

Alineamiento múltiple de secuencias

Algoritmos exactos, heurísticas, información contenida en un alineamiento

Fernán Agüero

Instituto de Investigaciones Biotecnológicas
Universidad Nacional de General San Martín

Qué es un alineamiento multiple?

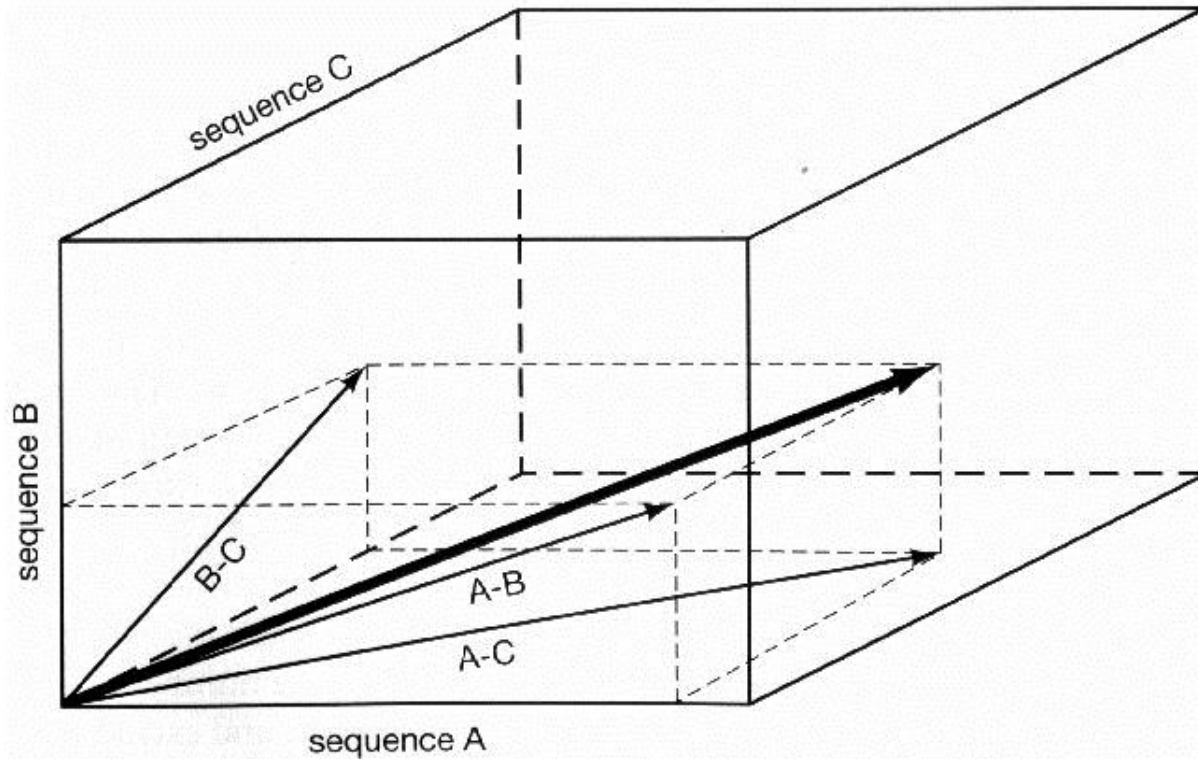
```
FHIT_HUMAN  -----MS-F RFGQHLLKP-SVVFL KTELSFALVNRKPVV PGHVLV...
APH1_SCHPO  -----MPKQ LYFSKFPPVG-SQVFY RTKLSAAFVNLKPIL PGHVLV...
HNT2_YEAST  MILSKTKKPKSMNKP IYFSKFLVT-EQVFY KSKYTYALVNLKPIV PGHVLI...
Y866_METJA  -----MCIF CKIINGEIP-AKVVY EDEHVLAFLDINPRN KGHTLV...
```

Y cómo sería un algoritmo de alineamiento multiple?

Un método de alineamiento múltiple debe alinear
todas las secuencias al mismo tiempo.

True multiple alignment

- **Cómo se resuelve un alineamiento múltiple de 3 secuencias?**
- **Usando dynamic programming en una matriz tridimensional**
- **El problema es el mismo: encontrar el camino óptimo en el espacio**



Multiple alignment

```
FHIT_HUMAN  -----MS-F RFGQHLLKP-SVVFL KTELSEALVNRKPVV PGHVLV...
APH1_SCHPO  -----MPKQ LYFSKFPGV-SQVFY RTKLSAAFVNLKPIL PGHVLV...
HNT2_YEAST  MILSKTKKPKSMNKP IYFSKFLVT-EQVFY KSKYTYALVNLKPIV PGHVLI...
Y866_METJA  -----MCIF CKIINGEIP-AKVYV EDEHVLAFLDINPRN KGHTLV...
```

**Un método de alineamiento múltiple verdadero,
alinea todas las secuencias al mismo tiempo.**

**Pero no existe un método computacional que pueda
realizar esto en tiempo razonable para más de 3
secuencias cortas**

Complejidad del algoritmo DP

- El número de comparaciones que DP tiene que hacer para llenar la matriz (sin usar heurísticas y excluyendo gaps) es el producto de las longitudes de las dos secuencias
- La complejidad del algoritmo crece en forma exponencial con el número de secuencias
- Alinear dos secuencias de longitud 300 implica realizar 90,000 comparaciones
- Alinear tres secuencias de longitud 300 implica realizar 27,000,000 comparaciones

MSA: global optimal MSAs

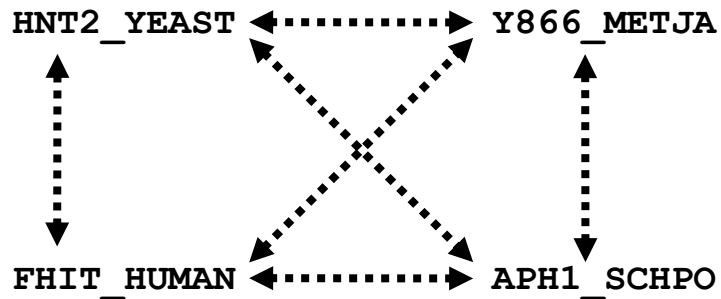
- Needleman-Wunsch o Smith Waterman extendido a una matriz *n-dimensional*
- MSA (Lipman et al. 1989)
 - <http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/msa.html>
 - **Multidimensional dynamic programming**
 - **Usa heurísticas para reducir el espacio de búsqueda**
 - **Varios programas:**
 - msa_50_150 - Alinea no más de 50 secuencias. (c/u < 150 residuos)
 - msa_25_500 - Alinea no más de 25 secuencias (c/u < 500 residuos)
 - msa_10_1000 - Alinea no más de 10 secuencias (c/u < 1000 residuos)
- **Otras heurísticas**
 - **Divide and conquer**
 - Progressive Multiple Sequence Alignments
 - Iterative MSAs
 - Etc.

MSA: progressive multiple alignments

- Alinear todas las secuencias de a pares
- Usar los scores para construir un árbol filogenético
- Alinear secuencialmente (siguiendo el orden que sugiere el árbol) las secuencias para producir un MSA
- No es un verdadero MSA
- Las secuencias **siempre** se alinean de a pares

MSA: progressive multiple alignments

Align all pairs of sequences.

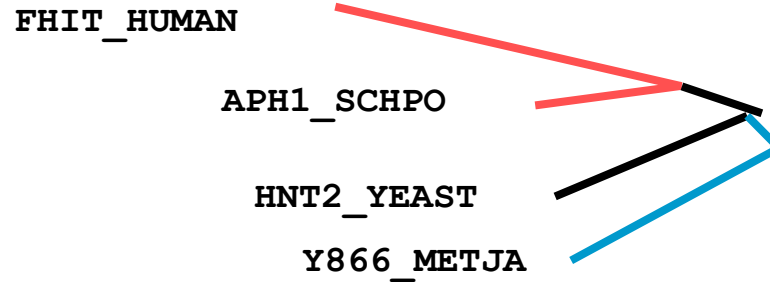


Pairwise alignments: compute distance matrix

	FHIT_HUMAN	APH1_SCHPO	HNT2_YEAST	Y866_METJA
FHIT_HUMAN				
APH1_SCHPO	395			
HNT2_YEAST	316	380		
Y866_METJA	290	300	340	

Progressive multiple alignments

Guide Tree




Pairwise alignments: compute distance matrix

	FHIT_HUMAN	APH1_SCHPO	HNT2_YEAST	Y866_METJA
FHIT_HUMAN				
APH1_SCHPO	395			
HNT2_YEAST	316	380		
Y866_METJA	290	300	340	

Multiple alignment

```
FHIT_HUMAN MSFR FGQHLLKP-SVVFL KTELSEALVNRPVV PGHVLV...
APH1_SCHPO MPKQ LYFSKFPVGSQVFY RTKLSAAFVNLPIL PGHVLV...
HNT2_YEAST MILSKTKKPKSMNKPIYFSKFLVTEQVFYKSKYTYALVNLKPIVPGHVLI...
Y866_METJA MCIF CKIINGEIPAKVVYEDEHVLAFLDINPRNKGHTLV...
```



**Alinear las dos
secuencias más
cercanas**

El alineamiento genera un consenso que se utiliza para alinear las secuencias que quedan.

