

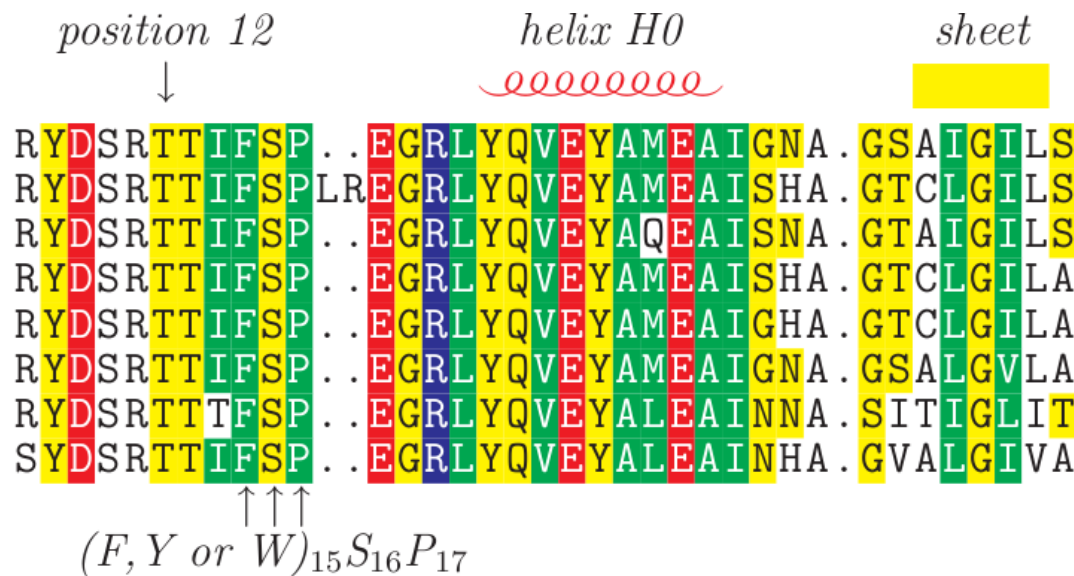
Alineamiento múltiple de secuencias

Algoritmos exactos, heurísticas,
información contenida en un alineamiento

Fernán Agüero

Instituto de Investigaciones Biotecnológicas

Universidad Nacional de San Martín



Alineamientos múltiples

Qué es un alineamiento múltiple?

Example: A multiple sequence alignment corresponding to the WW domain
(Source: SMART database)

O54971/1-33	PLPPGWEKRT	DSN-GRVYFV	N---HNTRIT	QWEDPRS
O43165/1-33	GLPSGWEERK	DAK-GRYYV	N---HNNRTT	TWTRPIM
NED4_HUMAN/1-33	PLPPGWEERT	HTD-GRIFYI	N---HNIKRT	QWEDPRL
O14326/1-33	PLPSGWEML	TNS-ARVYFV	D---HNTKTT	TWDDPRL
O43165_2/1-33	FLPPGWEML	APN-GRPFFI	D---HNTKTT	TWEDPRL
PIN1_HUMAN/1-34	KLPPGWEKRM	SRSSGRVYYF	N---HITNAS	QWERPSG
NED4_HUMAN_1/1-0	PLPPGWEERQ	DIL-GRYYV	N---HESRRT	QWKRPTP
O75853/1-33	PLPPGWEVRS	TVS-GRIFYFV	D---HNNRTT	QFTDPRL
PUB1_SCHPO_2/1-0	RLPPGWERRT	DNL-GRYYV	D---HNTRST	TWIRPNL
YA65_CHICK/1-33	PLPPGWEAK	TPS-QQRYFL	N---HIDQTT	TWQDPK
I83196_2/1-33	GLPPGWEKQ	DDR-GRSYYV	D---HNSKTT	TWSKPTM
YA65_MOUSE/1-33	PLPDGWEQAM	TQD-GEVYYI	N---HKNKTT	SWLDPRL

- Alineamiento de 3 o más secuencias (DNA o proteína)
- Se asume una relación evolutiva entre las secuencias

Alineamientos múltiples

Importancia: muchos métodos en bioinformática usan alineamientos múltiples como *input*.

La calidad de los alineamientos es clave!

Divergencia evolutiva: muchas especies tienen variantes de la misma proteína, todas con esencialmente la *misma* función molecular, pero con secuencias de amino ácidos *diferentes*.

Homología: genes (ADN) y proteínas que evolucionaron a partir de un mismo gen o proteína ancestral se dice que son *homólogos*.

Homólogos, ortólogos, parálogos

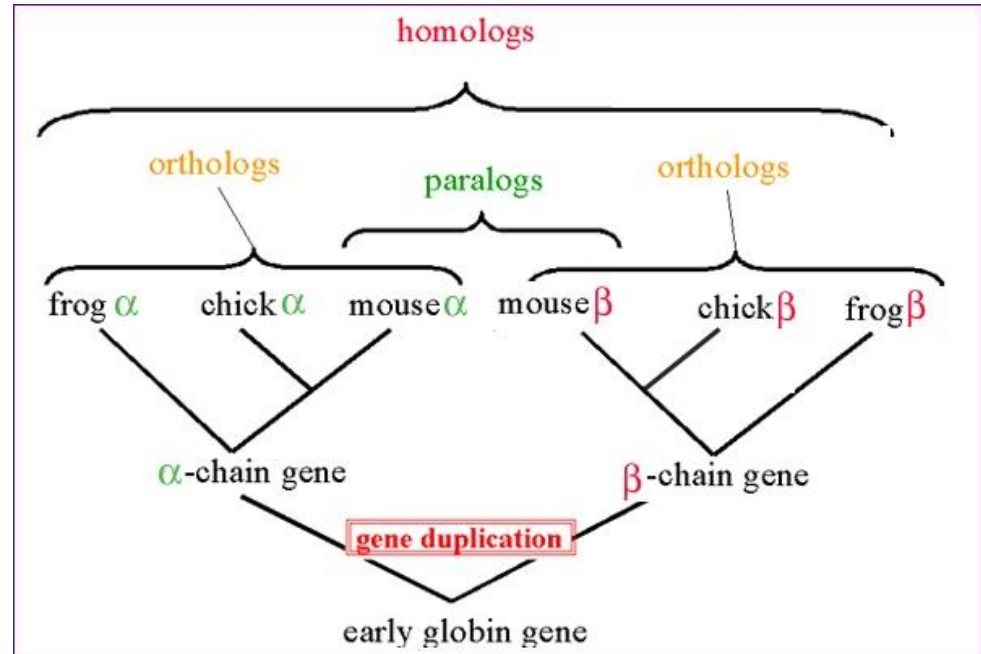
Un par de genes **homólogos** se dice que son **ortólogos** cuando se separaron por especiación.

los genes ortólogos tienden a conservar la misma función molecular en diferentes especies

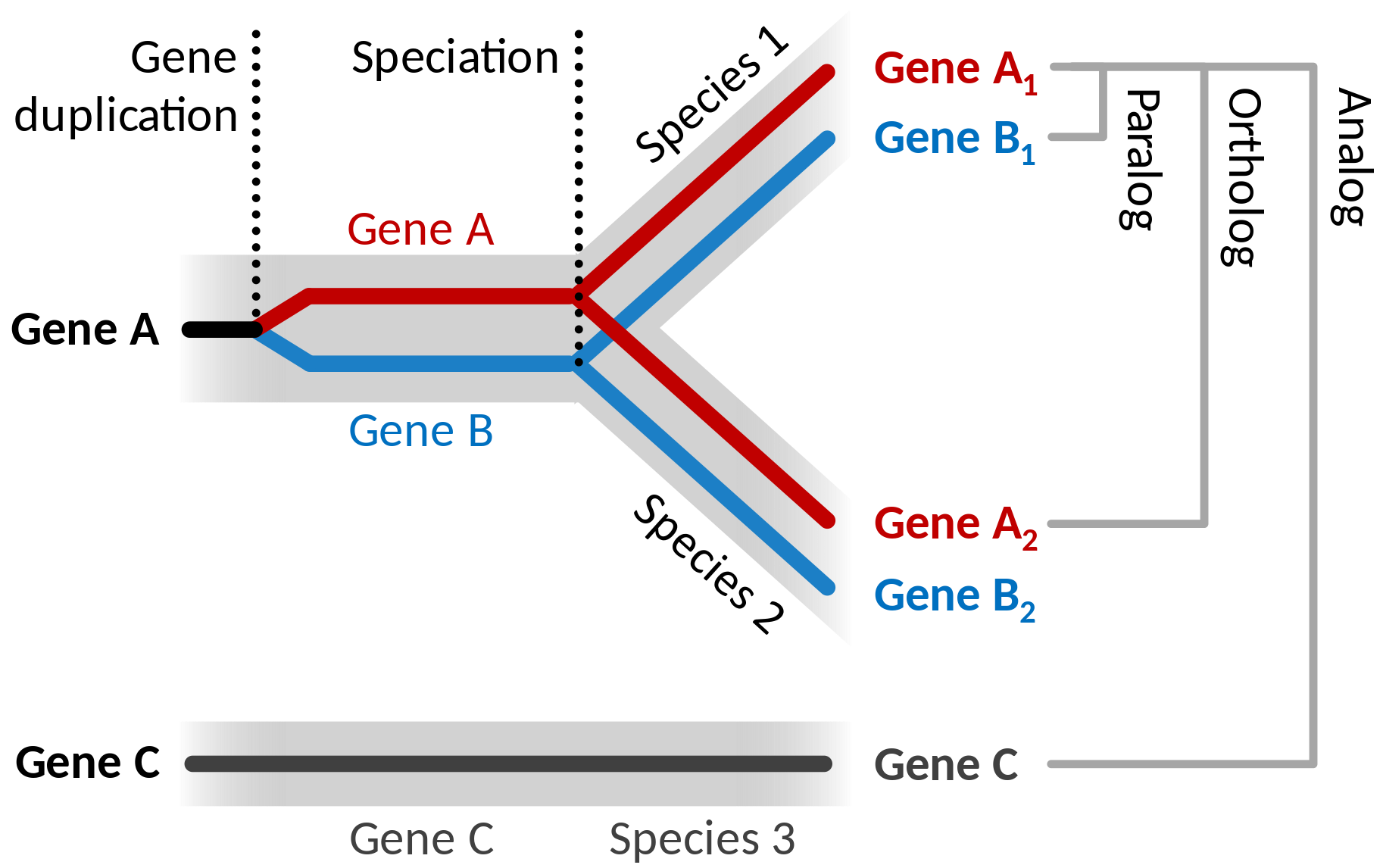
Un par de genes **homólogos** se dice que son **parálogos** cuando se generaron por duplicación dentro de la misma especie.

los genes parálogos tienden a desarrollar o evolucionar hacia funciones diferentes

PERO! Caveat emptor: cuidado con las generalizaciones!



Homólogos, ortólogos, parálogos



By Thomas Shafee - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=70715956>

In-paralogs, Out-paralogs

Box 1: Relationships between genes

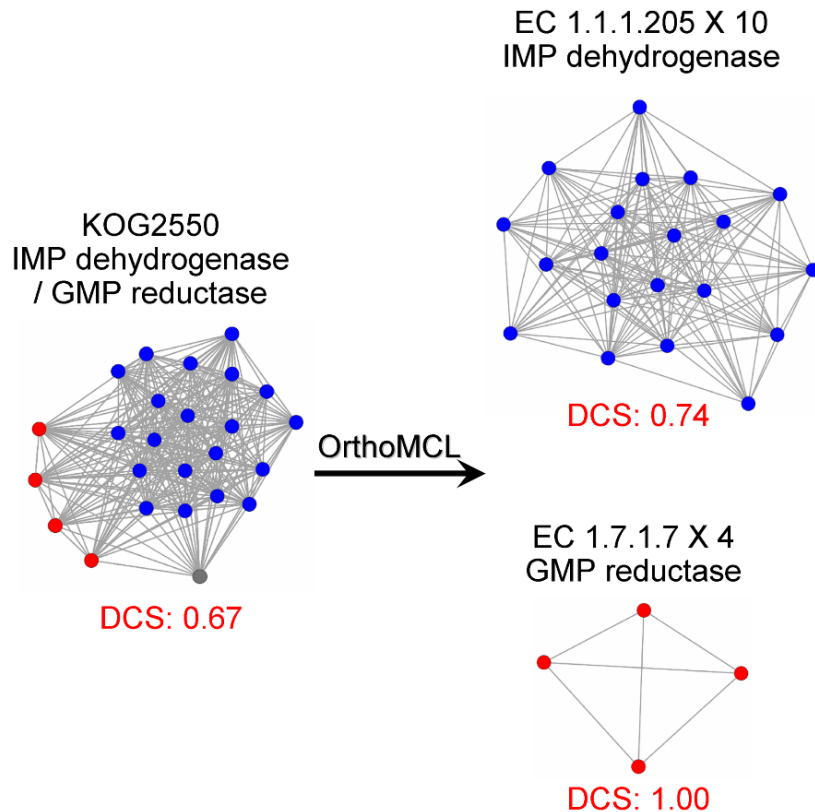
- Homology: genes that share a common origin.
- Analogy: non-homologous genes that perform the same function as a result of convergent evolution.
- Orthology: genes arising by speciation at their most recent point of origin.
- Paralogy: genes arising by duplication at their most recent point of origin.
- Xenology: genes arising by HGT from another organism.
- In-/Out-paralogy: paralogous genes arising from lineage-specific duplication(s) after/before a given speciation event.
- Co-orthology: in-paralogous genes that are collectively, but not individually, orthologous to genes in other lineages (due to their common origin by speciation).
- Orthologous group: collection of all descendants of an ancestral gene that diverged from (after) a given speciation event.

