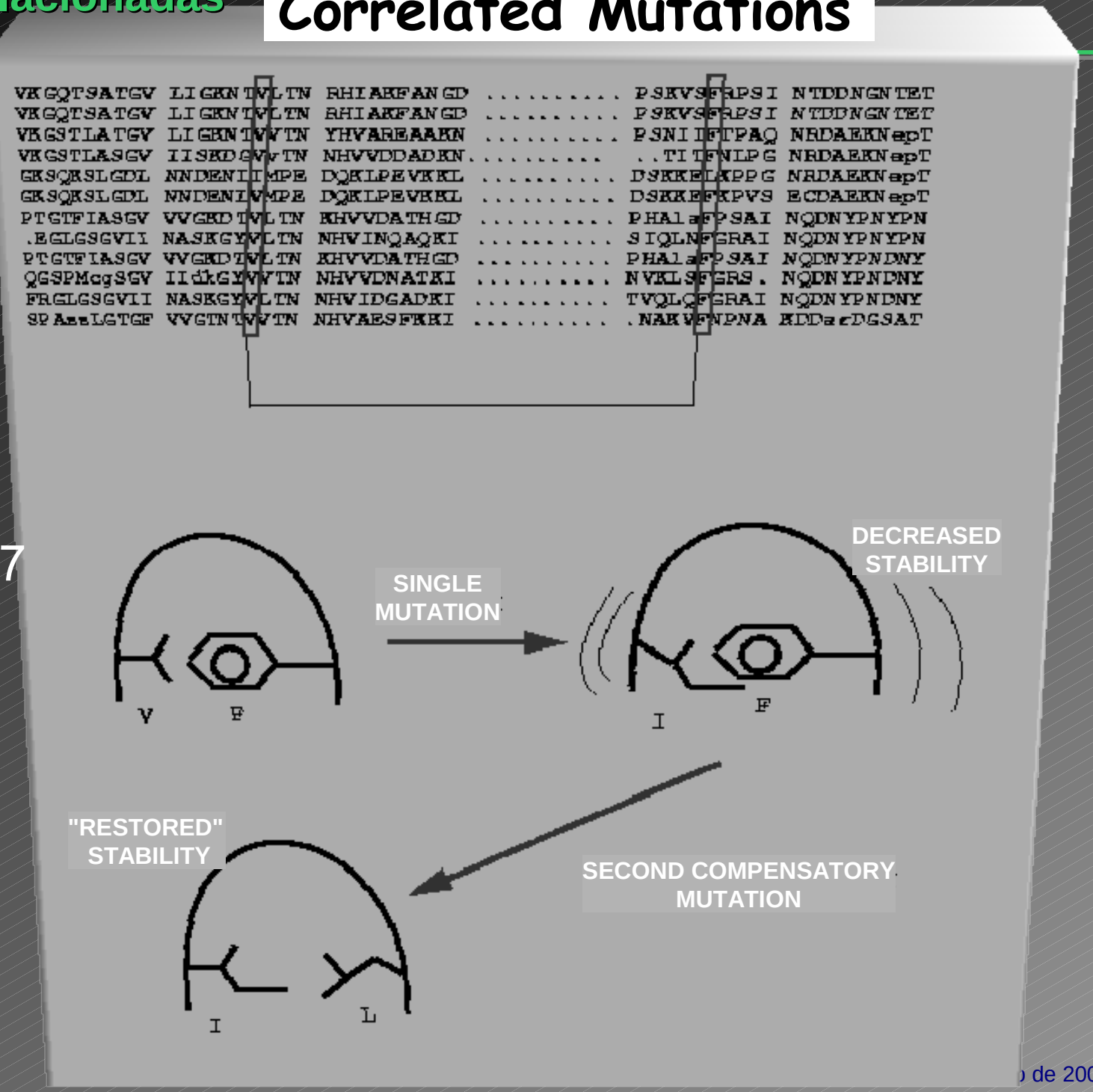
Extracción de información evolutiva a partir de alineamientos múltiples de proteínas