Desde el punto de vista del alineamiento del primer par, el gap puede insertarse en cualquier lugar

Multiple alignment

```
FHIT_HUMAN  -----MSF RFGQHLLKP-SVVFL KTELSFALVNRKPVV PGHVLV...
APH1_SCHPO  -----MPK QLYFSKFPVGSQVFY RTKLSAAFVNLKPIL PGHVLV...
HNT2_YEAST  MILSKTKKPKSMNK PIYFSKFLVTEQVFY KSKYTYALVNLKPIV PGHVLI...
Y866_METJA  MCIF  CKIINGEIPAKVVYEDEHVLAFDINPRNKGHTLV...
```

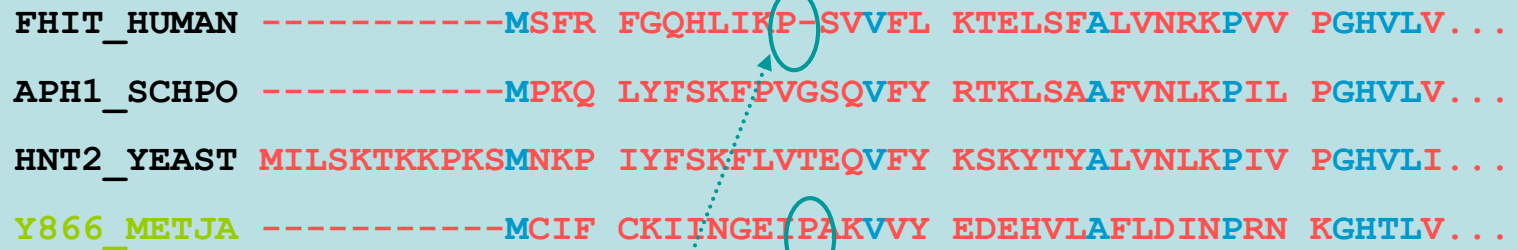


**Alinear las dos
secuencias más
cercanas**

Una vez insertado el gap no
se puede mover porque es
parte del consenso.

Multiple alignment

```
FHIT_HUMAN  -----MSFR FGQHLLKP-SVVFL KTELSFALVNRKPVV PGHVLV...
APH1_SCHPO  -----MPKQ LYFSKEFPVGSQVFY RTKLSAAFVNLKPIL PGHVLV...
HNT2_YEAST  MILSKTKKPKSMNKP IYFSKFLVTEQVFY KSKYTYALVNLKPIV PGHVLI...
Y866_METJA  -----MCIF CKIINGEIPAKVVY EDEHVLAFLDINPRN KGHTLV...
```



**Alinear la
secuencia
siguiente**

Con suerte, el resultado llegue a ser *similar* al resultado que obtenido por un verdadero método de alineamiento múltiple.

Debido al orden de los alineamientos, la posición del gap no puede cambiarse para alinear estas dos Prolinas (lo cual hubiera resultado en un score mayor).

Clustalw is a progressive multiple alignment tool.

- **Adaptive** gap opening and extension scores
- Choice of DNA or protein gap penalty alignments.
- Available on the web or on PC / Mac / unix.

<http://www.ebi.ac.uk/Tools/msa/clustalw2/> (No longer maintained)

<http://dot.imgen.bcm.tmc.edu:9331/multi-align/options/clustalw.html>

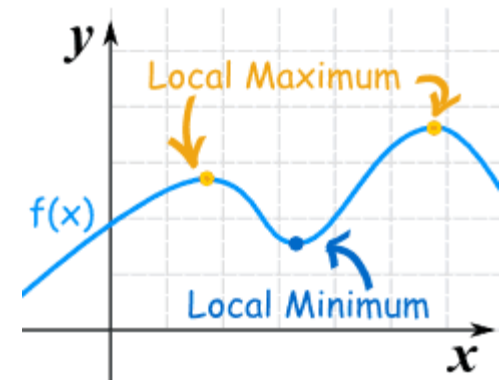
New version ClustalO (Omega)

<http://www.ebi.ac.uk/Tools/msa/clustalo/>

Usa una versión modificada del algoritmo basada en profiles-HMM (se van a ver más adelante en la material)

MSA: métodos iterativos

- Comienzan con un alineamiento multiple inicial
 - Se puede obtener, por ej, usando un método progresivo
- Se optimiza el alineamiento en forma iterativa
- Distintos programas implementan distintas estrategias
- Se realinean subgrupos de secuencias en forma repetida, buscando optimizar el score final del MSA
 - MultAlin (Corpet 1988)
 - PRRP (Gotoh, 1996)
 - DIALIGN (Morgenstern et al. 1996)
 - SAGA (algoritmo genético)
- Como todos los métodos de optimización, pueden quedar atrapados en mínimos locales



- **SAGA (Notredame & Higgins, 1996)**
 - **Sequence Alignment by Genetic Algorithm**
 - **Genera diferentes MSAs por rearrreglos que simulan inserciones de gaps similares a los que ocurren durante la replicación del DNA**
 - **El proceso continúa hasta que converge en un score que no puede ser mejorado**
 - **Los MSAs no tienen garantía alguna de ser óptimos**
 - **Sin embargo, los alineamientos que produce este método son similares a los que se obtienen por otros métodos**

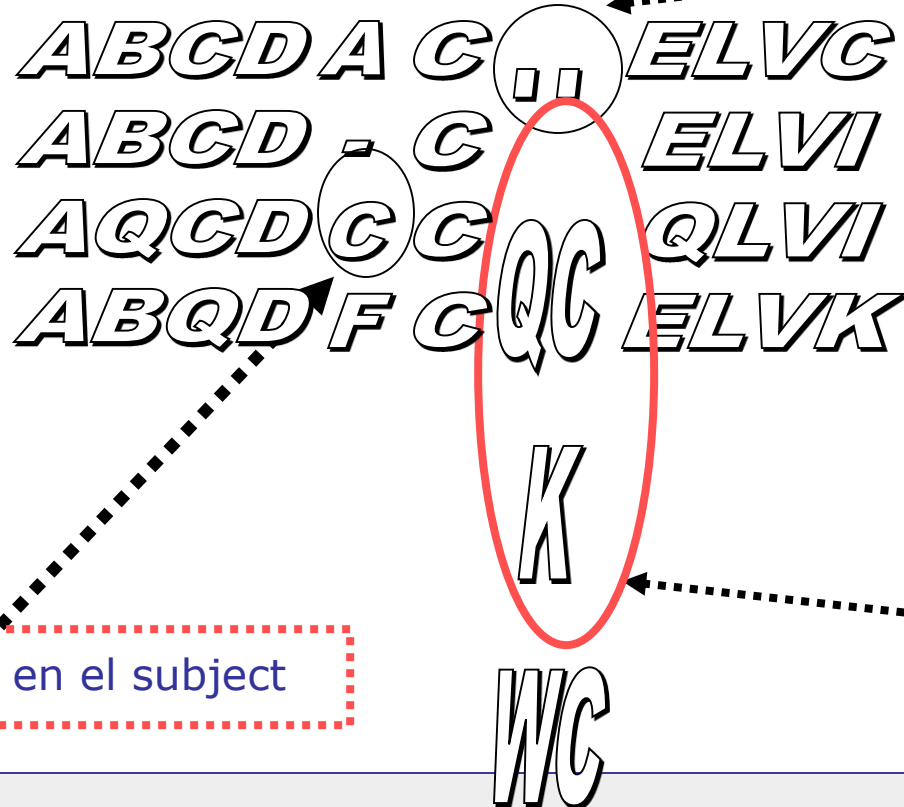
Query-anchored alignments (master slave)

Clustalw:

Produce MSAs

Blast:

No produce MSAs, pero puede mostrar los alineamientos de a pares de una forma que parece un alineamiento múltiple, aunque todas las secuencias estén alineadas con la primera.!



Los gaps en el query quieren decir que nada se pudo alinear en este lugar.

Esta columna no está alineada. Se muestra por conveniencia

Bases de datos de alineamientos

- **Pir-ALN (obsolete)**

- <http://www-nbrf.georgetown.edu/pirwww/search/textpiraln.html>
- **Alineamientos anotados derivados de PIR**
- **Incluye alineamientos al nivel de superfamilia, familia y dominio**
- **3983 alineamientos, 1480 superfamilias, 371 dominios**

- **Protomap (obsolete)**

- <http://www.protomap.cs.huji.ac.il>
- **Clasificación automática de proteínas en Swissprot en grupos (clusters) de proteínas relacionadas**
- **Tiene organización jerárquica para distinguir sub y super familias**

- **COG (still available)**

- <http://www.ncbi.nlm.nih.gov/COG>
- **Clusters of Orthologous Groups of Proteins**
- **Proteomas completos**
- **Contiene alineamientos de cada COG**



- **BLOCKS**

- *Blocks are ungapped multiple sequence alignments representing conserved protein regions*
- <http://blocks.fhcrc.org/blocks> (no existe mas)
- **Representan regiones conservadas de un MSA global**
- **No incluyen gaps**
- **Una serie de blocks conservados pueden describir la pertenencia o no a una familia**
- **Pueden buscar usando una secuencia**
- **Pueden usar un MSA para generar blocks**

Información representada en un MSA

- **Un MSA contiene información acerca de las secuencias que lo componen**
- **Si representa a una familia de proteínas:**
 - **regiones conservadas**
 - **residuos conservados**
- **Qué cosas podemos hacer con esta información?**
 - **Muchas**
- **Qué cosas no deberíamos hacer con esta información?**
 - **Generar un consenso**

Consensos

- Un consenso derivado de un MSA contiene para cada posición el residuo más frecuente

OPS2_DROME	MERSHLPETP	FDLAHSGP--	RFQ-AQSSGN	GSV---LDNV	LPDMAHLVNP
OPS2_DROPS	MERSLLPEPP	LAMALLGP--	RFE-AQTGGN	RSV---LDNV	LPDMAPLVNP
OPS2_LIMPO	-----	-MANQLSY--	SSLGWPYQPN	ASV---VDTM	PKEMLYMIHE
OPS2_HEMSA	----MTNATG	PQMAYYGA--	ASMDFGYPEG	VSI---VDFV	RPEIKPYVHQ
OPS2_SCHGR	-----	-MVNTTDFYP	VPAAMAYESS	VGLPLLGNV	PTEHLDLVHP
OPS2_PATYE	----MPFPLN	RTDTALVISP	SEFRIIGIFI	SICCIIGVLG	NLLIIIVFAK
Consenso	MERSMLPETP	?MMA?LGP?P	...		