David M. Kristensen, Yuri I. Wolf, Arcady R. Mushegian, Eugene V. Koonin,
Computational methods for Gene Orthology inference,
Briefings in Bioinformatics 12: 379–391.

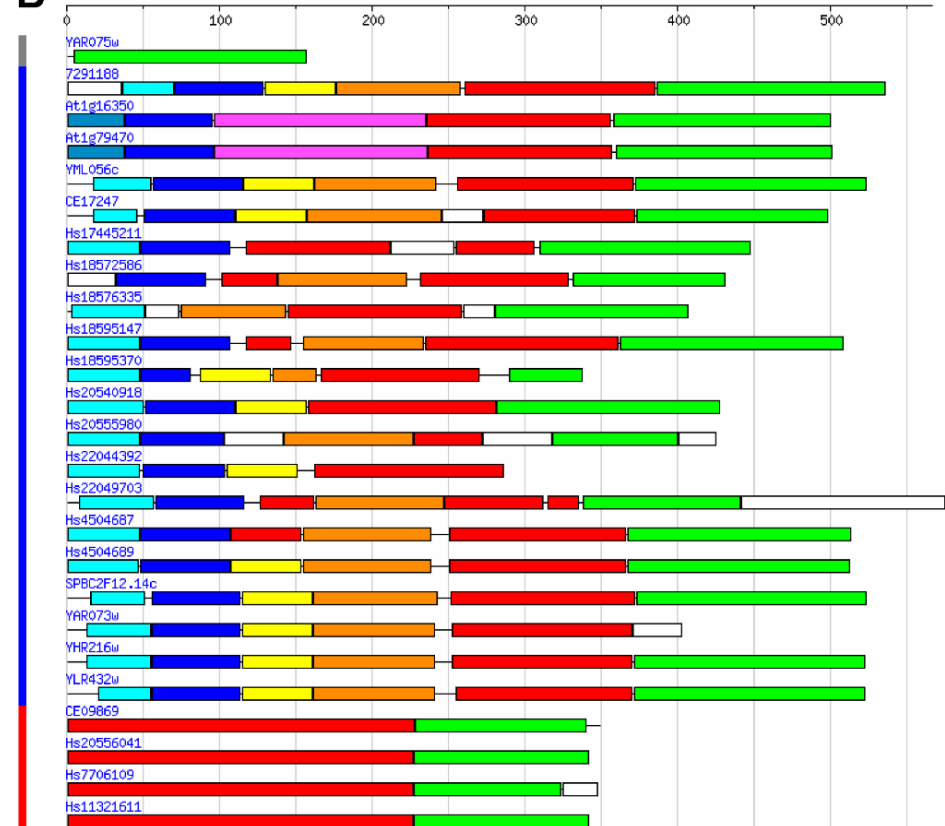
Agrupamiento de Ortólogos

Hay distintas aplicaciones que buscan *agrupar* ortólogos

A



B



Chen F et al 2007 PLOS One 2: e383
DOI: 10.1371/journal.pone.0000383

Y cómo sería un *algoritmo* de alineamiento multiple?

Como en muchos otros algoritmos necesitamos:

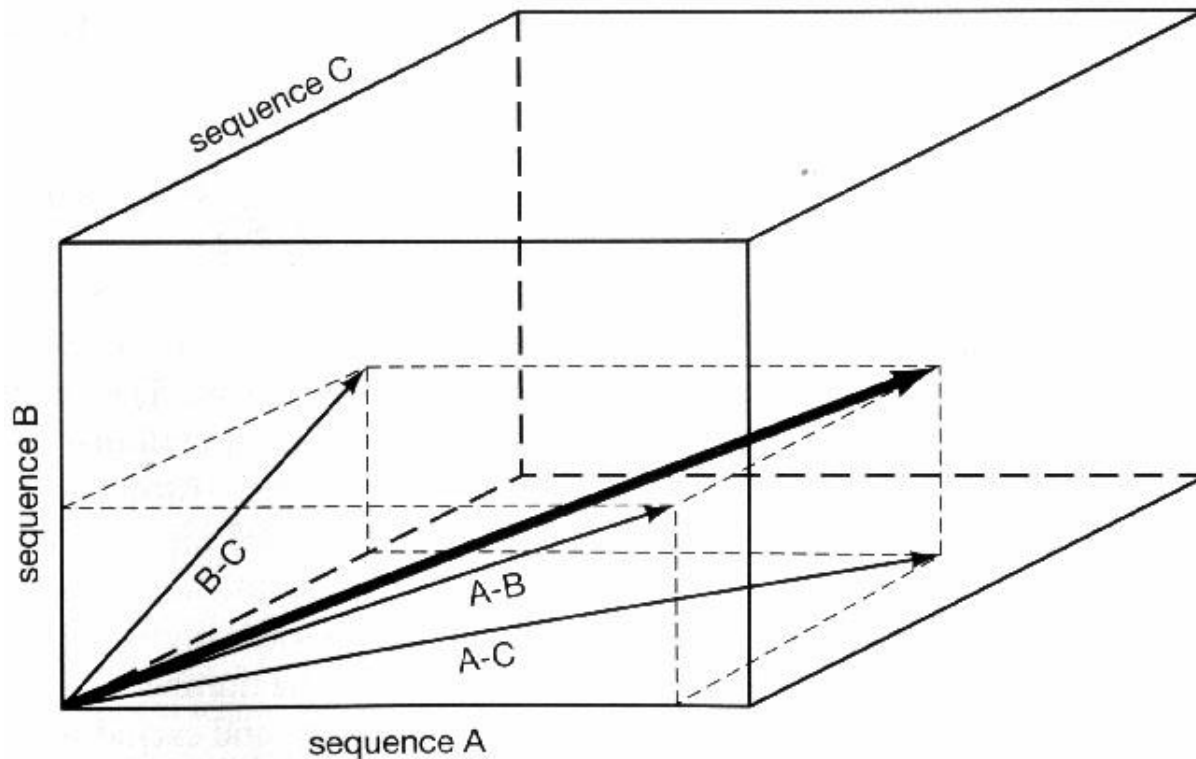
- Una **función objetivo** (métrica)
- Un **procedimiento** para **optimizar** la función objetivo

Para dos alinear **dos** secuencias:

- Una **función objetivo** (métrica)
 - *Sistema de puntajes (Scoring), Matrices (ej BLOSUM62)*
- Un **procedimiento** para **optimizar** la función objetivo
 - *Dynamic programming (ej Needleman-Wunsch)*

Algoritmo exacto

- **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 LYFSKFVPG-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 (Dynamic programming)

- **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

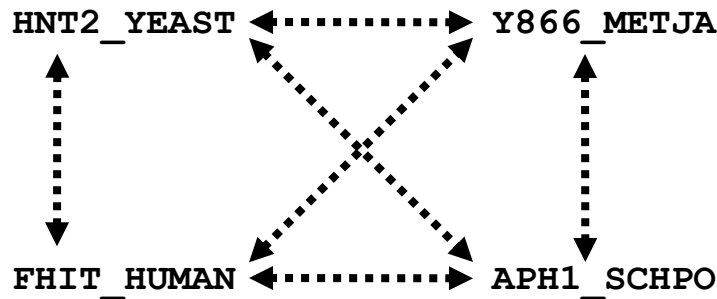
- Needleman-Wunsch o Smith Waterman extendido a una matriz ***n-dimensional***
- MSA (Lipman et al. 1989)
 - <http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/msa.html>
 - **Multidimensional dynamic programming**
 - **Usa heurísticas para reducir el espacio de búsqueda**
 - **Varios programas:**
 - msa_50_150 - 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 ...

Algoritmo

1. Alinear todas las secuencias de a pares
 2. Usar los scores para construir un árbol filogenético
 3. 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.

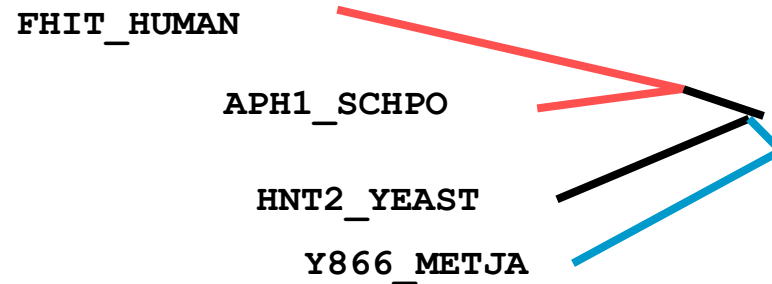


Pairwise alignments: compute distance matrix

	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...
```

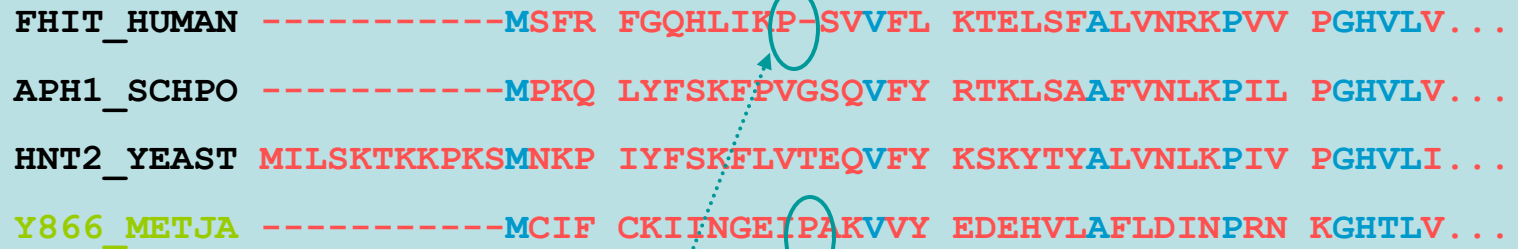


**Alinear las dos
secuencias más
cercanas**

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...
```