Information extracted from multiple sequence alignments

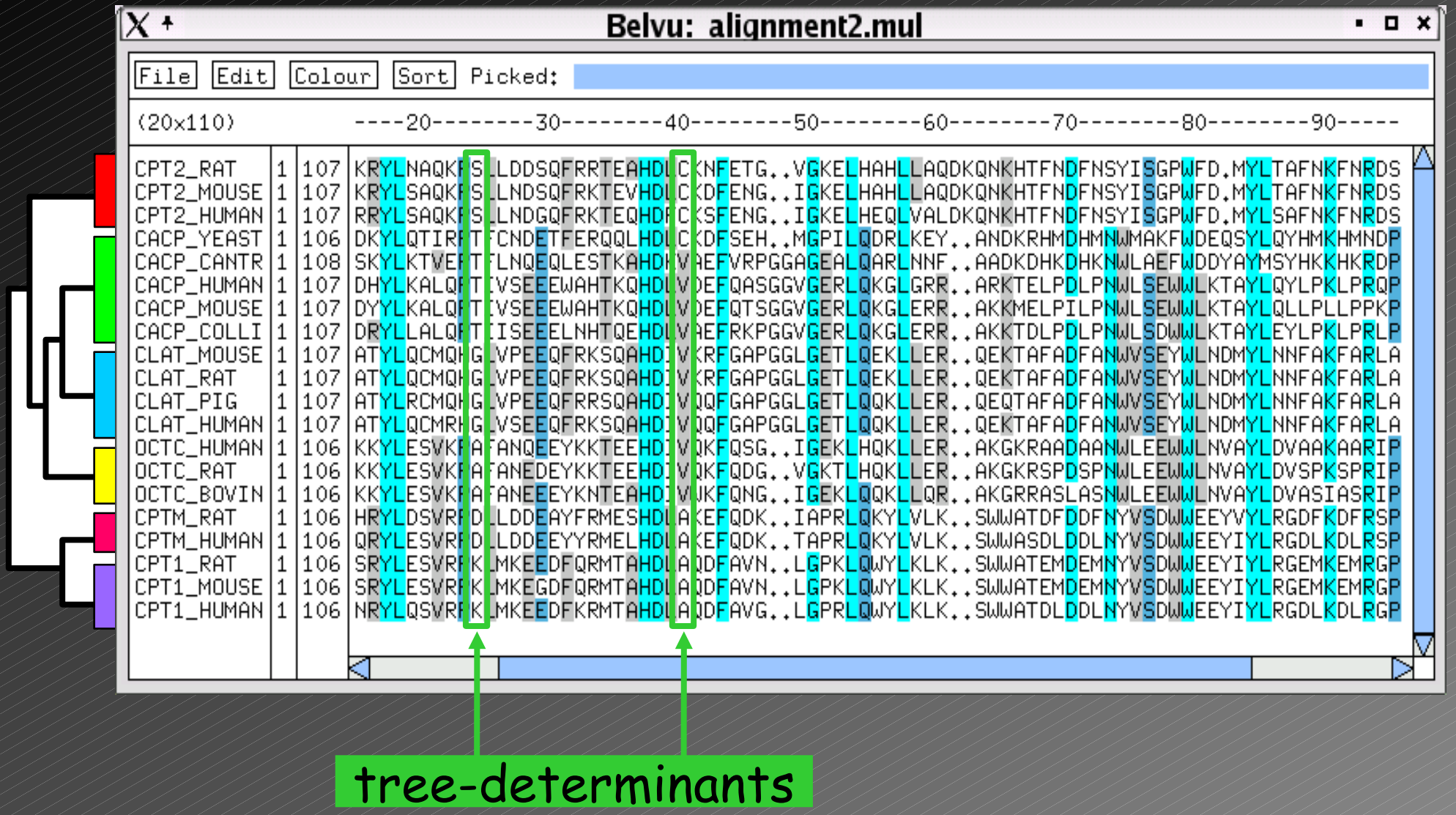
Mutaciones correlacionadas

Correlated Mutations



Pazos et al.
J. Mol. Biol., 1997

Extracción de información evolutiva

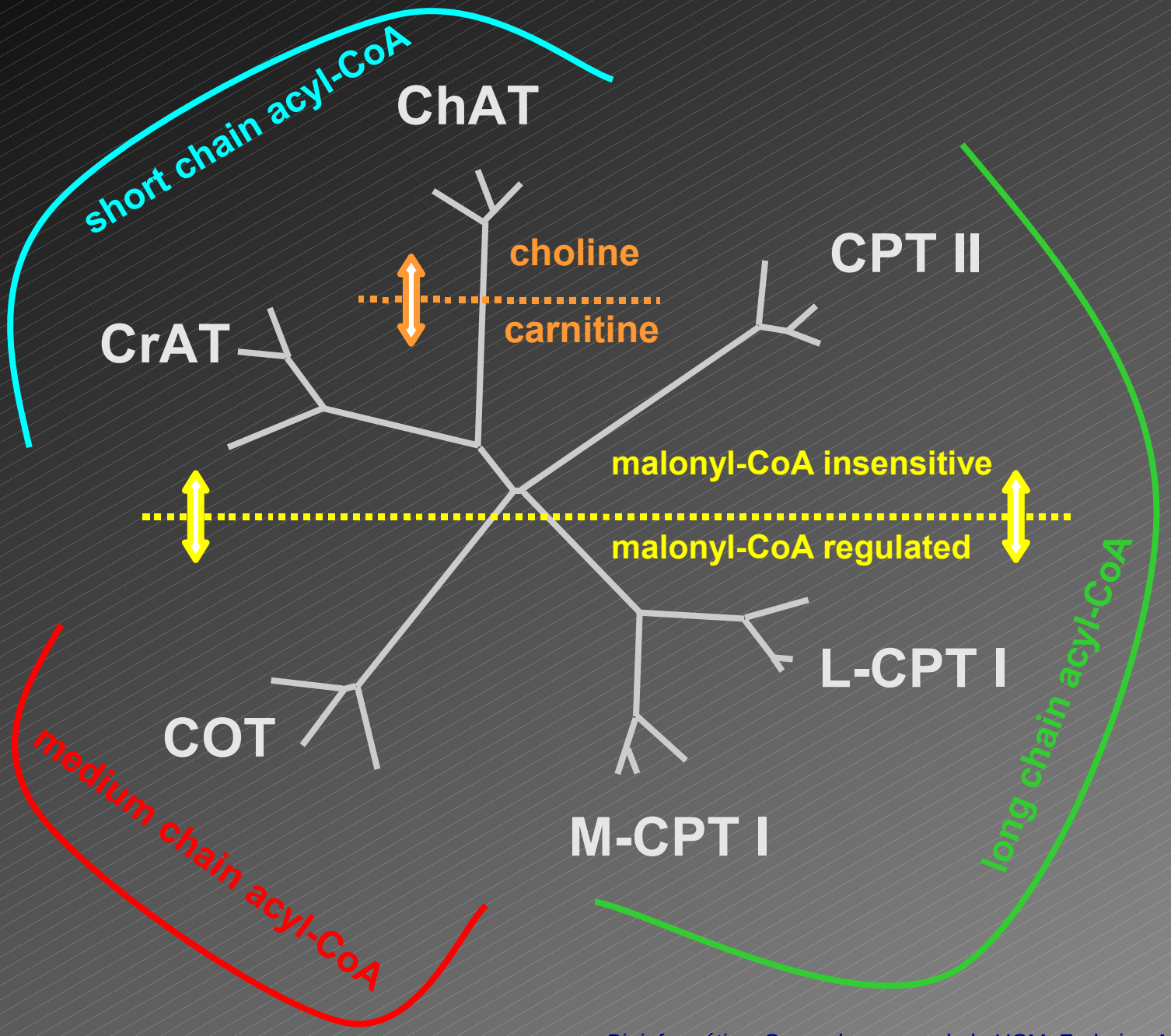


tree-determinants

Information extracted from multiple sequence alignments