Problemas!

Usos de los MSAs

- **Para extraer / generar**
 - **Patterns/Motifs**
 - **Profiles**
 - **Fingerprints**
 - **Position Specific Scoring Matrices**
 - **HMMs**
- **Para qué extraer / generar patterns, motifs, etc, etc?**
 - **Para clasificar**
 - **Para alinear secuencias**
 - **Para buscar secuencias similares por métodos más sensibles**

Webster's New Collegiate Dictionary:

mo-tif *n*[F, motive, motif] **1 a:** a usu. recurring salient thematic element in a work of art; *esp:* a dominant idea or central theme

- En secuencias biológicas un **motif** es un patrón recurrente (común) en una serie de secuencias relacionadas
- Los MSAs permiten distinguir regiones de evolución lenta (conservadas) y otras de evolución más rápida en un grupo de secuencias
- Cómo describir/representar las características salientes de un motif?

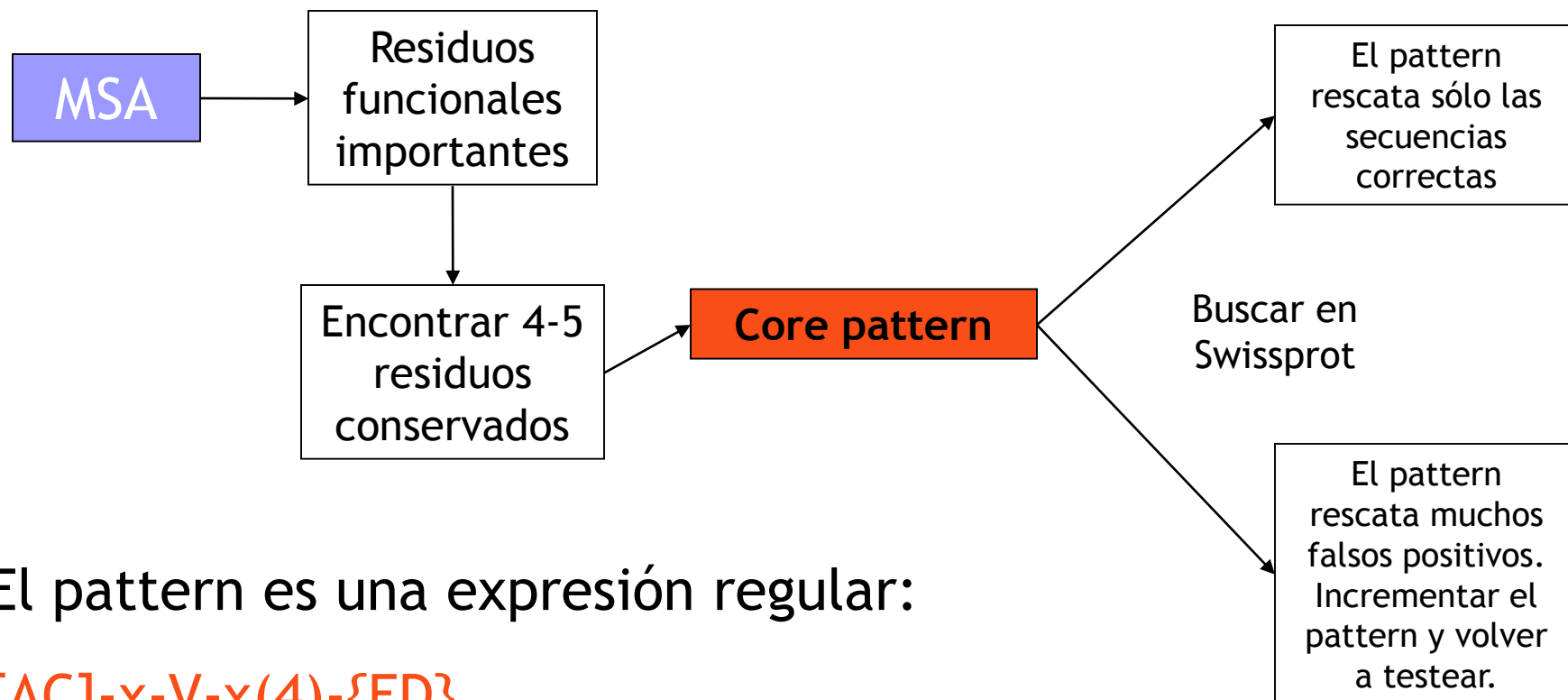
- **Patterns**

- **Descripción (usando una sintaxis particular) de una región corta que tenga relevancia funcional**
- **Cómo se construye un pattern**
 - A partir de la literatura. Se testea contra Swissprot
 - A partir de
 - Enzyme catalytic sites
 - Prosthetic group attachment sites (heme, pyridoxal-phosphate, biotin, etc)
 - Amino acids involved in binding a metal ion
 - Cysteines involved in disulfide bonds
 - Regions involved in binding a molecule (ADP/ATP, GDP/GTP, calcium, DNA, etc.) or another prote



<http://www.expasy.ch/prosite>

Patterns



El pattern es una expresión regular:

[AC]-x-V-x(4)-{ED}

ala/cys-any-val-any-any-any-any-(any except glu or asp)



<http://www.expasy.ch/prosite>

NiceSite view of PROSITE: [PDOC00101](#) (documentation)

Pyruvate kinase active site signature

PROSITE cross-reference(s)

[PS00110:PYRUVATE_KINASE](#)

Documentation

Pyruvate kinase (EC [2.7.1.40](#)) (PK) [1] catalyzes the final step in glycolysis, the conversion of phosphoenolpyruvate to pyruvate with the concomitant phosphorylation of ADP to ATP. PK requires both magnesium and potassium ions for its activity. PK is found in all living organisms. In vertebrates there are four, tissues specific, isozymes: L (liver), R (red cells), M1 (muscle, heart, and brain), and M2 (early fetal tissues). In *Escherichia coli* there are two isozymes: PK-I (gene *pykF*) and PK-II (gene *pykA*). All PK isozymes seem to be tetramers of identical subunits of about 500 amino acid residues.

As a signature pattern for PK we selected a conserved region that includes a lysine residue which seems to be the acid/base catalyst responsible for the interconversion of pyruvate and enolpyruvate, and a glutamic acid residue implicated in the binding of the magnesium ion.

Description of pattern(s) and/or profile(s)

Consensus pattern	[LIVAC]-x-[LIVM](2)-[SAPCV]-K-[LIV]-E-[NKRST]-x-[DEQHS]-[GSTA]-[LIVM] [K is the active site residue] [E is a magnesium ligand]
Sequences known to belong to this class detected by the pattern	ALL.
Other sequence(s) detected in SWISS-PROT	1.

Last update

July 1999 / Pattern and text revised.

References

[1]
Muirhead H.
Biochem. Soc. Trans. 18:193-196(1990).

Copyright

This PROSITE entry is copyright by the Swiss Institute of Bioinformatics (SIB). There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.isb-sib.ch/announce/> or email to license@isb-sib.ch).

Pattern-Hit Initiated BLAST

<http://www.ncbi.nlm.nih.gov/blast/html/PHIsyntax.html>

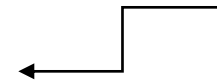
**Combina búsqueda por motivos (usa sintaxis de Prosite)
con BLAST (PSI-BLAST en realidad, ver más adelante)**

Profiles

- Representan un MSA en forma de tabla
- Cada posición en el alineamiento corresponde a una fila en el profile
- Para cada posición en el alineamiento el profile contiene la información de frecuencias de aminoácidos que ocurren en esa posición
- Esta información se encuentra representada en forma de scores y penalties e incluye a gaps
- **Un profile no es otra cosa que una serie de matrices de scoring, una para cada posición en el alineamiento**

MSA

1	2	3	4	5	6	7	8	9
---	---	---	---	---	---	---	---	---



1
2
3
4
5
6
7
8

Profile

Un MSA particular

ATP binding RNA helicase ("DEAD" box family)