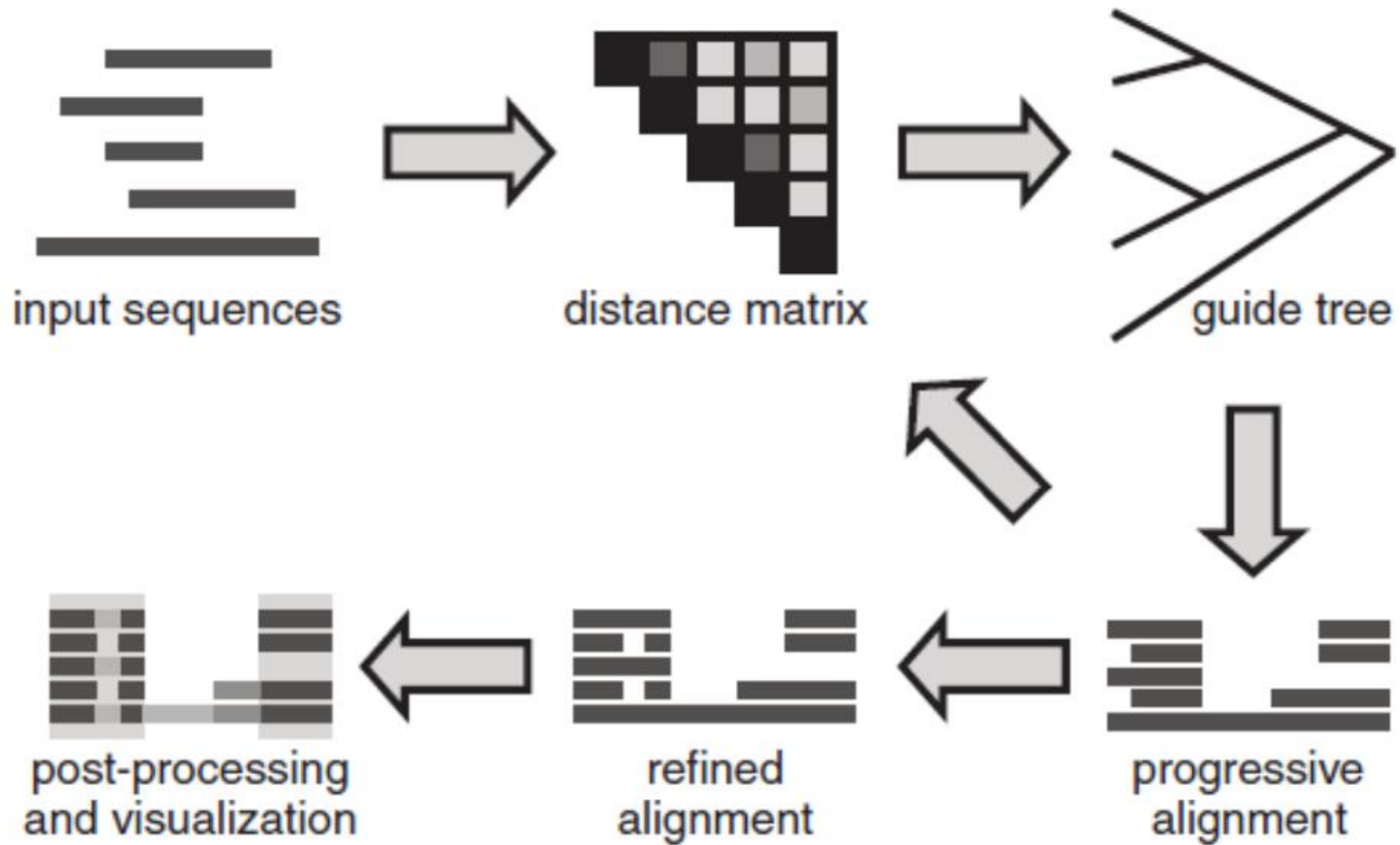


**Alinear la
secuencia
siguiente**

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).

Resumen de alineamientos progresivos



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.clustal.org/clustal2/>

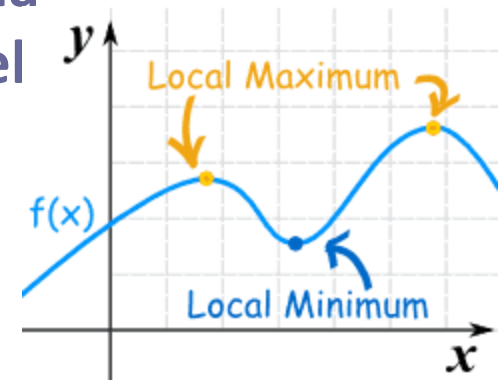
New version ClustalO (Omega)

<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)
 - MAFFT (Katoh, 2002)
- Como todos los métodos de optimización, pueden quedar atrapados en mínimos locales

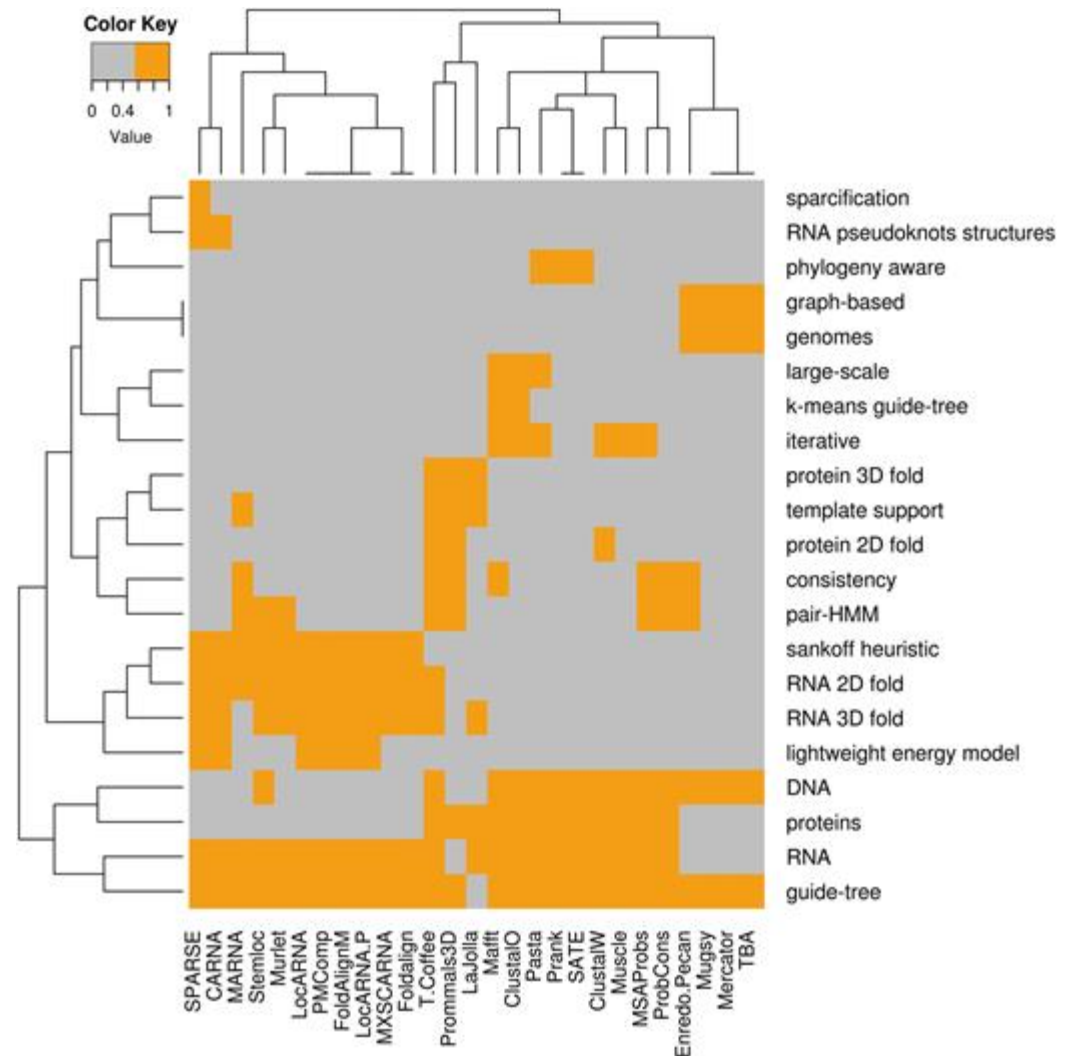


- **SAGA (Notredame & Higgins, 1996)**
 - **Sequence Alignment by Genetic Algorithm**
 - **Genera diferentes MSAs por rearrreglos que simulan inserciones de gaps similares a los que ocurren durante la replicación del DNA**
 - **El proceso continúa hasta que converge en un score que no puede ser mejorado**
 - **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**

Otros algoritmos más recientes

- **T-Coffee**
- **MUSCLE**
- **MAFFT**
- **ProbCons**

Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, Cedric Notredame, Multiple sequence alignment modeling: methods and applications, *Briefings in Bioinformatics*, Volume 17, Issue 6, November 2016, Pages 1009–1023, <https://doi.org/10.1093/bib/bbv099>



Visualización y Edición de Alineamientos

Herramienta	URL
Jalview	https://jalview.org
SeaView	http://doua.prabi.fr/software/seaview



protdna.mase

File ▾ Edit ▾ Props ▾ Sites ▾ Species ▾ Footers ▾ Search: Goto: Trees ▾ Help

sel=0 310 Seq:26 Asp Pos:337|178 [AK045539.PE1] 402

AB035322 GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCTGATGAGAGAAGCTGGGGTTC

AB036423.IBA1 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATAATGCTGGAGAAACTTGGGGTT

AB036423.PE2 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATAATGCTGGAGAAACTTGGGGTT

AB094629.IBA2 GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCTGATGAGAGAAGCTGGGGTTC

AB128049.AIF1 GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAACGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGGGTTC

AF074959.PE1 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATAATGCTGGAGAAACTTGGGGTT

AF109719.PE7 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATAATGCTGGAGAAACTTGGGGTT

AF129756.PE18 GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTTC

AF299325 -----ATGGAGTTTGATCTGAATGGCAATGGTGATATTGATATTATGTCCCTGAAACGAATGCTGGAGAAGCTTGGGGTTC

AF299327 -----ATGGAGTTTGATCTGAATGGCAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTTC

AF299328 -----ATGGAGTTTGATCTGAATGGCAATGGAGATATCGATATTATGTCCCTGAAGCGAATGCTGGAGAAACTTGGGGTT

AF348450.PE1 GCCTTCAAGAAGAAATACATGGAGTTTGACCTGAATGAAGATGGAGGTATCGATATCATGTCCCTGAAGCGAATGATGGAGAAACTTGGGGTT

AJ506968.AIF1 ATGTTTAAAAATAAATACATGGAGTTTGATCTCAATGATCAAGGAGACATAGACATAATGGGTTTAAACGGATGCTTGAAAACTTGGAGTTC

AK006184.PE1 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATATTATGTCCCTGAAGCGAATGCTGGAGAAACTTGGGGTT

AK006562.PE1 GCCTTCAAGGTGAAGTACATGGAGTTTGATCTGAATGGAAATGGAGATATCGATATTATGTCCCTGAAGCGAATGCTGGAGAAACTTGGGGTT

AK022845.PE1 GGCTTCAAAGAGAAATACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AK028955.PE1 GGCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGGGTTC

AK045539.PE1 GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGGGTTC

AK091912.PE1 GGCTTCAAAGAGAAATACATGGAGTTTGACCTGAATAATGAAGGCGAGATTGATCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AK128526.PE1 GGCTTCAAAGAGAAATACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL136566.PE1 GGCTTCAAAGAGAAATACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL157938.C9ORF58 GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL157938.PE2 GCCTTCAAAGAGAAATACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL157938.PE3 GGCTTCAAAGAGAAATACATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL157938.PE7 -----ATGGAGTTTGACCTGAACAATGAAGGCGAGATTGACCTGATGTCTTTAAAGAGGATGATGGAGAAGCTTGGTGTC

AL662801.PE45 GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTTC

AL662801.PE46 -----ATGGAGTTTGACCTTAATGGAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTTC

AL662847.AIF1 GGCTTCAAAGAGAAATACATGGAGTTTGACCTTAATGGAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTTC

AL662847.PE40 -----ATGGAGTTTGACCTTAATGGAATGGCGATATTGATATCATGTCCCTGAAACGAATGCTGGAGAAACTTGGAGTTC

)(<-+ _



- **BLOCKS**

- *Blocks are ungapped multiple sequence alignments representing conserved protein regions*
- <http://blocks.fhcrc.org/blocks> (no existe más)
- SeqFire
 - <http://www.seqfire.org/> (modulo "Conserved block")
- Gblocks
 - Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis
 - https://home.cc.umanitoba.ca/~psgendb/doc/Castresana/Gblocks_documentation.html

- **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 / Weight matrices**
 - **HMMs**
- **Para qué extraer / generar patterns, motifs, etc?**
 - **Para clasificar**
 - **Para alinear secuencias**
 - **Para buscar secuencias similares por métodos más sensibles**

Usos de los MSAs

- Para extraer / generar
 - **Patterns/Motifs** 
 - Profiles
 - Fingerprints
 - Position Specific Scoring Matrices / Weight matrices
 - HMMs
- Para qué extraer / generar patterns, motifs, 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?