Carnitine/choline acyl transferases

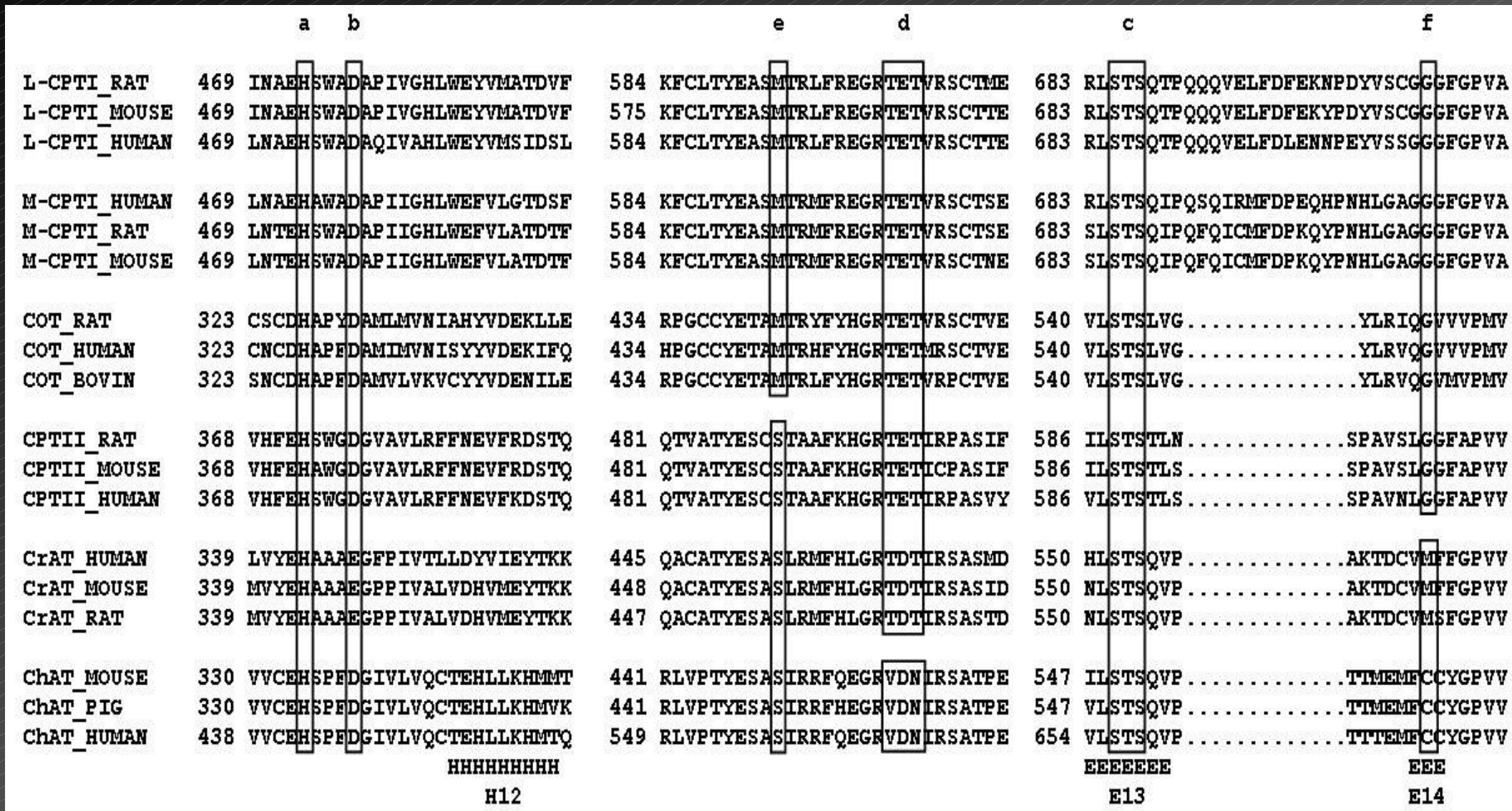
F.G. Hegardt

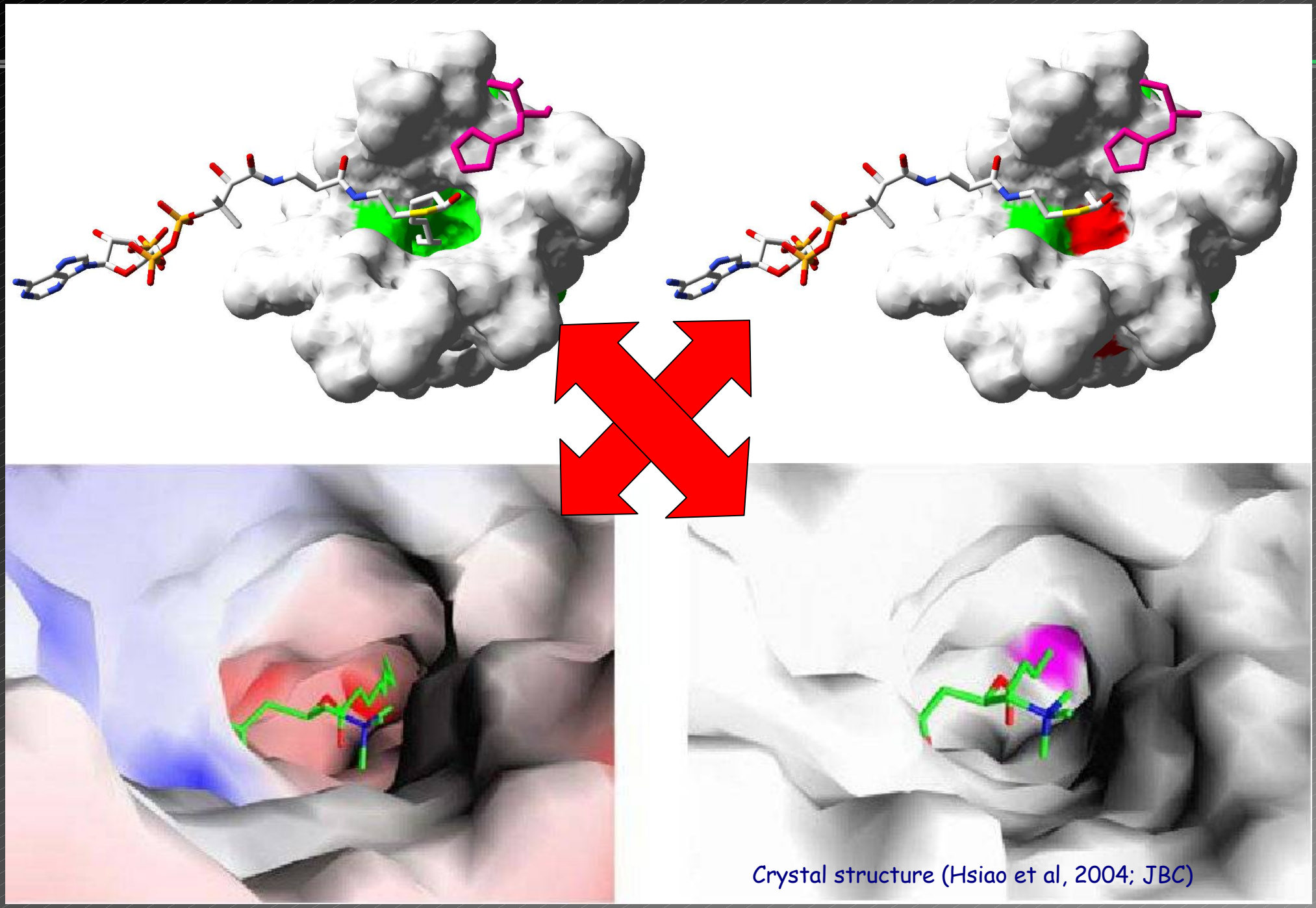


Carnitine-Choline: Thr/Glu/Thr vs. Val/Asp/Asn

Malonyl-CoA regulation: Met vs. Ser

Short vs. Long substrate: Gly vs. Met




Crystal structure (Hsiao et al, 2004; JBC)

¿Cómo comparar secuencias? - Resumen

-por pares

- -alineamiento de dos secuencias
- -búsqueda en bases de datos con BLAST.

-muchas a la vez

- -alineamiento múltiple con Clustalw.

-con patrones, perfiles y hmm's

- -búsqueda en bases de datos con PSI-BLAST.
- -bases de datos de interés:
  - · PROSITE
  - · PFam
  - · InterPro

Agradecimientos

Algunas figuras han sido tomadas de...

-Paulino Gómez Puertas — *Centro de Biología Molecular "Severo Ochoa"*

-Eduardo López-Viñas — *Centro de Biología Molecular "Severo Ochoa"*

-Alberto Pascual — *UCM - Centro Nacional de Biotecnología*

-Manuel José Gómez — *Centro de Astrobiología*