rhle_ecoli	GVDVLVA	TPG	RLLDLEHQNAVKLDQV	EILVL	DEADR	MLDMGFIHDI
dbp2_schpo	GVEICIA	TPG	RLLDMLDSNKTNLRRV	TYLVL	DEADR	MLDMGFEPQI
dbp2_yeast	GSEIVIA	TPG	RLIDMLEIGKTNLKR	TYLVL	DEADR	MLDMGFEPQI
dbpa_ecoli	APHIIVA	TPG	RLLDHLQKGTVSLDAL	NTLVM	DEADR	MLDMGFSDAI
rm62_drome	GCEIVIA	TPG	RLIDFLSAGSTNLKRC	TYLVL	DEADR	MLDMGFEPQI
p68_human	GVEICIA	TPG	RLIDFLECGKTNLRR	TYLVL	DEADR	MLDMGFEPQI
rhlb_ecoli	GVDILIG	TTG	RLIDYAKQNHINLGAI	QVVVL	DEADR	MYDLGFIKDI
yn21_caeel	RPHIIVA	TPG	RLVDHLENTK	...GFNLKAL	KFLIM	DEADR	ILNMDFEVEL
yhm5_yeast	KPHIIVA	TPG	RLMDHLENTK	...GFSLRKL	KFLVM	DEADR	LLDMEFGPVL
me31_drome	KVQLIIA	TPG	RILDLMDDKVADMSHC	RILVL	DEADK	LLSLDFQGML
drsl_yeast	RPDIVIA	TPG	RFIDHIRNSA	...SFNVDSV	EILVM	DEADR	MLEEGFQDEL
if4a_rabbit	APHIIVG	TPG	RVFDMNRRYLSPKYI	KMFVL	DEADE	MLSRGFKDQI
if41_human	APHIIVG	TPG	RVFDMNRRYLSPKYI	KMFVL	DEADE	MLSRGFKDQI
vasa_drome	GCHVVIAT	TPG	RLLDVFVDRTFITFEDT	RFVVL	DEADR	MLDMGFSEDM
srm_b_ecoli	NQDIVVA	TTG	RLQYIKEENFDCRAV	ETLIL	DEADR	MLDMGFAQDI
dead_ecoli	GPQIVVG	TPG	RLLDHLKRGLDLSKL	SGLVL	DEADE	MLRMGFIEDV
if4a_orysa	GVHVVG	TPG	RVFDMNRRQLRPDYI	KMFVL	DEADE	MLSRGFKDQI
dead_klepn	GPQIVVG	TPG	RLLDHLKRGLDLSKL	SGLVL	DEADE	MLRMGFIEDV
pl10_mouse	GCHLLVA	TPG	RLVDMMERGKIGLDFC	KYLVL	DEADR	MLDMGFEPQI
p54_human	TVHVVIAT	TPG	RILDLIKKGAKVDHV	QMIVL	DEADK	LLSQDFVQIM
if4a_drome	GCHVVG	TPG	RVYDMINRKLRTQYI	KLFVL	DEADE	MLSRGFKDQI
ded1_yeast	GCDLLVA	TPG	RLNDLLERGKISLANV	KYLVL	DEADR	MLDMGFEPQI
ms16_yeast	RPNIVIA	TPG	RLIDVLEKYS	...NKFFRFV	DYKVL	DEADR	LLEIGFRDDL
pr28_yeast	GCDILVA	TPG	RLIDSLENHLLVMKQV	ETLVL	DEADK	MYDLGFEDQV
if4n_human	GQHVVG	TPG	RVFDMIRRRSLRTRAI	KMLVL	DEADE	MLNKGFEQI
an3_xenla	GCHLLVA	TPG	RLVDMMERGKIGLDFC	KYLVL	DEADR	MLDMGFEPQI
dbp1_yeast	GCDLLVA	TPG	RLNDLLERGKVSLANI	KYLVL	DEADR	MLDMGFEPQI
if4a_yeast	DAQIVVG	TPG	RVFDNIQRRRFRTDKI	KMFIL	DEADE	MLSSGFKEQI
spb4_yeast	RPQILIG	TPG	RVLDFLQMPAVKTSAC	SMVVM	DEADR	LLDMSFIKDT
if4a_caeel	GIHVVG	TPG	RVGDMINRNALDTSRI	KMFVL	DEADE	MLSRGFKDQI
pr05_yeast	GTEIVVA	TPG	RFIDILTLND	.GKLLSTKRI	TFVVM	DEADR	LFDLGFEPQI
if42_mouse	APHIVVG	TPG	RVFDMNRRYLSPKWI	KMFVL	DEADE	MLSRGFKDQI
dhh1_yeast	TVHILVG	TPG	RVLDLASRKVADLSDC	SLFIM	DEADK	MLSRDFKTI
db73_drome	KADIVVT	TPG	RLVDHLHATK	...GFCLKSL	KFLVL	DEADR	IMDAVFQNW
yk04_yeast	GCNFIIG	TPG	RVLDHLQNTK	VIKEQLSQSL	RYIVL	DEGDK	LMELGFDETI
ybz2_yeast	SGQIVIA	TPG	RFLELLEKDN	.TLIKRFSKV	NTLIL	DEADR	LLQDGHFDEF
yhw9_yeast	KPHFIIA	TPG	RLAHHIMSSG	DDTVGGMLRA	KYLVL	DEADI	LLTSTFADHL
glh1_caeel	GATIIIVG	TVG	RIKHFCEEGTIKLDKC	RFFVL	DEADR	MIDAMGFGTD

Un profile generado a partir del MSA

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	Gap	Len	..
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11	100	100	
P	18	0	13	0	0	-12	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1	100	100	
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27	100	100	
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11	100	100	
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8	100	100	
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9	100	100	
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10	100	100	
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10	100	100	
P	31	6	7	6	6	-41	19	11	-9	6	-16	-11	0	89	17	17	24	22	9	-50	-48	12	100	100	
G	70	60	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30	100	100	
! 11																									
R	-30	10	-30	0	0	-50	-30	50	-30	80	-40	20	10	30	40	150	10	-10	-30	140	-60	20	100	100	
L	-2	-17	-15	-18	-12	38	-13	-9	38	-12	49	39	-15	-9	-9	-15	-11	0	38	6	12	-10	100	100	
L	0	-12	-15	-14	-9	32	-12	-7	32	-7	41	35	-11	-9	-6	-12	-9	0	29	6	9	-7	100	100	
D	15	58	-27	78	54	-52	35	27	-12	16	-26	-21	38	6	41	3	9	10	-12	-57	-25	50	100	100	
L	-5	-5	-7	-8	-4	24	-12	13	13	-6	25	17	-1	-7	0	-2	-8	-3	10	11	17	-2	100	100	
L	3	-13	-13	-13	-8	31	-11	-8	34	-9	41	36	-12	-7	-5	-13	-8	2	31	-1	8	-6	100	100	
E	6	19	-15	23	27	-21	9	15	-6	18	-8	-1	16	6	23	12	6	5	-6	-15	-16	25	100	100	
K	3	14	-12	11	12	-16	2	10	-5	23	-7	4	15	6	15	22	8	3	-5	7	-15	14	100	100	
G	11	17	0	16	14	-16	19	5	-6	11	-11	-5	16	9	8	4	14	15	-1	-13	-14	11	100	100	
T	12	9	-1	7	7	-8	9	2	4	12	0	4	10	5	4	3	9	12	7	-8	-8	5	100	100	
! 21																									
D	1	1	0	2	1	-1	1	0	1	0	0	0	1	0	1	0	0	1	2	-3	-1	1	22	22	
T	2	2	0	3	2	-2	3	0	2	0	0	0	1	1	1	-1	1	4	2	-5	-2	2	22	22	
K	0	1	-3	0	1	0	0	0	1	4	1	3	1	0	1	1	0	3	1	0	-2	1	22	22	
G	3	3	0	4	4	-1	6	-1	3	0	1	1	3	1	1	-2	4	3	5	-6	-3	2	22	22	
L	5	-6	-4	-7	-4	16	-2	-4	21	-4	23	17	-5	-4	-4	-8	-2	4	19	0	6	-4	22	22	
B	5	16	-6	15	11	-15	10	6	-3	16	-8	-1	15	4	9	10	12	7	-2	-3	-11	10	100	100	
L	1	-13	-12	-14	-9	27	-8	-7	24	-8	36	30	-10	-5	-7	-10	-4	7	23	6	9	-8	100	100	
D	7	19	-7	22	17	-22	13	7	-6	19	-11	-3	14	8	15	14	17	6	-5	-5	-18	16	100	100	
K	11	10	-3	10	9	-12	5	9	-4	16	-6	0	10	6	11	12	10	4	-4	3	-8	10	100	100	
V	7	-10	11	-11	-10	14	0	-8	31	-11	19	16	-10	0	-10	-12	2	8	34	-22	9	-10	100	100	
K	8	9	-4	9	9	-13	11	1	0	16	-4	4	8	7	8	11	13	12	3	-2	-15	8	100	100	
L	3	4	-9	3	6	3	-2	8	9	7	10	10	5	0	8	3	0	5	7	-2	0	7	100	100	
L	1	-13	-13	-13	-9	32	-11	-7	32	-9	42	36	-12	-7	-6	-13	-9	3	33	2	8	-7	100	100	
*	99	0	25	208	120	94	137	44	181	105	256	94	41	62	64	144	59	99	162	3	35	0			

Usos de los profiles

- **También conocidos como**
 - **Position-Specific Scoring Matrix (PSSM)**
- **Derivación de motifs (patterns)**
- **Generación de un MSA**
 - **partiendo de un MSA que se supone representativo de una familia o grupo de proteínas, se genera un profile**
 - **el profile se usa para generar alineamientos nuevos con proteínas no representadas originalmente en el profile**
 - **Más sensible que una matriz de scoring sitio-inespecífica**
- **Búsqueda de secuencias similares en bases de datos**
 - **El 'query' no es una secuencia, sino el profile**

Position-Specific-Iterated BLAST

<ftp://ftp.ncbi.nih.gov/blast/documents/blastpgp.html>

- 1. La 1ra iteración es un BLAST tradicional**
- 2. A partir de los hits se calcula un MSA y a partir del MSA se deriva un profile (PSSM)**
- 3. A partir de la segunda iteración, se usa la PSSM como query**

Profile HMMs

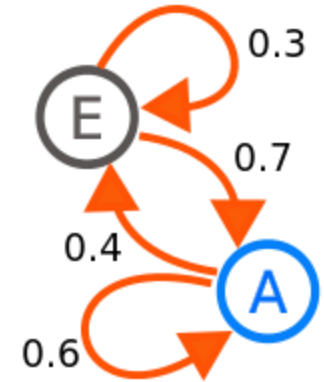
- La información contenida en un profile puede representarse de otras formas
- Los profiles originales contienen scores y penalidades basados en las frecuencias de ocurrencia
- Un profile (o un MSA) también puede representarse como una cadena de eventos con probabilidades de ocurrencia (Markov Chain)
- **Veamos un ejemplo!**

Markov Chains: una pequeña intro

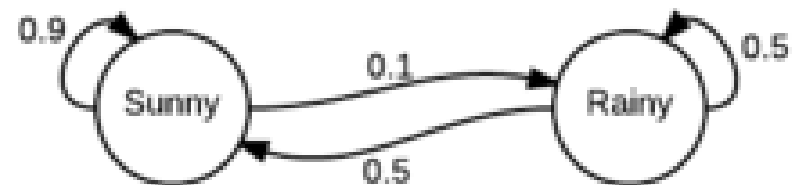
Una cadena de Markov es un sistema matemático que *transita* entre distintos *estados*, de acuerdo a probabilidades