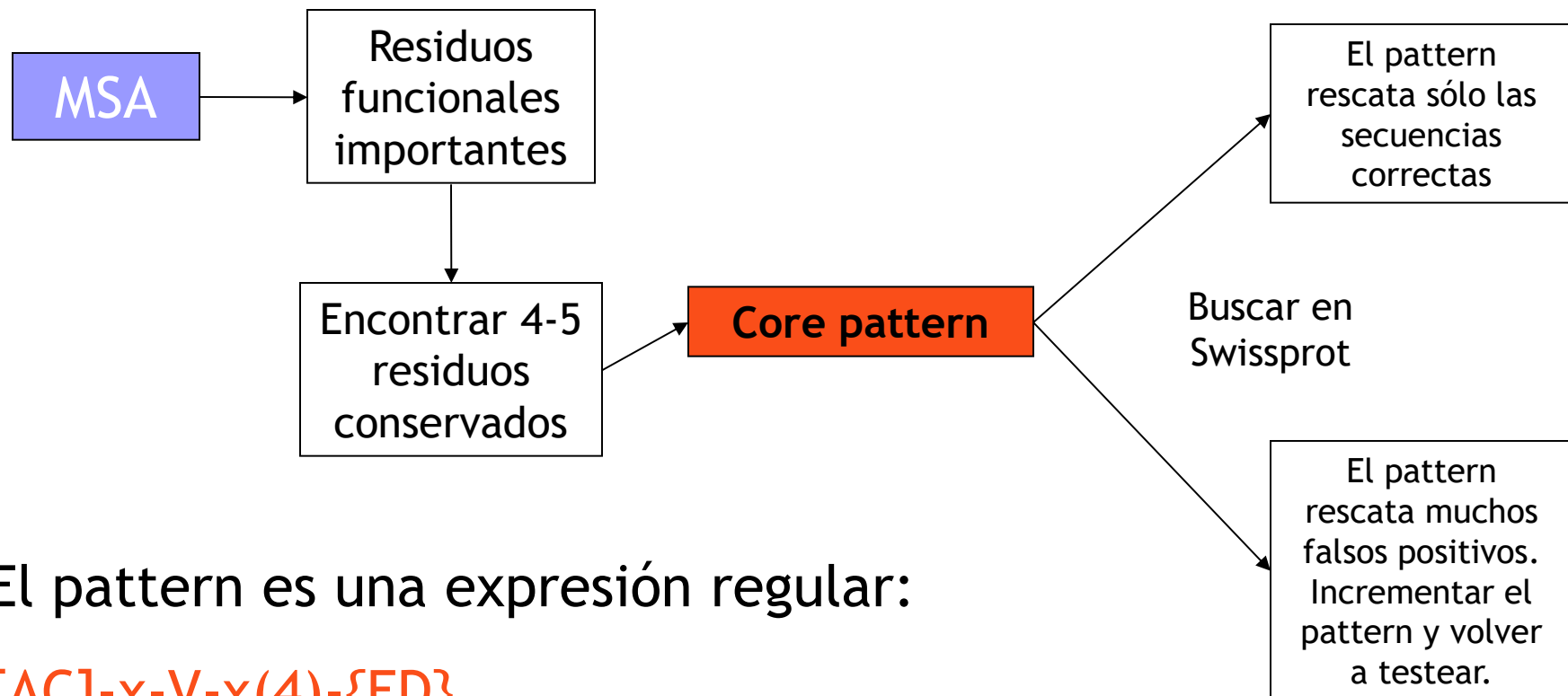
- **Patterns**

- **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>

General information about the entry

Entry name [info]	PYRUVATE_KINASE
Accession [info]	PS00110
Entry type [info]	PATTERN
Date [info]	01-APR-1990 CREATED; 01-JUL-1999 DATA UPDATE; 03-AUG-2022 INFO UPDATE.
PROSITE Doc. [info]	PDOC00101

Name and characterization of the entry

Description [info]	Pyruvate kinase active site signature.
Pattern [info]	[LIVAC]-x-[LIVM](2)-[SAPCV]-K-[LIV]-E-[NKRST]-x-[DEQHS]-[GSTA]-[LIVM].

Numerical results [\[info\]](#)

Numerical results for UniProtKB/Swiss-Prot release **2022_03** which contains **568'002** sequence entries.

Total number of hits	84 in 84 different sequences
Number of true positive hits	79 in 79 different sequences
Number of 'unknown' hits	0
Number of false positive hits	5 in 5 different sequences
Number of false negative sequences	22
Number of 'partial' sequences	4
Precision (true positives / (true positives + false positives))	94.05 %
Recall (true positives / (true positives + false negatives))	78.22 %



ScanProsite tool

This form requires to have JavaScript enabled to work correctly.

This form allows you to scan proteins for matches against the [PROSITE collection of motifs](#) as well as against your own patterns.

- ☐ Option 1 - Submit PROTEIN sequences to scan them against the PROSITE collection of motifs.
- ☒ **Option 2 - Submit MOTIFS to scan them against a PROTEIN sequence database.**
- ☐ Option 3 - Submit PROTEIN sequences and MOTIFS to scan them against each other.

[Reset](#)

STEP 1 - Enter a MOTIF or a combination of MOTIFS [Examples](#) [\[help\]](#)

[LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV]

Supported input:

- A PROSITE accession e.g. [PS50240](#) or identifier e.g. [TRYPSIN_DOM](#)
- Your own pattern e.g. [P-x\(2\)-G-E-S-G\(2\)-\[AS\]](#)

» [More](#)

» [Options](#) [\[help\]](#)

STEP2 - Select a PROTEIN sequence database [\[help\]](#)

- ☒ UniProtKB
 - ☒ Swiss-Prot ☒ Include isoforms
 - ☐ TrEMBL (sequences belonging to reference proteomes only)
 - ☐ PDB
 - ☐ Your protein database
 - ☐ Randomized UniProtKB/Swiss-Prot
-
- ☐ Exclude fragments (concerns UniProtKB only)

Pattern-Hit Initiated BLAST

Combina búsqueda por motivos (usa sintaxis de Prosite) con BLAST (PSI-BLAST en realidad)

Lo vamos a ver más adelante!

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

KIKSRFGHLVKCSMVNTKFGELPKAEIVGVYMKIHKTEEGEIVGLHQAQVPEI
QRDCRPFILLSLGSSELIQVRKEKFYDMVDEETRAKIIKMDVDYPSDEDLCSQSF
LKENDYIVFRKDLLRLLEPLNKSPFIPVQTKKKEIYNHKSFLDLCSLEKKVK
QHYPFLAPQKYLPLRVVQAISAPRHKIQELLQYKNAGVL

Query subrange [?](#)
From
To

Or, upload file No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Databases

☒ Standard databases (nr etc.): New ☐ Experimental databases

[Try experimental clustered nr database](#) [?](#)
For more info see [What is clustered nr?](#)

Standard

Database [?](#)

Organism Optional

☐ exclude [Add organism](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

☐ Exclude Optional

☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Program Selection

Algorithm

☐ Quick BLASTP (Accelerated protein-protein BLAST)

☐ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☒ PHI-BLAST (Pattern Hit Initiated BLAST)

Enter a PHI pattern [?](#)

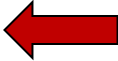
☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

Search database nr using PHI-BLAST (Pattern Hit Initiated BLAST)

☐ Show results in a new window

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

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



- 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 ANNUM (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

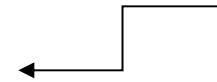
- Para extraer / generar
 - Patterns/Motifs
 - Fingerprints
 - Profiles 
 - HMMs
 - Position Specific Scoring Matrices / Weight matrices
- Para qué extraer / generar patterns, motifs, etc?
 - Para clasificar
 - Para alinear secuencias
 - Para buscar secuencias similares por métodos más sensibles

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   GVDVLVATPG  RLLDLEHQNA  ....VKLDQV  EILVLDEADR  MLDMGFIHDI
dbp2_schpo   GVEICIAATPG  RLLDMLDSNK  ....TNLRRV  TYLVLDEADR  MLDMGFEPQI
dbp2_yeast   GSEIVIAATPG  RLIDMLEIGK  ....TNLKR  TYLVLDEADR  MLDMGFEPQI
dbpa_ecoli   APHIIVATPG  RLLDHLQKGT  ....VSLDAL  NTLVMEADR  MLDMGFSDAI
rm62_drome   GCEIVIAATPG  RLIDFLSAGS  ....TNLKRC  TYLVLDEADR  MLDMGFEPQI
p68_human   GVEICIAATPG  RLIDFLECGK  ....TNLRRT  TYLVLDEADR  MLDMGFEPQI
rhlb_ecoli   GVDILIGTTG  RLIDYAKQNH  ....INLGAI  QVVVLDEADR  MYDLGFIKDI
yn21_caeel   RPHIIVATPG  RLVDHLENTK  ...GFNLKAL  KFLIMDEADR  ILNMDFEVEL
yhm5_yeast   KPHIIATPG  RLMDHLENTK  ...GFSRLKL  KFLVMDEADR  LLDMEFGPVL
me31_drome   KVQLIIATPG  RILDLMDDKV  ....ADMSHC  RILVLDEADK  LLSLDFQGML
drsl_yeast   RPDIVIAATPG  RFIDHIRNSA  ...SFNVDSV  EILVMDEADR  MLEEGFQDEL
if4a_rabbit  APHIIVATPG  RVFDMLNRRY  ....LSPKYI  KMFVLDEADE  MLSRGFKDQI
if41_human   APHIIVATPG  RVFDMLNRRY  ....LSPKYI  KMFVLDEADE  MLSRGFKDQI
vasa_drome   GCHVVIAATPG  RLLDFVDRTF  ....ITFEDT  RFVVLDEADR  MLDMGFSEDV
srmb_ecoli   NQDIVVATTTG  RLLQYIKEEN  ....FDCRAV  ETLILDEADR  MLDMGFAQDI
dead_ecoli   GPQIVVATPG  RLLDHLKRG  ....LDLSKL  SGLVLDEADE  MLRMGFIEDV
if4a_orysa   GVHVVVATPG  RVFDMLRRQS  ....LRPDYI  KMFVLDEADE  MLSRGFKDQI
dead_klepn   GPQIVVATPG  RLLDHLKRG  ....LDLSKL  SGLVLDEADE  MLRMGFIEDV
pl10_mouse   GCHLLVATPG  RLVDMMERGK  ....IGLDFC  KYLVLDEADR  MLDMGFEPQI
p54_human    TVHVVIATPG  RILDLIKKG  ....AKVDHV  QMIVLDEADK  LLSQDFVQIM
if4a_drome   GCHVVVATPG  RVYDMINRKL  ....RTQYI  KLFVLDEADE  MLSRGFKDQI
ded1_yeast   GCDLLVATPG  RLNDLLERGK  ....ISLANV  KYLVLDEADR  MLDMGFEPQI
ms16_yeast   RPNIVIAATPG  RLIDVLEKYS  ...NKFFRFV  DYKVLDEADR  LLEIGFRDDL
pr28_yeast   GCDILVATPG  RLIDSLENHL  ....LVMKQV  ETLVLDEADK  MYDLGFEDQV
if4n_human   GQHVVAGTGG  RVFDMIRRRS  ....LRTRAI  KMLVLDEADE  MLNKGFEQI
an3_xenla    GCHLLVATPG  RLVDMMERGK  ....IGLDFC  KYLVLDEADR  MLDMGFEPQI
dbp1_yeast   GCDLLVATPG  RLNDLLERGK  ....VSLANI  KYLVLDEADR  MLDMGFEPQI
if4a_yeast   DAQIVVATPG  RVFDNIQRRR  ....FRTDKI  KMFILDEADE  MLSSGFKEQI
spb4_yeast   RPQILIGTGG  RVLDFLQMPA  ....VKTSAC  SMVVMDEADR  LLDMSFIKDT
if4a_caeel   GIHVVVATPG  RVGDMINRNA  ....LDTSRI  KMFVLDEADE  MLSRGFKDQI
pr05_yeast   GTEIVVATPG  RFIDILTND  .GKLLSTKRI  TFVVMDEADR  LFDLGFEQI
if42_mouse   APHIVVATPG  RVFDMLNRRY  ....LSPKWI  KMFVLDEADE  MLSRGFKDQI
dhh1_yeast   TVHILVATPG  RVLDLASRKV  ....ADLSDC  SLFIMDEADK  MLSRDFKTI
db73_drome   KADIVVTATPG  RLVDHLHATK  ...GFCLSL  KFLIVDEADR  IMDAVFQNL
yk04_yeast   GCNFIIGTGG  RVLDDLQNTK  VIKEQLSQSL  RYIVLDEGDK  LMELGFDETI
ybz2_yeast   SGQIVIAATPG  RFLELLEKDN  .TLIKRFSKV  NTLILDEADR  LLQDGHFDEF
yhw9_yeast   KPHFIIATPG  RLAHHIMSSG  DDTVGGLMRA  KYLVLDEADI  LLTSTFADHL
glh1_caeel   GATIIIVATGG  RIKHFCEEGT  ....IKLDKC  RFFVLDEADR  MIDAMGFGTD
```

Un profile generado a partir del MSA

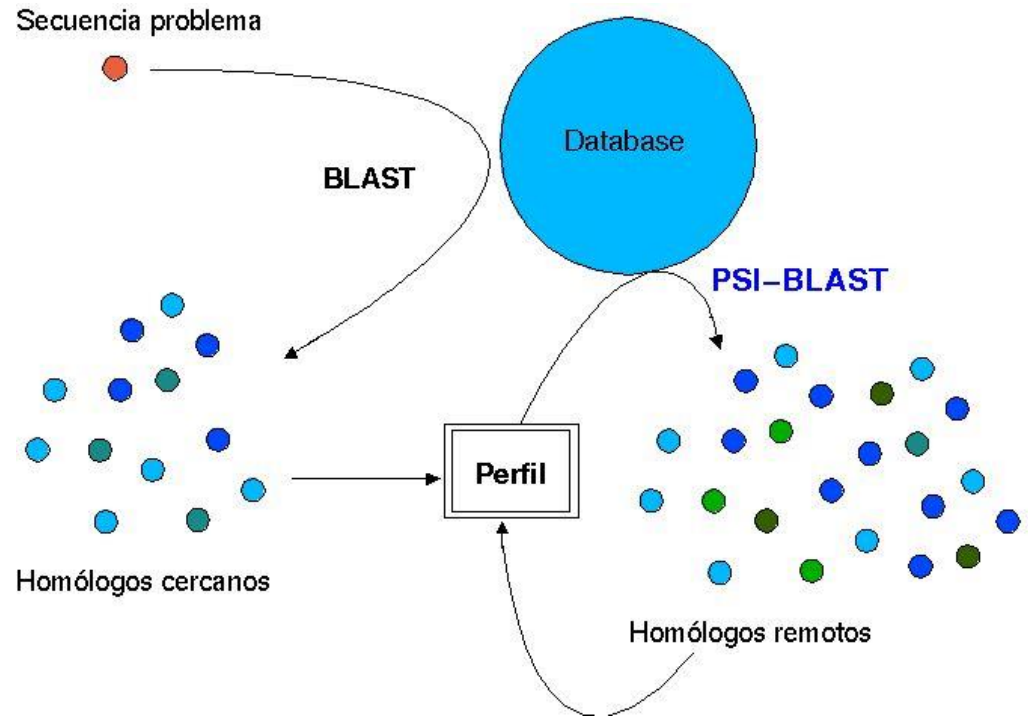
Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	Gap	Len	..
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11	100	100	
P	18	0	13	0	0	-12	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1	100	100	
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27	100	100	