Es un proceso azaroso y sin memoria

El próximo estado del sistema sólo depende del estado actual y no de la secuencia de estados precedentes (historia)

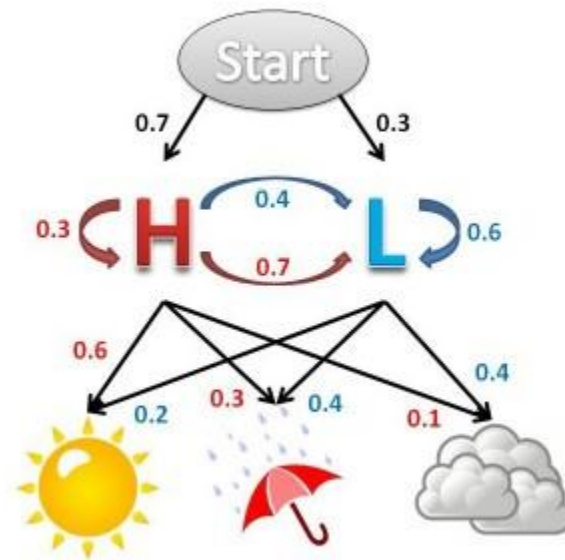
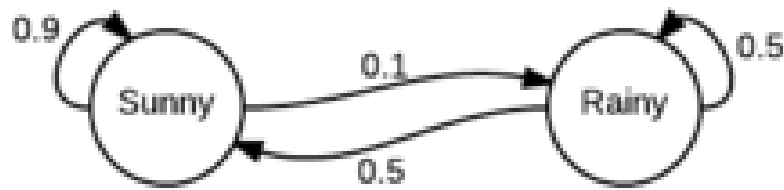


Markov Chain, Wikipedia. http://en.wikipedia.org/wiki/Markov_chain



Hidden Markov Models

Un modelo de Markov es un modelo probabilístico de algún Sistema, en donde existen estados no observables (ocultos).



Profile HMMs

El modelo se
inicia con
transiciones
equiprobables

Y se **entrena** con
un alineamiento

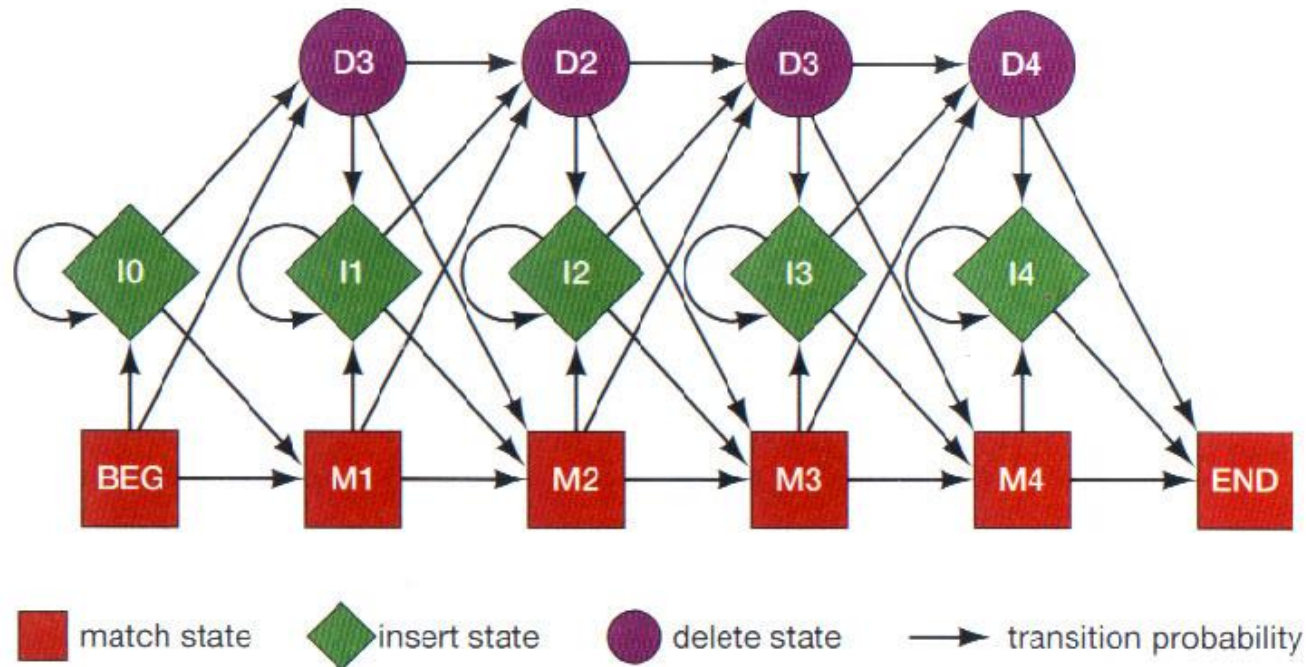
El profile ahora
está codificado en
forma de **estados**
y **probabilidades**
de transición

A. Sequence alignment

N	•	F	L	S
N	•	F	L	S
N	K	Y	L	T
Q	•	W	-	T

RED POSITION REPRESENTS ALIGNMENT IN COLUMN
GREEN POSITION REPRESENTS INSERT IN COLUMN
PURPLE POSITION REPRESENTS DELETE IN COLUMN

B. Hidden Markov model for sequence alignment





- **HMMER**
 - **<http://hmmer.wustl.edu>**
- **Paquete de programas para trabajar con profile HMMs**
 - **genera profile HMMs a partir de MSAs**
 - **usa los HMMs para realizar búsquedas en bases de datos de secuencias**
 - **puede buscar en bases de datos de profile HMMs a partir de una secuencia**



- **Una base de datos de profile HMMs**
- **(y de MSAs)**
 - **Wellcome Trust Sanger Institute**
 - **Stockholm Bioinformatics Centre**
 - **Janelia Farm**
- **Representan dominios proteicos**
- **Pueden buscar**
 - **a partir de palabras clave**
 - **a partir de una secuencia**
- **Pfam 27.0 (Marzo 2013, 14831 families)**



Sequence information

Alignment

☒ Seed (12) ☐ Full (28)

Format:

Hyperlinked plain text

[Retrieve alignment](#)

Visualize domain structures

☒ Seed (12) ☐ Full (28)

display per page.

[Retrieve domain structures](#)

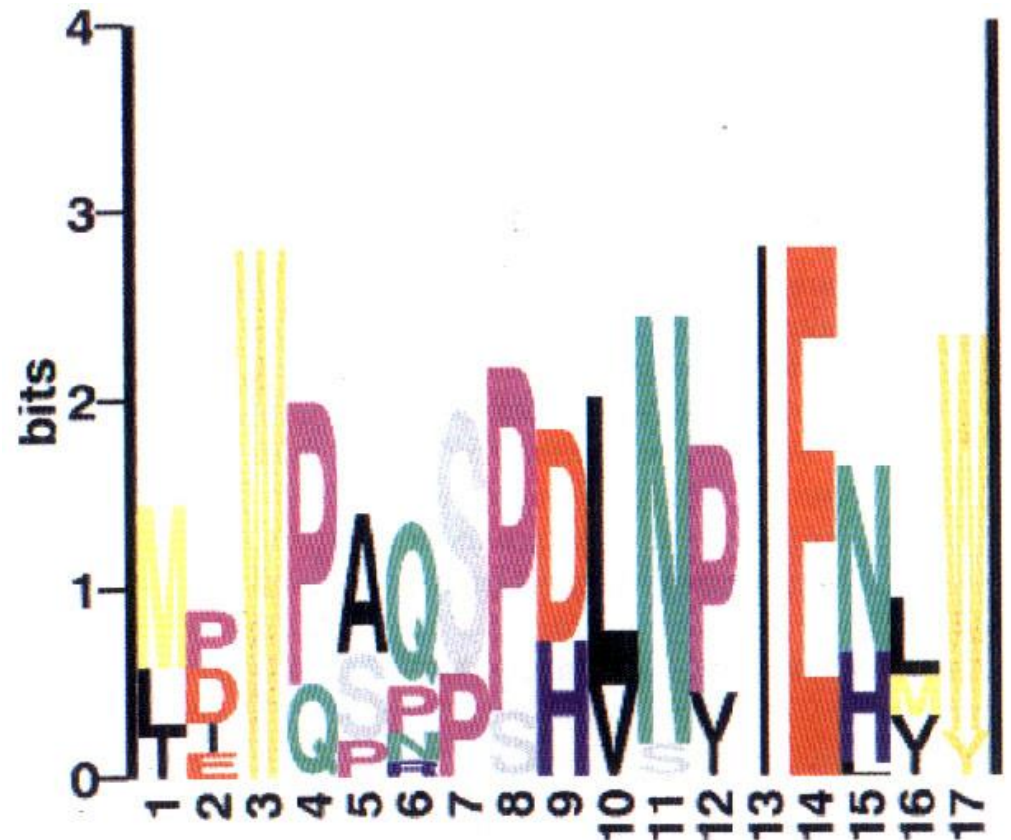
Species distribution

Tree depth:

[View species tree](#)

- Los motifs se pueden representar de distintas maneras (patterns por ejemplo)
- Sin embargo, los patterns no les dan peso a las distintas sustituciones
- [AC]-x-V-x(4)-{ED}
- Una **Position Specific Scoring Matrix** es una descripción de un motif en términos de una matriz

- Evaluar la información que contiene una PSSM usando Sequence Logos
- <http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html>



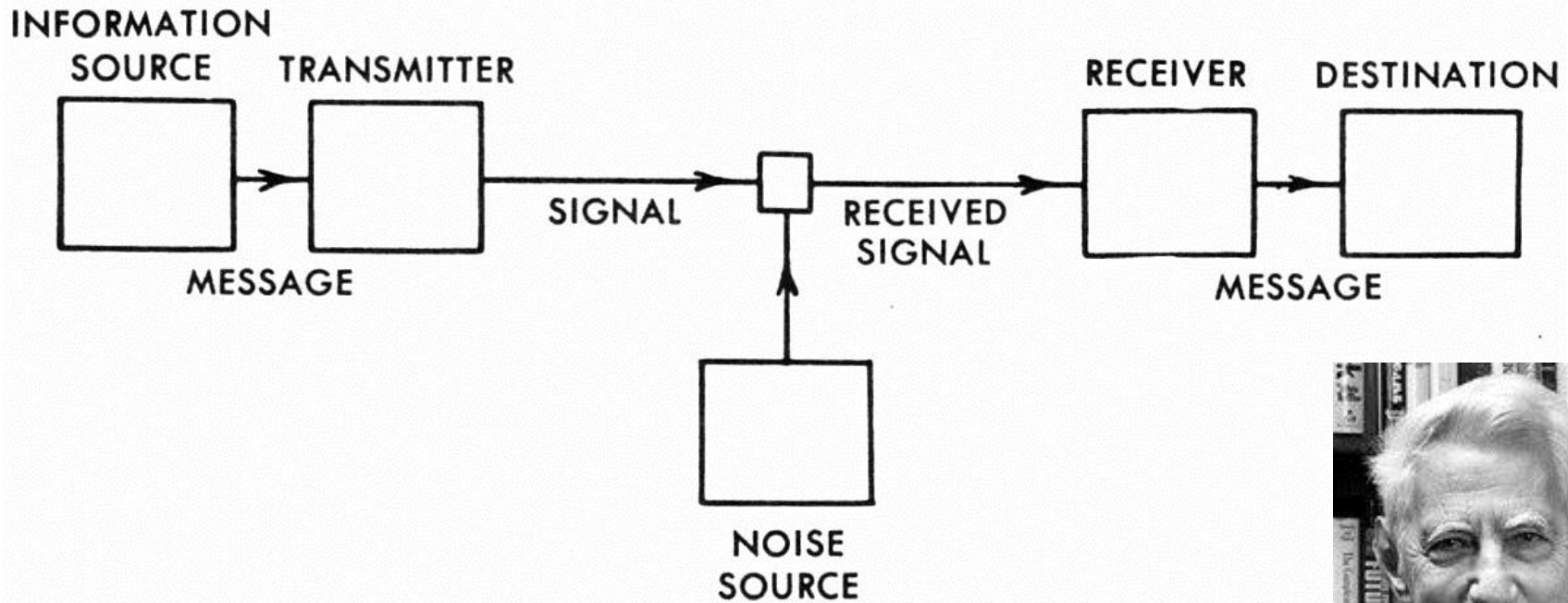
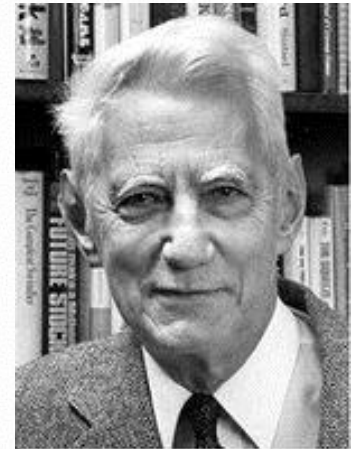


Fig. 1. — Schematic diagram of a general communication system.

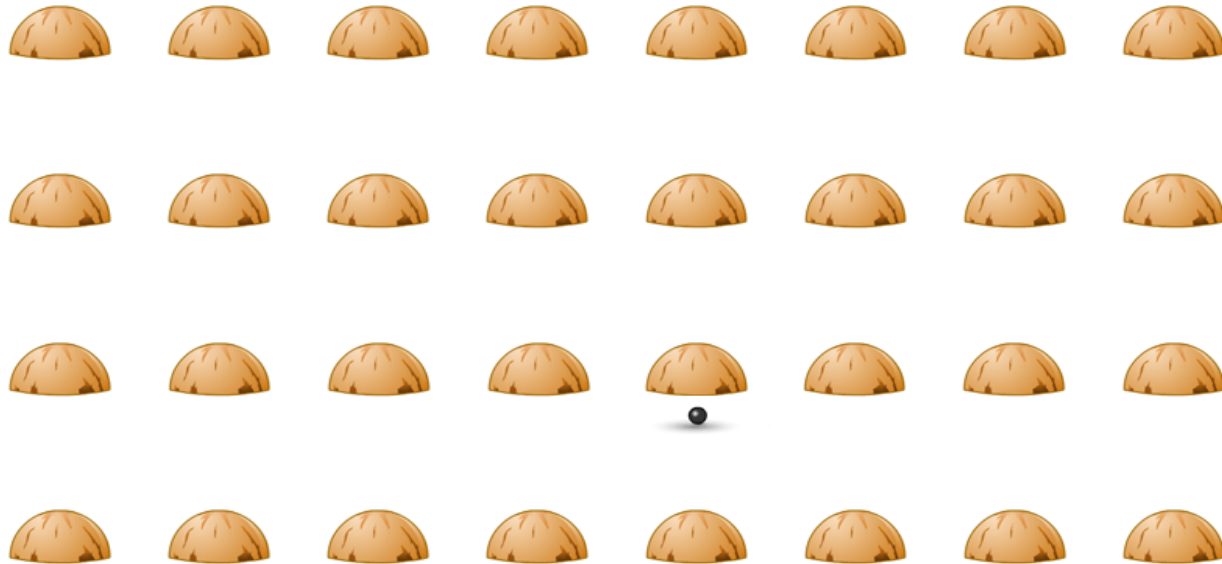


Claude Shannon

Information theory

- **Entropía:** medida de desorden de un sistema
- **La termodinámica** provee herramientas para calcular entropía
- **El desorden implica falta de *información* sobre el estado exacto de un sistema**
- **Claude Shannon / Leon Brillouin**
 - **Information theory**
 - **La Información es una combinación de**
 - Certain + Uncertain, Expected + Unexpected
 - **El grado de *sorpresa* que genera un evento que ya ocurrió es *cero***
 - **Si se reporta un evento poco probable, la información que se provee es *mayor***
 - **La información se incrementa cuando la probabilidad baja**

Shell game



Shell Game (Thimblorig)

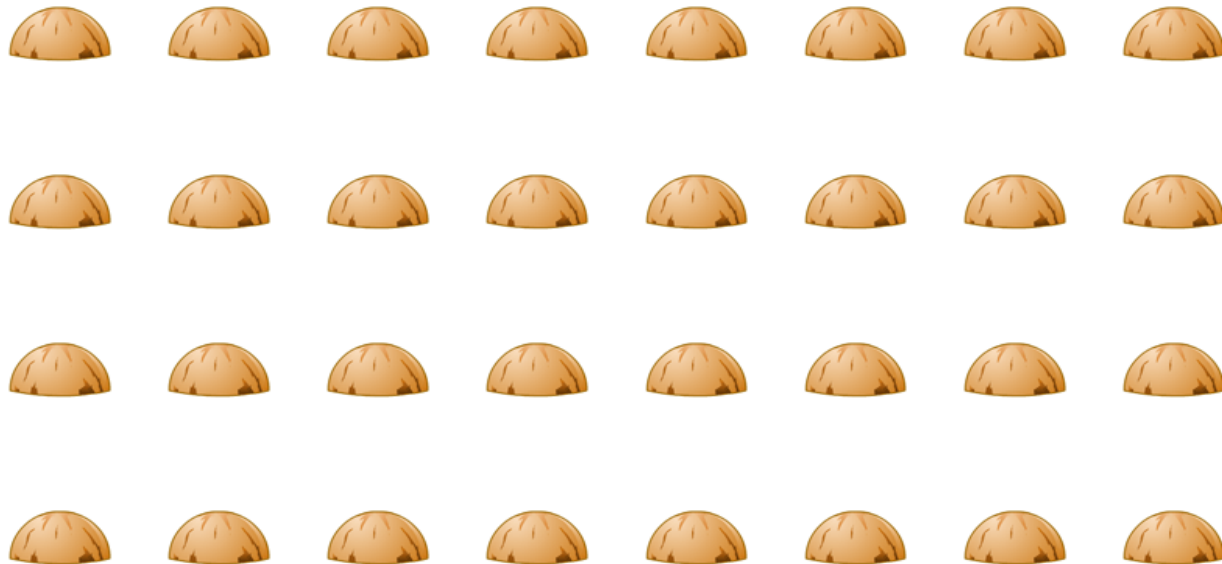
Adivinar en qué taza / nuez
está escondida la bolita.

Uncertainty

Si hay 64 nueces, cuántas
preguntas hay que hacer
para llegar a la respuesta?

Probability

$$p(object) = 1/64$$



Shell game



Shell Game (Thimblergig)

Adivinar en qué taza /
nuez está escondida
la bolita.

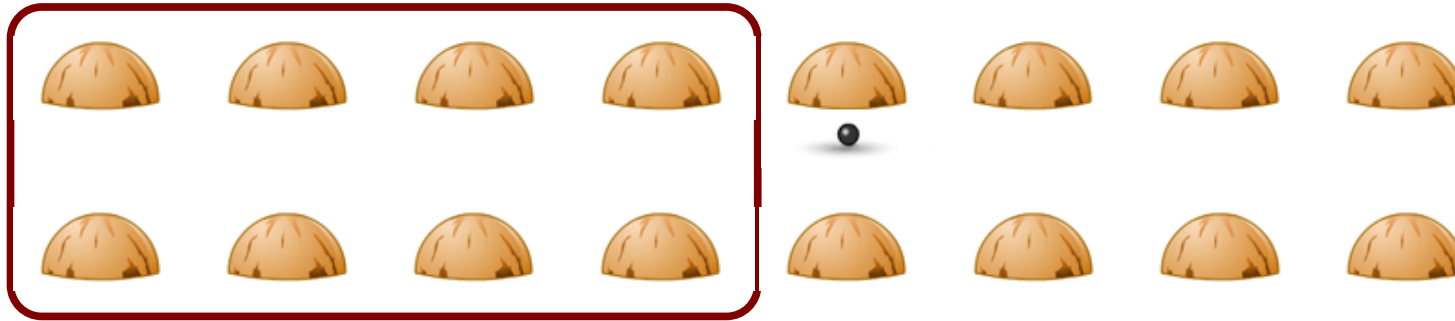
Uncertainty

Si hay 64 nueces,
cuántas preguntas
hay que hacer para
llegar a la respuesta?