I	-1	-12	6	-13	-11	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

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



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

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

Profile HMMs

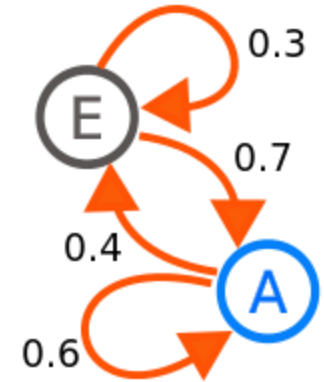
- La información contenida en un profile puede representarse de otras formas
- Los profiles originales contienen scores y penalidades basados en las frecuencias de ocurrencia
- Un profile (o un MSA) también puede representarse como una cadena de eventos con probabilidades de ocurrencia (Markov Chain)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

Markov Chains: una pequeña intro

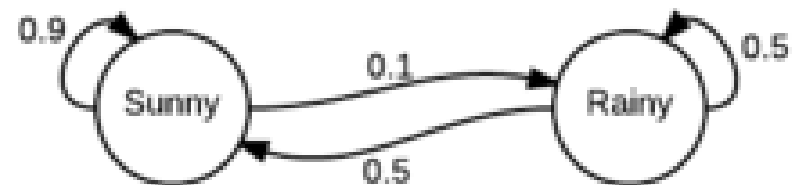
Una cadena de Markov es un sistema matemático que *transita* entre distintos *estados*, de acuerdo a probabilidades

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)

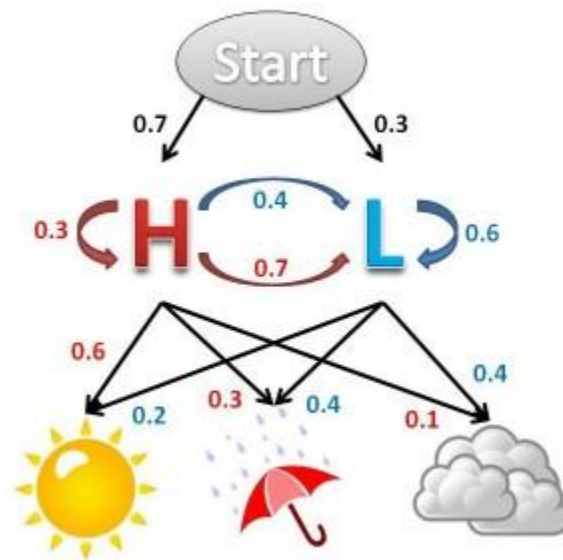
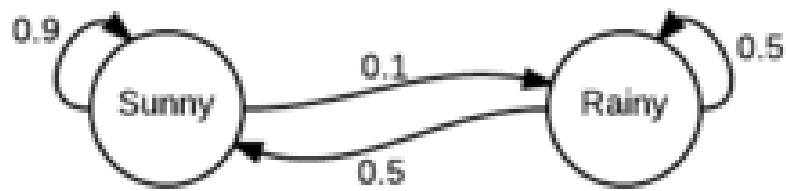


Markov Chain, Wikipedia. http://en.wikipedia.org/wiki/Markov_chain



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

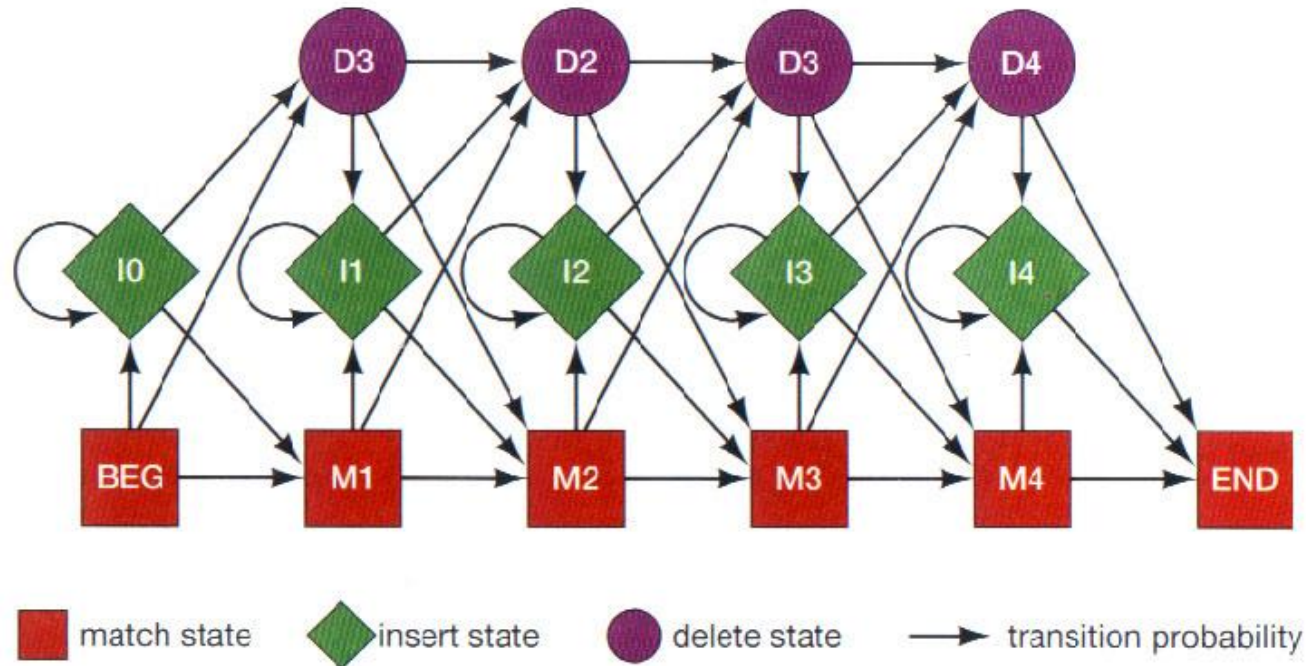
El profile ahora
está codificado en
forma de **estados**
y **probabilidades**
de transición

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.org>



- **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 35.0 (Noviembre 2021, 19632 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](#)

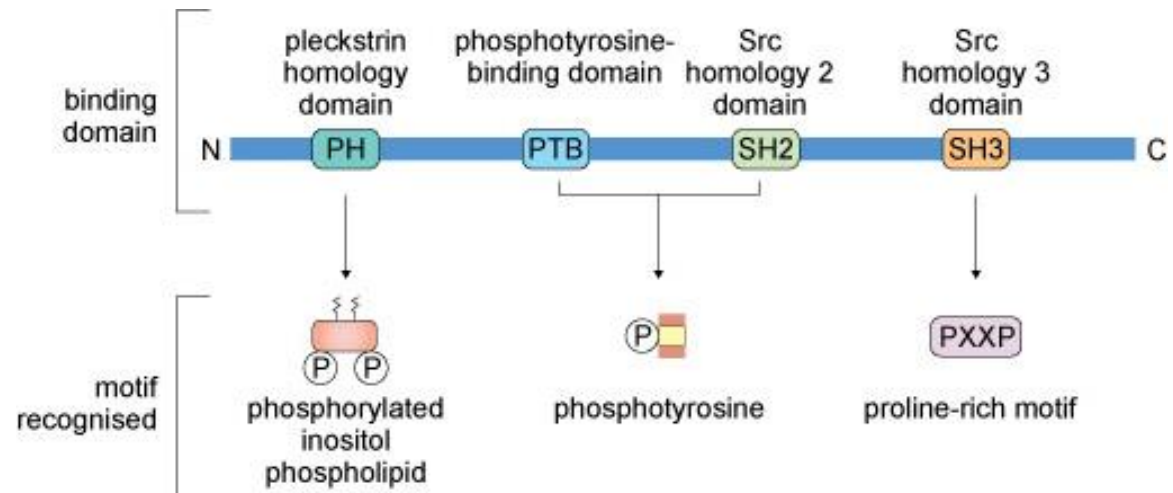
Usos de los MSAs

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

Por qué Weight Matrices (o PSSMs)?

La gran mayoría de los motivos biológicos pueden caracterizarse por **motifs**

- **Modificaciones post-traduccionales (ej Asn X Ser)**
- **Signal Peptides**
- **T-cell epitopes**
- **Transcription factor binding sites (ej TATA box)**
- **SH2/SH3 domain binding**
- **MHC binding**
- ...



Identificando motivos

Peptide

LMLSLFEQSLSCQAQ

QGTDATKSIIFEAEER

RLEEAQAYLAAGQHD

EISELRTKVQEQQKQ

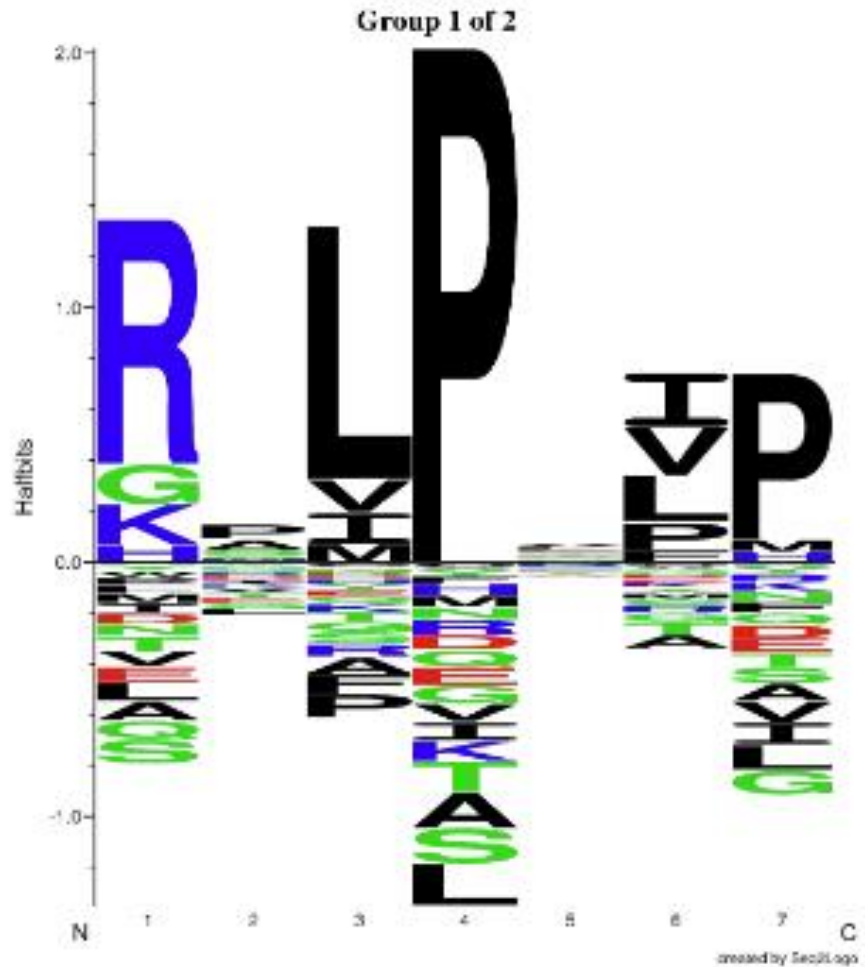
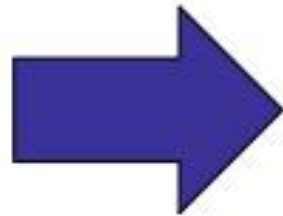
FAGAKKIFGSLAFLP

VRASSRVSGSFPEDS

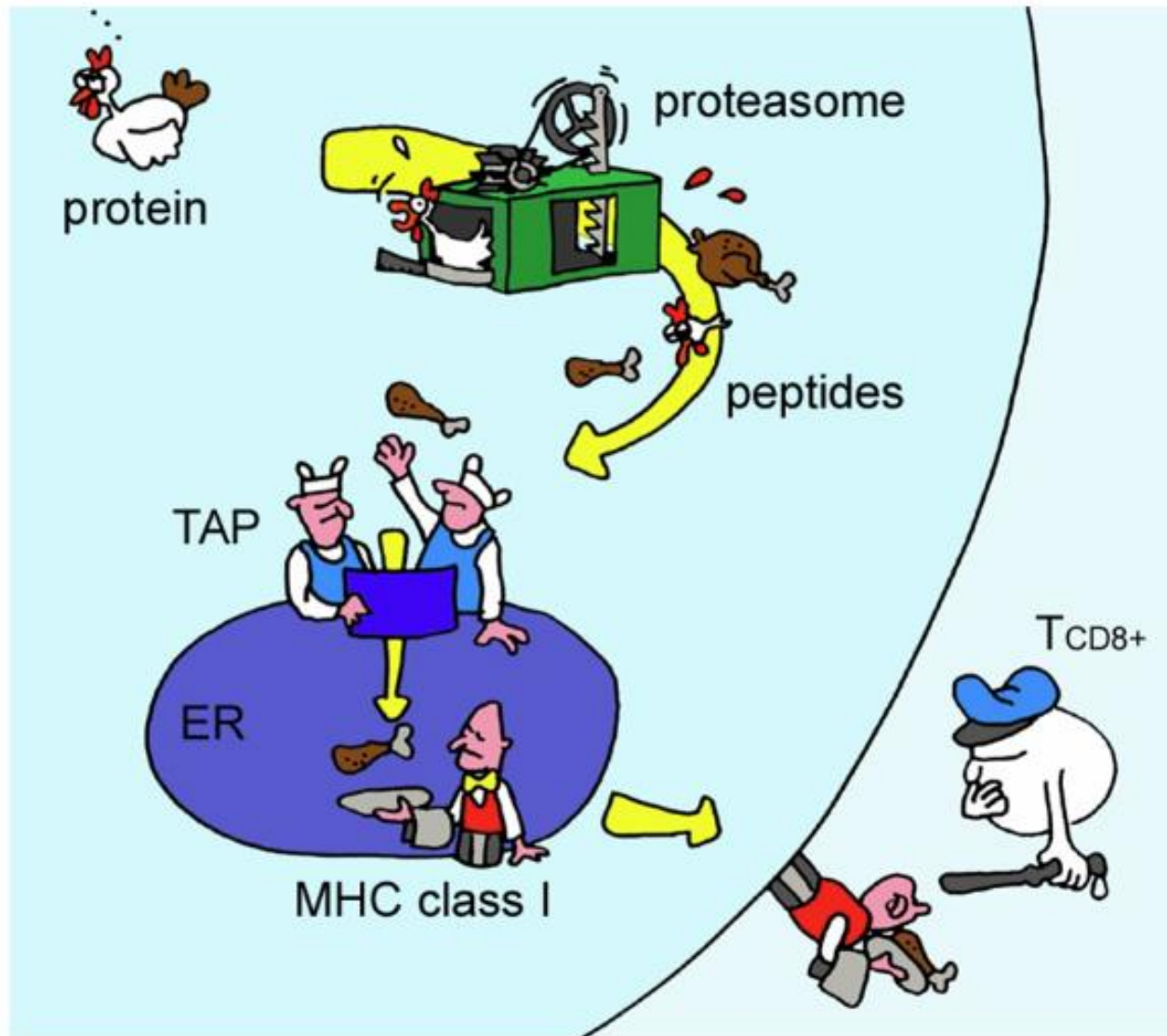
CKAFFKRSIQGHNDY

CEGCKAFFKRSIQGH

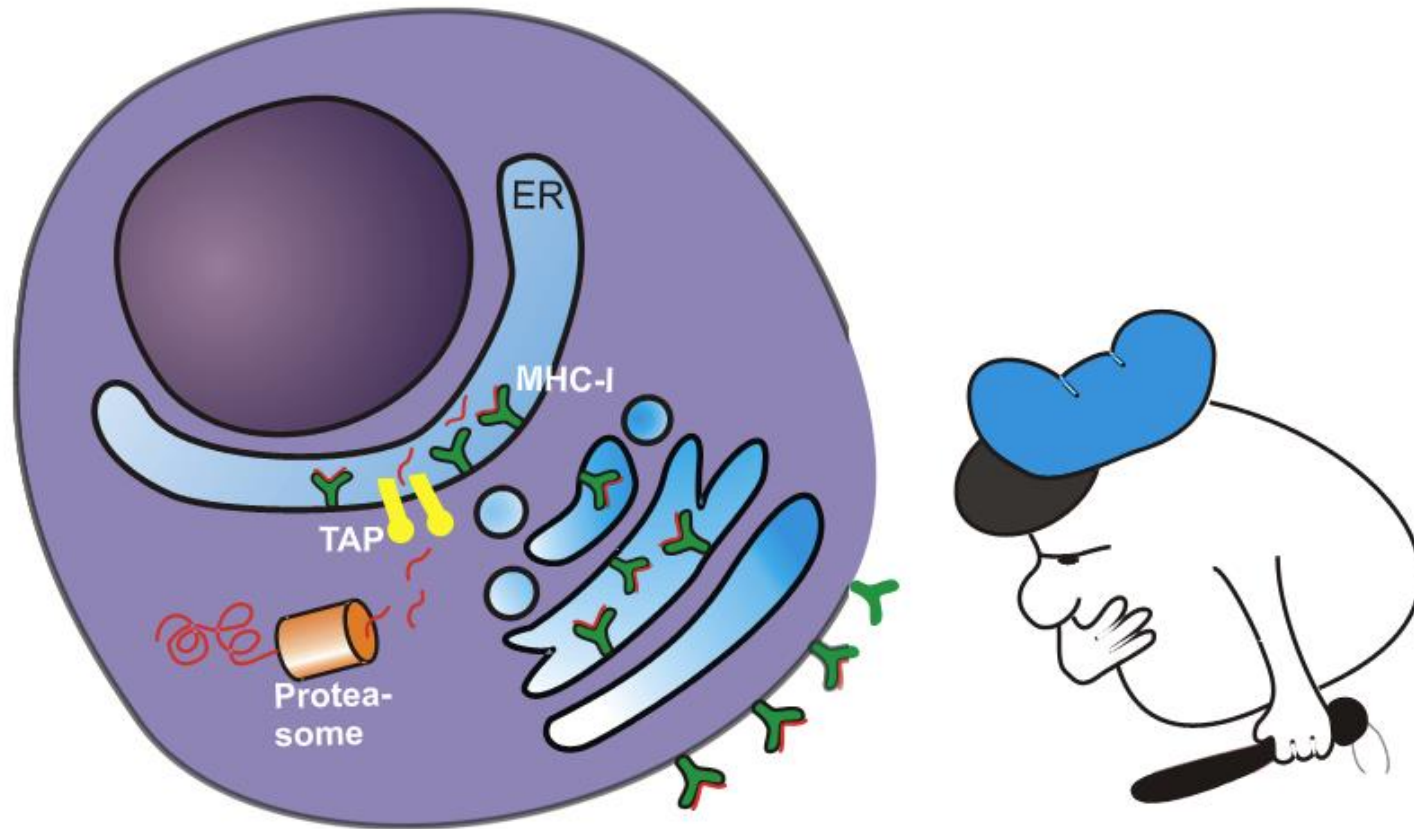
RLSEADIRGFVAAVV



Binding de péptidos en el contexto MHC



Patrullaje inmune



Encuentro con la muerte

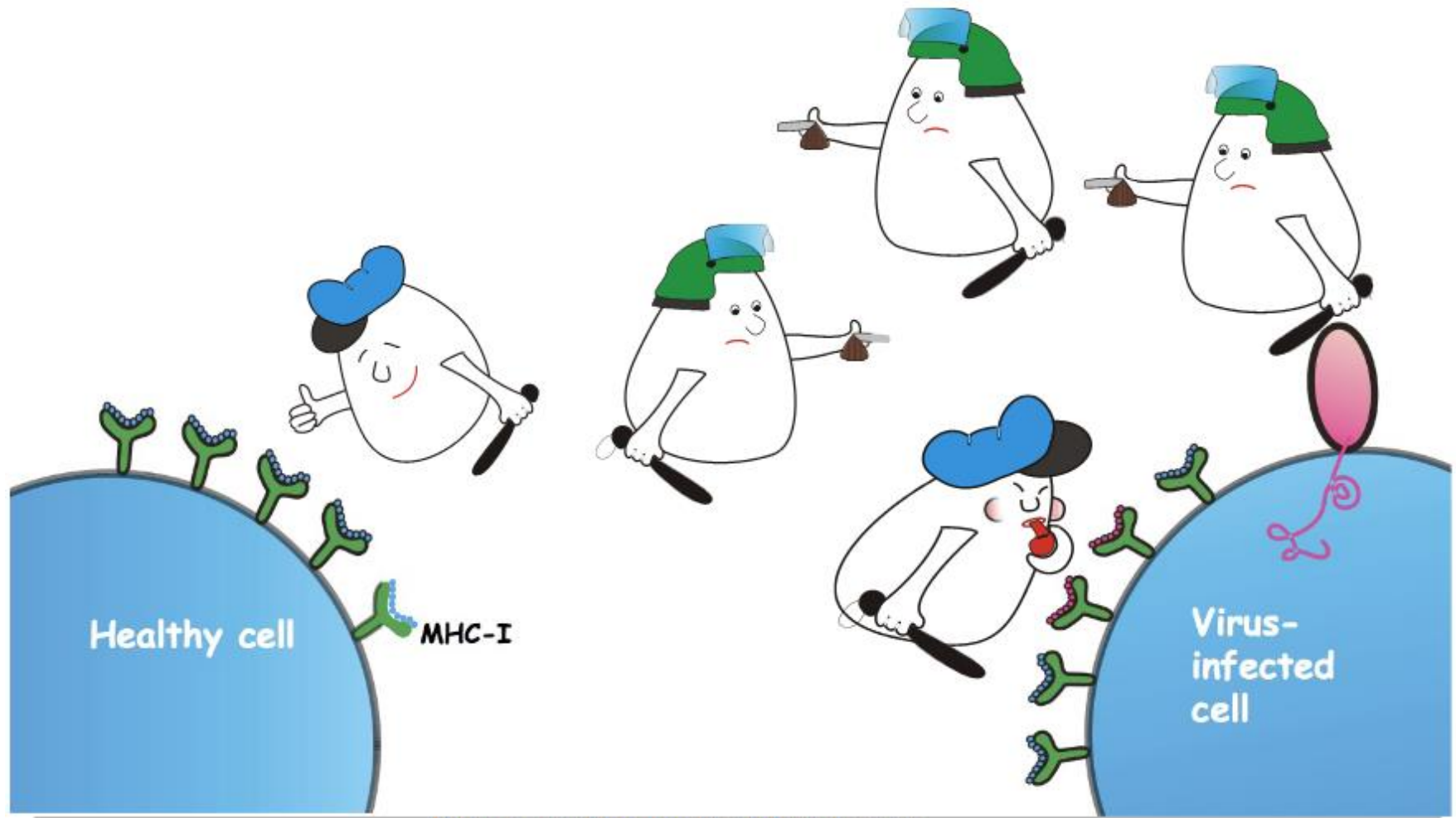


Figure courtesy Mette Voldby Larsen

Objetivos

- Entender los conceptos utilizados en la construcción de **matrices de pesos**
- **Visualización de motivos de binding**
 - Construcción de Sequence Logos
- **Construir una matriz de pesos**

PSSMs – Weight matrices

- 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 de pesos

Información de secuencia

Datos experimentales: secuencias de péptidos que se unen a un determinado receptor MHC

Que información podemos obtener de las secuencias?

Un péptido debe tener L (Leu) en la posición 2 (P2) para unirse al MHC.

En la posición 1 (P1) observamos Ala (A), Phe (F), Trp (W), y Tyr (Y).

Qué posición tiene más información?

ALAKAAAAM
FLAKAAAAN
WLAKAAAAR
ALAKAAAAT
YLAKAAAAY
FLNERPILT
WLLGFVFTM
YLNAAVKVV
ALNEPVLLL
... ..
... ..
WLVPFIVSV