Probability

$$p(\text{object}) = 1/64$$

Shell game



Shell Game (Thimblig)

Adivinar en qué taza /
nuez está escondida
la bolita.

Uncertainty

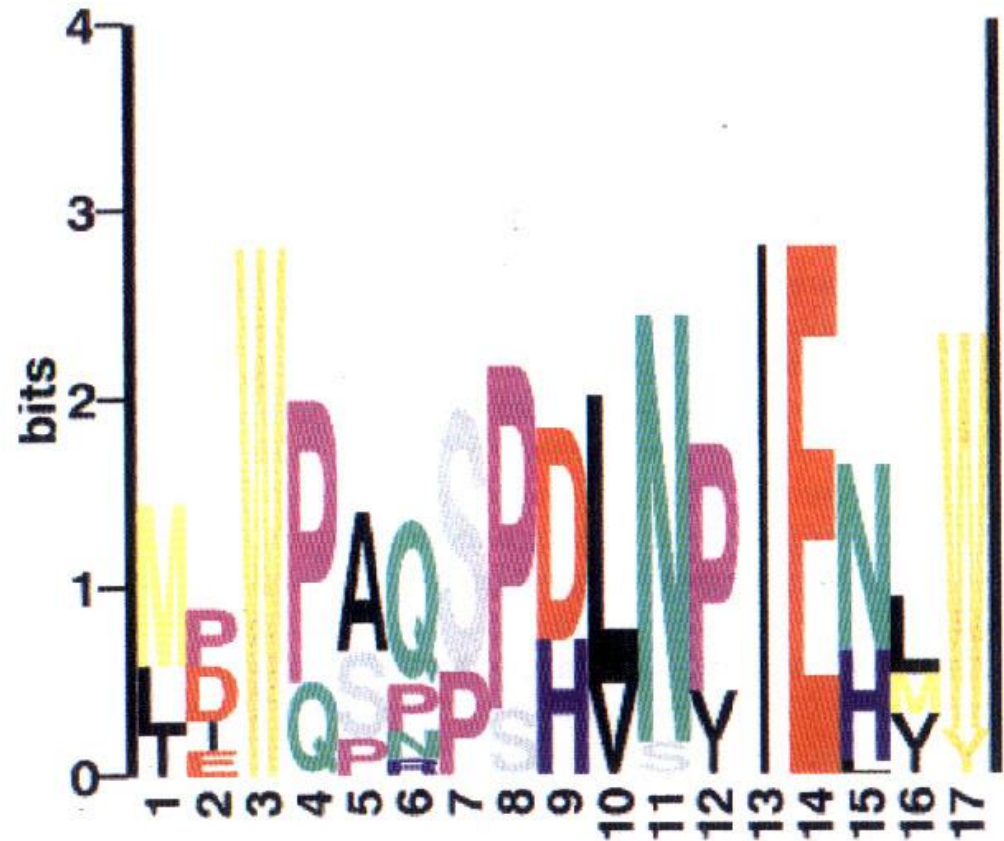
Si hay 64 nueces,
cuántas preguntas
hay que hacer para
llegar a la respuesta?

Probability

$p(object) = 1/64$

- Las preguntas secuenciales reducen las posibilidades (incertidumbre) de 64 a 32, luego a 16, 8, 4, 2, y finalmente 1.
- 6 preguntas son suficientes (peor caso) para encontrar la bolita.
- Esta es una manera de cuantificar la incertidumbre
- La incertidumbre también se puede calcular a partir de las probabilidades
 - Uncertainty = $-\log_2(1/64) = 6$

- **Information content of a PSSM**
 - Objetivo: conocer qué residuo pertenece a cada columna en el motivo
 - 20 residuos (20 posibilidades),
 $\log_2(20) = 4.32$
- **Sequence Logos**
 - Forma de visualización desarrollada por Tom Schneider
 - Grafica la cantidad de información (*disminución* en la incertidumbre) que nos da la matriz para cada posición





- **Protein Fingerprints DB**

- <http://www.bioinf.man.ac.uk/dbbrowser/PRINTS>

- **Qué es un fingerprint?**

- Una serie de motifs conservados en un orden particular
 - Se utilizan para predecir la ocurrencia de motifs similares en una secuencia
 - Importa la presencia y el orden de los motifs
 - Una proteína de la misma familia tiene todos los motifs en orden.
 - En el caso de una superfamilia, miembros de distintas familias pueden tener matchs parciales contra el fingerprint

SUMMARY INFORMATION

9 codes involving 8 elements
 0 codes involving 7 elements
 10 codes involving 6 elements
 29 codes involving 5 elements
 5 codes involving 4 elements
 4 codes involving 3 elements
 10 codes involving 2 elements

COMPOSITE FINGERPRINT INDEX

8	9	9	9	9	9	9	9	9
7	0	0	0	0	0	0	0	0
6	0	10	10	10	10	10	0	10
5	0	29	7	28	29	29	0	23
4	0	4	1	5	5	5	0	0
3	0	0	1	4	3	3	0	1
2	0	9	1	1	0	1	0	8

1	1	2	3	4	5	6	7	8

True positives..

ANX1_HUMAN	ANX1_BOVIN	ANX1_CAVCU	ANX1_RAT
ANX1_RABIT	ANX1_MOUSE	ANX1_COLLI	ANX1_COLLI
ANX1_RODSP			

Subfamily: Codes involving 6 elements

Subfamily True positives..

093446	ANX2_HUMAN	ANX2_CHICK	ANX2_RAT
ANX2_BOVIN	ANX2_MOUSE	ANX2_XENLA	ANX2_XENLA
093444	ANX5_BOVIN		

Subfamily: Codes involving 5 elements

Subfamily True positives..

093447	ANX3_RAT	ANX5_CHICK	ANX6_MOUSE
035639	ANX4_MOUSE	ANX4_HUMAN	ANX4_RAT
ANXA_BOVIN	ANXB_BOVIN	ANX4_PIG	ANX4_BOVIN
ANXA_RABIT	ANX6_HUMAN	ANX4_CANFA	ANXA_HUMAN
ANX6_RAT	ANX5_RAT	ANX3_HUMAN	ANX5_MOUSE
ANXA_MOUSE	ANX5_HUMAN	ANXD_HUMAN	093445
ANX7_HUMAN	ANX7_MOUSE	ANX6_CHICK	ANXX_DROME
ANXD_CANFA			

Subfamily: Codes involving 4 elements

Subfamily True positives..

ANX8_HUMAN	035640	ANXC_HYDAT	ANX5_CYNPY
Q27512			

Subfamily: Codes involving 3 elements

Subfamily True positives..

ANX7_XENLA	Q27473	ANX7_DICDI	059907
----------------------------	------------------------	----------------------------	------------------------

Subfamily: Codes involving 2 elements

Subfamily True positives..

Q27864	081536	081535	076027
Q43863	024131	Q42657	024132
082090	065848		

[Q27864](#)[081536](#)[081535](#)[076027](#)[Q43863](#)[024131](#)[Q42657](#)[024132](#)[082090](#)[065848](#)

NEX1 ANNEXIN - CAENORHABDITIS ELEGANS.

ANNEXIN P34 - LYCOPERSICON ESCULENTUM (TOMATO).

ANNEXIN P35 - LYCOPERSICON ESCULENTUM (TOMATO).

ANNEXIN 31 (ANNEXIN XXXI) - HOMO SAPIENS (HUMAN).

ANNEXIN P33 - ZEA MAYS (MAIZE).

ANNEXIN - NICOTIANA TABACUM (COMMON TOBACCO).

ANNEXIN - CAPSICUM ANNUUM (BELL PEPPER).

ANNEXIN - NICOTIANA TABACUM (COMMON TOBACCO).

FIBER ANNEXIN - GOSSYPIMUM HIRSUTUM (UPLAND COTTON).

ANNEXIN - MEDICAGO TRUNCATULA (BARREL MEDIC).

SCAN HISTORY

OWL21_1	2	100	NSINGLE
OWL26_0	1	100	NSINGLE
SPTR37_9f	2	122	NSINGLE

INITIAL MOTIF SETS

ANNEXINI1 Length of motif = 16 Motif number = 1
 Annexin type I motif I - 1

	PCODE	ST	INT
FLKQAWFIENEEQEYV	ANX1_HUMAN	6	6
FLKQARFLENQEYV	ANX1_MOUSE	6	6
FLKQAYFIDNQEYV	ANX1_CAVCU	7	7
FLKQAWFMENLEQECI	ANX1_COLLI	7	7
FLKQACYIEKQEYV	ANX1_RAT	6	6

ANNEXINI2 Length of motif = 23 Motif number = 2
 Annexin type I motif II - 1

	PCODE	ST	INT
MVKGVDDEATIIDILTKRNNAAQRQ	ANX1_HUMAN	55	33
MVKGVDDEATIIDILTKRTNAQRQ	ANX1_MOUSE	55	33
TVKGVDDEATIIDILTKRNNAAQRQ	ANX1_CAVCU	56	33
TAKGVDDEATIIDIMTTRTNAQRQ	ANX1_COLLI	51	28
MVKGVDDEATIIDILTKRTNAQRQ	ANX1_RAT	55	33

ANNEXINI3 Length of motif = 17 Motif number = 3
 Annexin type I motif III - 1

	PCODE	ST	INT
LKKALTGHLEEVVLALL	ANX1_HUMAN	95	17
LRKALTGHLEEVVLALL	ANX1_MOUSE	95	17
LKKALTGHLEEVVLALL	ANX1_CAVCU	96	17
MKRVLKSHLEDVVVALL	ANX1_COLLI	91	17
LKKALTGHLEEVVLALL	ANX1_RAT	95	17

ANNEXINI4 Length of motif = 22 Motif number = 4
 Annexin type I motif IV - 1

	PCODE	ST	INT
LRAAMKGLGTDEDTLIEILASR	ANX1_HUMAN	122	10
LRGAMKGLGTDEDTLIEILTTR	ANX1_MOUSE	122	10
LRAAMKGLGTDEDTLIEILVSR	ANX1_CAVCU	123	10
LRACMKGHGTDEDTLIEILASR	ANX1_COLLI	118	10
LRAAMKGLGTDEDTLIEILTTR	ANX1_RAT	122	10

- **Integra varias otras bases de datos en un solo lugar y provee referencias a otras bases de datos (GO)**
 - **<http://www.ebi.ac.uk/interpro>**
 - **Prosites, PRINTS, Pfam, ProDom, SMART**

InterPro

InterPro Simple Search

You can use this page to search for InterPro, Pfam, PRINTS, Prosite, SWISS-PROT, TrEMBL accession numbers and names, database names, and entry_types. You may combine more than one search term with 'AND', '&', 'OR', '|', 'NOT' and '!'; you may also use wildcarded expressions (eg. *bar**).

Enter search terms here...

Search results for 'human transporter'

Click on the links below to jump to individual InterPro entries.

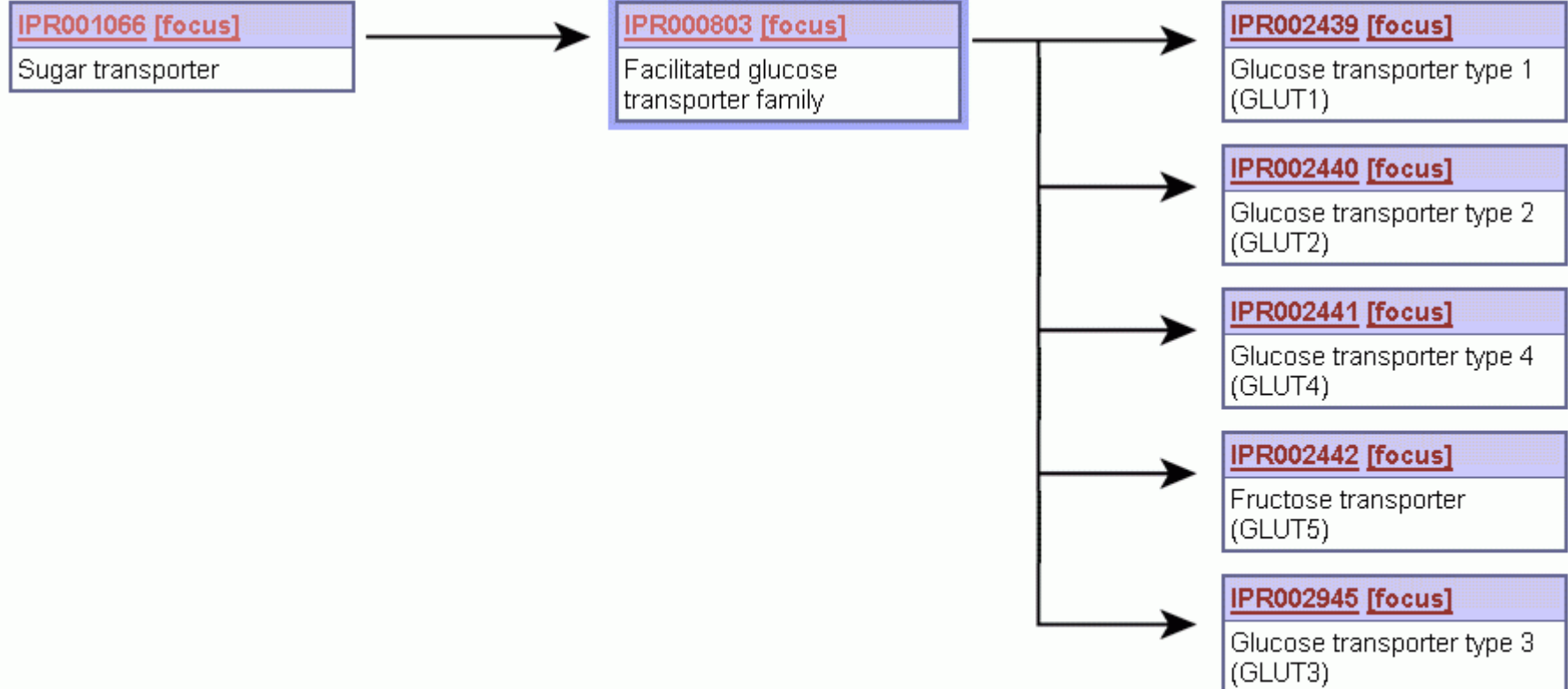
Entry	Entry name
IPR000076	K-Cl co-transporter
IPR000622	K-Cl Co-transporter type 1 (KCC1)
IPR000803	Facilitated glucose transporter family
IPR000849	GipT family of transporters
IPR001066	Sugar transporter
IPR001204	Phosphate transporter family
IPR001902	Sulfate transporter
IPR002259	Delayed-early response protein/equilibrative nucleoside transporter
IPR002293	Permease for amino acids and related compounds, family I
IPR002435	Noradrenaline neurotransmitter transporter
IPR002436	Dopamine neurotransmitter transporter
IPR002437	Serotonin (5-HT) neurotransmitter transporter

InterPro

Tree display for IPR000803

The tree below shows the selected InterPro entry, the path to the root of the tree, the immediate children and the immediate children of the selected entry's parent (i.e. the entry's siblings).

To return to the full entry for this accession number, click [here](#).



InterPro

InterPro - Proteins matching IPR000803

Table Graphical



Grid shows 10aa intervals, first mark at position 0. Move the mouse over a match to see more information in the status line of your browser window.

Item 21-40 of 91

< 1 2 3 4 5 >

Help for : graphic key - Netscape

Graphical match display legend

The table below shows the colour coding used in the graphical match display. The extent of the bars denotes the region on the protein sequence that the selected method [matches](#).

	True	Unknown
PRINTS		
PROSITE pattern		
PROSITE profile		
PFAM		
ProDom		n/a

See also :

Protein	Match Display
SWISS-PROT GTR2_HUMAN P11168	IPR000803 PR00172 GLUCTRNSPORT
	IPR001066 PS00216 SUGAR_TRANSPORT
	IPR001066 PS00217 SUGAR_TRANSPORT
	IPR001066 PR00171 SUGRTRNSPORT
	IPR001066 PF00083 sugar_tr
	IPR002440 PR01191 GLUCTRSPORT2
SWISS-PROT GTR3_HUMAN P11169	IPR000803 PR00172 GLUCTRNSPORT
	IPR001066 PS00216 SUGAR_TRANSPORT_1
	IPR001066 PS00217 SUGAR_TRANSPORT_2
	IPR001066 PR00171 SUGRTRNSPORT
	IPR001066 PF00083 sugar_tr

InterPro

Help for : table legend - Netscape

Tabular match display legend

The single letter codes after the amino acid ranges in this table denote the status of each individual match. Possible values are shown in the table below :

- T True
- F False Positive
- N False Negative
- P Partial
- ? Unknown

InterPro - Proteins matching IPR001066

Table [Graphical](#)



Item 401-420 of 1177

< [Previous](#) [21](#) [22](#) [23](#) [24](#) [25](#) [Next](#) >

	PS00216	PS00217	PR00171	PF00083
P39637 YWFA_BACSU				19-406 T
P39843 BMR2_BACSU	65-81 T			17-398 T
P39850 CAPA_STAAU		175-200 F		
P39924 HXTC_YEAST	370-387 T	169-194 T	68-78 T 164-183 T 328-338 T 423-444 T 446-458 T	60-521 T
P39932 STL1_YEAST	347-364 T	N		30-488 T
P40441 YIRO_YEAST	263-280 T	62-87 T		2-416 T
P40474 YIM1_YEAST	117-133 F			61-539 T
P40475 YIM0_YEAST	125-141 F			71-547 T
P40862 PROP_SALTY	P	P		
P40885 HXT9_YEAST	373-390 T	172-197 T	72-82 T 167-186 T 331-341 T	64-526 T

MSA: frecuencias de sustitución de aas

- **Un MSA es la base para determinar las frecuencias de sustitución de amino ácidos en un grupo particular de secuencias**
 - **frecuencias de sustitución globales**
 - Se utilizan para generar matrices de scoring:
 - Matrices PAM, BLOSUM, etc
 - Dan puntaje y penalizan por igual los mismos cambios, independientemente del contexto
 - **frecuencias de sustitución sitio por sitio**
 - Position Specific Scoring Matrices (PSSM)
 - Profiles

Cómo los uso?

- **Así como BLAST/FASTA pueden buscar sobre secuencias utilizando secuencias, distintos programas pueden buscar sobre secuencias usando**
 - **patterns**
 - **motifs**
 - **profiles**
 - **PSSMs**
 - **etc.**
- **Y en general también vale la inversa (buscar usando secuencias)**
- **Vamos a ver ejemplos en el TP de EMBOSS**

Bioinformatics. Sequence and Genome analysis. David W Mount, CSHL Press (2001)

Markov Chains, a visual explanation

<http://setosa.io/blog/2014/07/26/markov-chains/index.html>

Schneider Lab Home Page (Information Theory for Biology, Sequence Logos)

<http://schneider.ncifcrf.gov/>