Información experimental:
péptidos que se unen a un
determinado receptor MHC

Información e incertidumbre

Qué pasa si nos dan una secuencia X y queremos predecir si se va a unir a este receptor MHC?

Cuánta **incertidumbre** tenemos antes de saber cuál va a ser el residuo en P1? Y en P2?

Cuánta incertidumbre perdimos (y cuánta **información** ganamos) si sabemos que el residuo en P1 es Phe (F)?
Y si el residuo en P2 es Leu (L)?

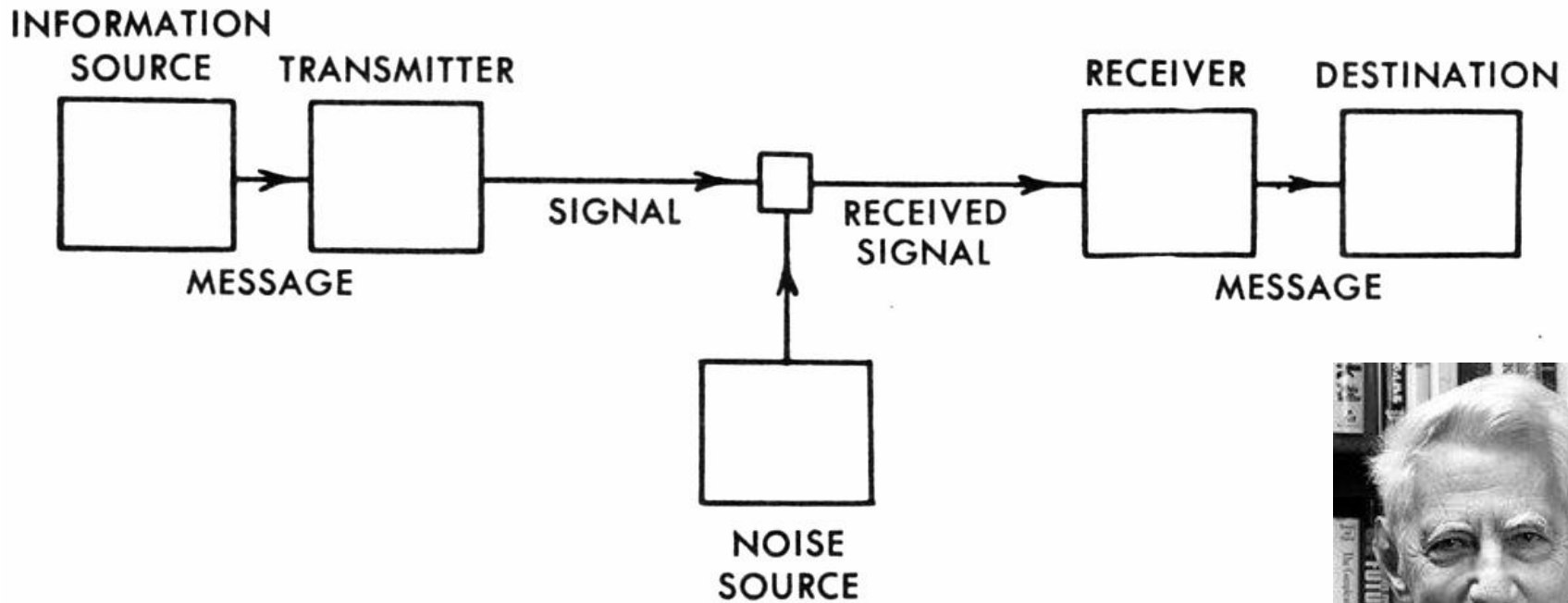
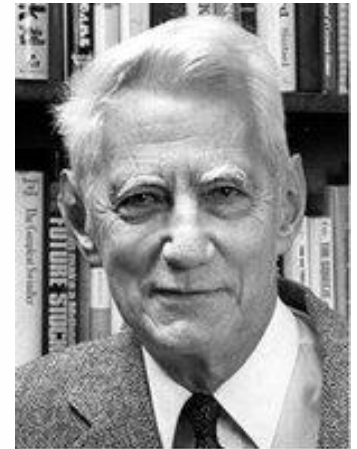


Fig. 1. — Schematic diagram of a general communication system.



Claude Shannon

Information theory

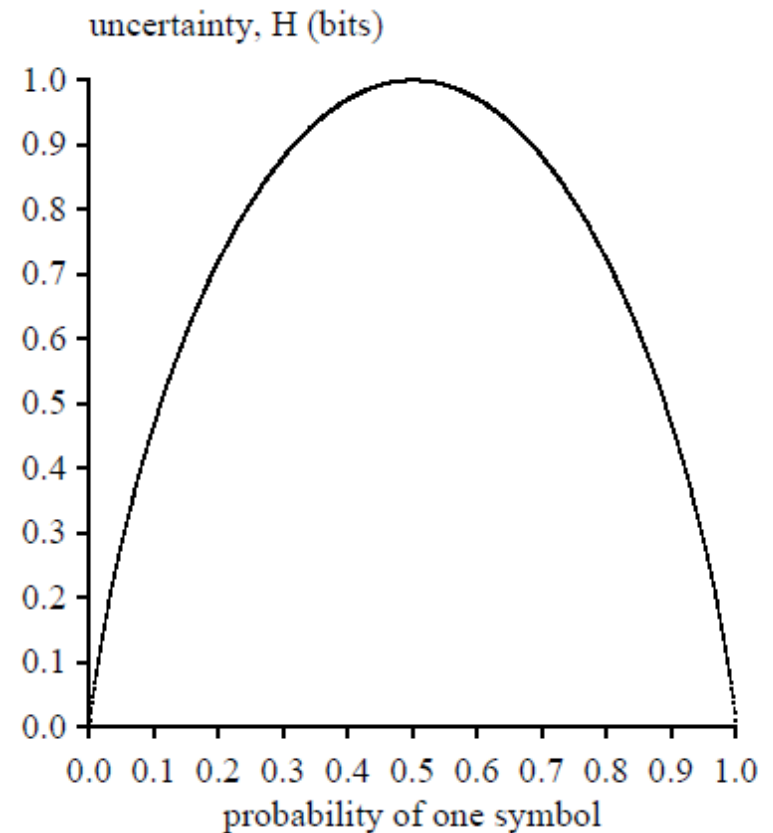
- **Entropía: medida de desorden de un sistema**
- **La termodinámica provee herramientas para calcular entropía**
- **El desorden implica falta de *información* sobre el estado exacto de un sistema**
- **Claude Shannon / Leon Brillouin**
 - **Information theory**
 - **La Información es una combinación de**
 - Certain + Uncertain, Expected + Unexpected
 - **El grado de *sorpresa* que genera un evento que ya ocurrió es *cero***
 - **Cómo cambia la información con la probabilidad de ocurrencia?**

Midiendo incertidumbre

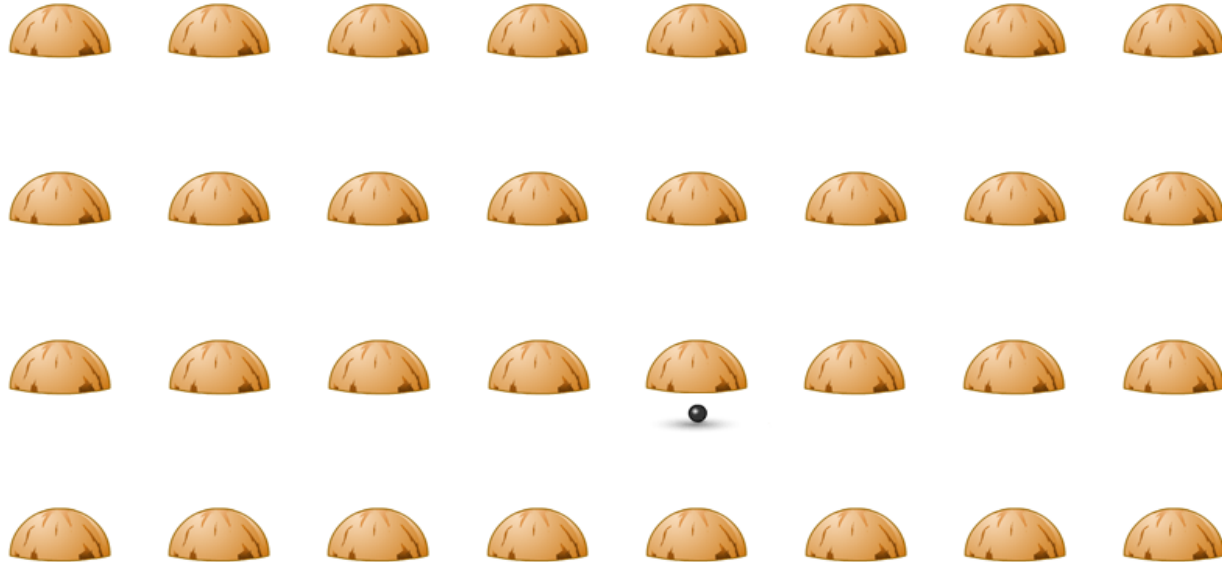
Para un sistema que puede producir solamente dos **símbolos** con distintas probabilidades, podemos medir la incertidumbre (H , uncertainty)

Ejemplos de sistemas binarios:

- cara, ceca
- SI, NO
- VERDADERO, FALSO
- 0, 1
- Activo, inactivo



Shell game



Shell Game (Thimblorig)

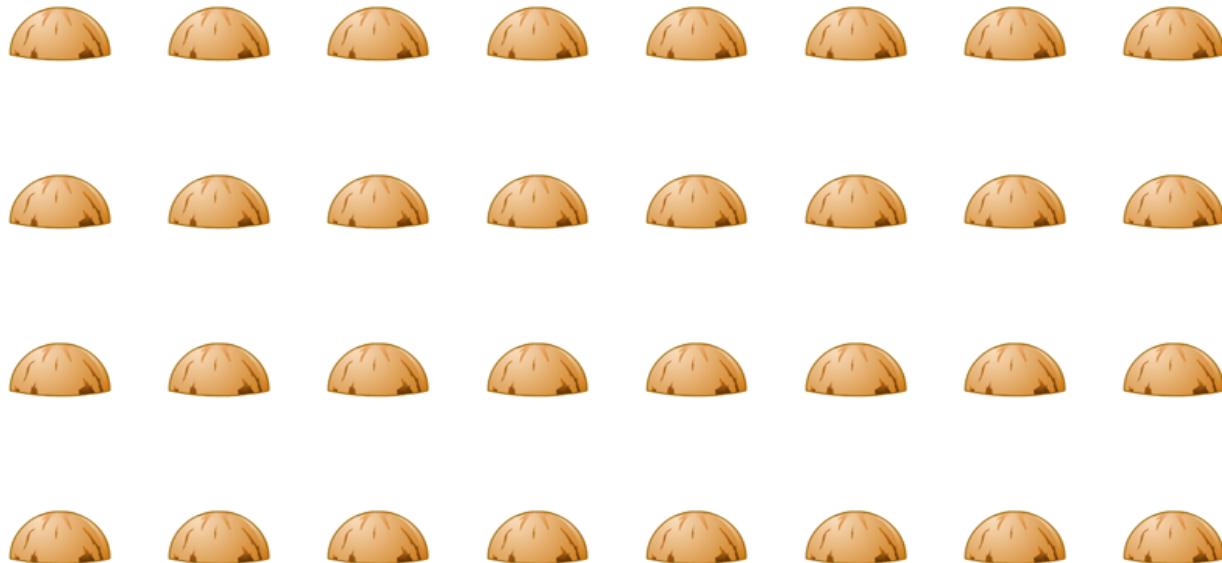
Adivinar en qué taza / nuez
está escondida la bolita.

Uncertainty

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

Probability

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



Shell game



Shell Game (Thimblergig)

Adivinar en qué taza /
nuez está escondida
la bolita.

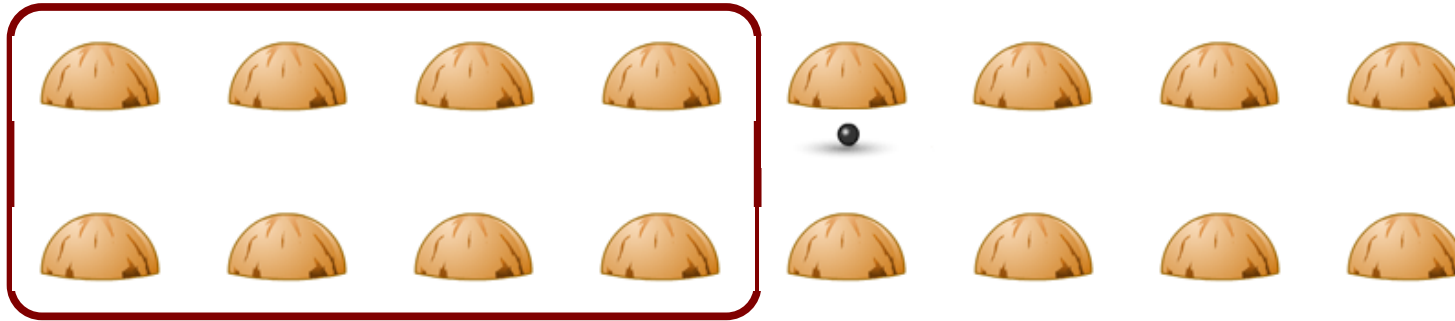
Uncertainty

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

Probability

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

Shell game



Shell Game (Thimblrig)

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

$p(object) = 1/64$

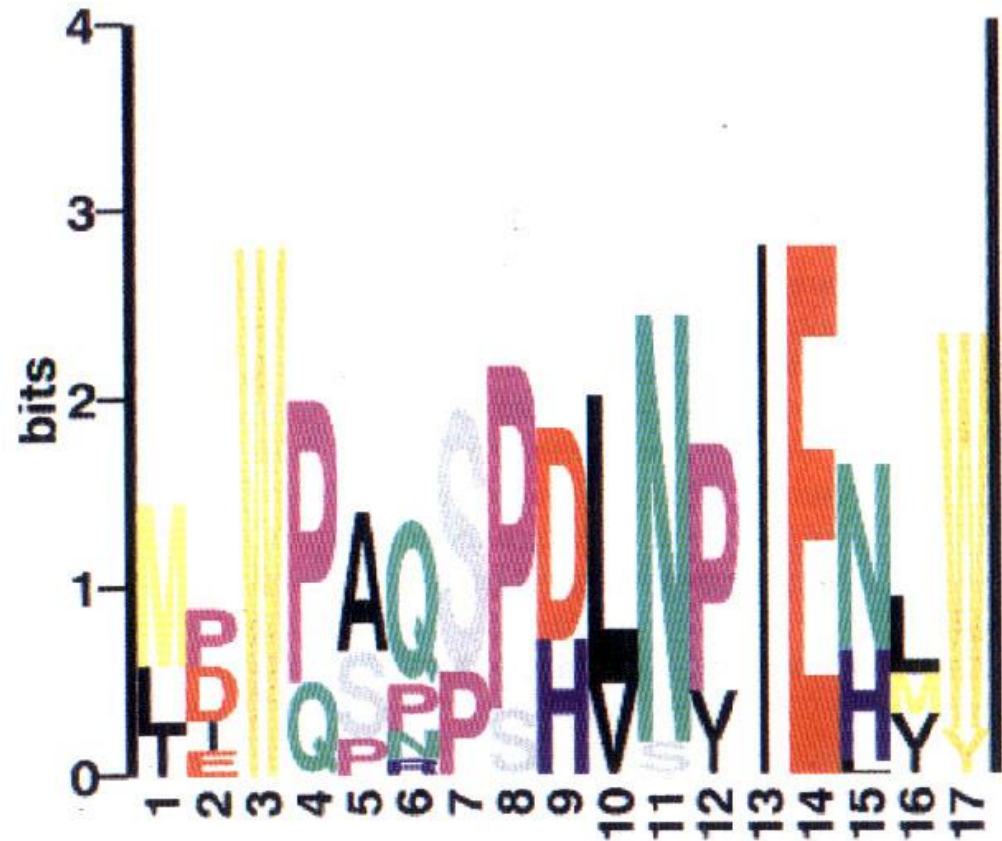
- Las preguntas secuenciales reducen las posibilidades (incertidumbre) de 64 a 32, luego a 16, 8, 4, 2, y finalmente 1.
- 6 preguntas son suficientes (peor caso) para encontrar la bolita.
- Esta es una manera de cuantificar la incertidumbre
- La incertidumbre también se puede calcular a partir de las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



- **Integra varias otras bases de datos en un solo lugar y provee referencias (links)**
 - **<http://www.ebi.ac.uk/interpro>**
 - **CATH, CDD, Hamap, NCBI Fam, Prosite, PRINTS, Pfam, Panther, PirSF, ProDom, SMART ...**

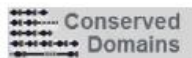


Databases in InterPro



CATH-Gene3D ⓘ ↗

4.3.0
7k entries



CDD ⓘ ↗

3.20
18k entries



HAMAP ⓘ ↗

2023_05
2k entries



NCBIfam ⓘ ↗

15.0
7k entries



PANTHER ⓘ ↗

18.0
16k entries



Pfam ⓘ

37.0
22k entries



PIRSF ⓘ ↗

3.10
3k entries



PRINTS ⓘ ↗

42.0
2k entries



PROSITE
profiles ⓘ ↗

2023_05
1k entries



PROSITE
patterns ⓘ ↗

2023_05
1k entries



SFLD ⓘ

4
299 entries



SMART ⓘ ↗

9.0
1k entries



SUPERFAMILY ⓘ ↗

1.75
2k entries

Search by protein families, domains, proteins, keywords, or GO terms



Examples: *IPR020422*, *kinase*, *O00167*, *PF02932*, *GO:0007165*, *1t2v*, *UP000005640*

Search

Clear

Powered by EBI search

1 - 20 of 181 entries



Export



Accession	Name	Source Database	Description
IPR040856	Glucose ABC transporter, C-terminal	INTERPRO	This is the C-terminal domain found at the ATPase subunit of the glucose ABC transporter from <i>Sulfolobus solfataricus</i> . Overall, the C-terminal domain (residues 243-353) contains only β -strands, which ...
PF17847	Glucose ABC transporter C-terminal domain	PFAM	This is the C-terminal domain found at the ATPase subunit of the glucose ABC transporter from <i>Sulfolobus solfataricus</i> . Overall, the C-terminal domain (residues 243-353) contains only beta-strands, whi...
TIGR00826	glucose PTS transporter subunit EIIB	NCBIFAM	The PTS Glucose-Glucoside (Glc) Family (TC 4.A.1) Bacterial PTS transporters transport and concomitantly phosphorylate their sugar substrates, and typically consist of multiple subunits or protein dom...
cd17435	Glucose transporter type 12 (GLUT12), a Class 3 GLUT, of the Major Facilitator Superfamily of transporters	CDD	Glucose transporter type 12 (GLUT12) is also called Solute carrier family 2, facilitated glucose transporter member 12 (SLC2A12). It is a facilitative glucose transporter, classified as a Class 3 GLUT...
TIGR01272	glucose/galactose MFS transporter	NCBIFAM	This model describes the glucose/galactose transporter in bacteria. This belongs to the larger facilitator superfamily. Disruption of the loci leads to the total loss of glucose or galactose uptake in...
cd17433	Glucose transporter type 8, a Class 3 GLUT, of the Major Facilitator Superfamily of transporters	CDD	Glucose transporter type 8 (GLUT8) is also called Solute carrier family 2, facilitated glucose transporter member 8 (SLC2A8) or glucose transporter type X1 (GLUTX1). It is classified as a Class 3 GLUT...

Rows per page: 20

Previous 1 2 3 4 5 6 7 8 9 10 Next



IPR040856

Glucose ABC transporter, C-terminal ★

InterPro entry ⓘ



Overview

Proteins 68

Domain Architectures 5

Taxonomy 94

Proteomes 12

Structures 5

AlphaFold 53

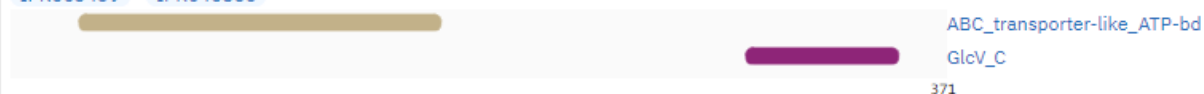
Pathways 7

5 domain architectures

Export ▾

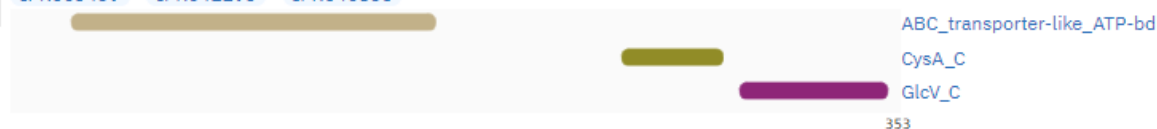
There are 56 proteins with this architecture (represented by Q97UF2):

IPR003439 - IPR040856



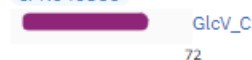
There are 5 proteins with this architecture (represented by A0A1C8ZTB7):

IPR003439 - IPR041193 - IPR040856



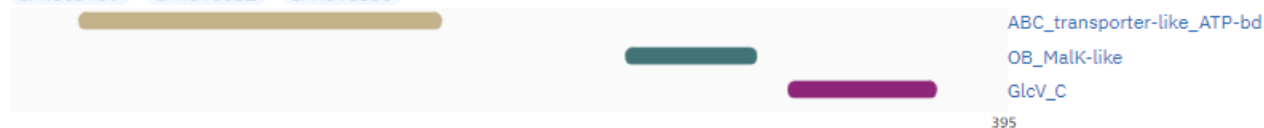
There are 4 proteins with this architecture (represented by C3NA79):

IPR040856



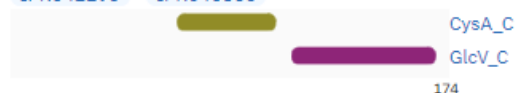
There are 2 proteins with this architecture (represented by A0A8J7YNG2):

IPR003439 - IPR040582 - IPR040856



There is 1 protein with this architecture (represented by A0A550D0W8):

IPR041193 - IPR040856



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

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/>