

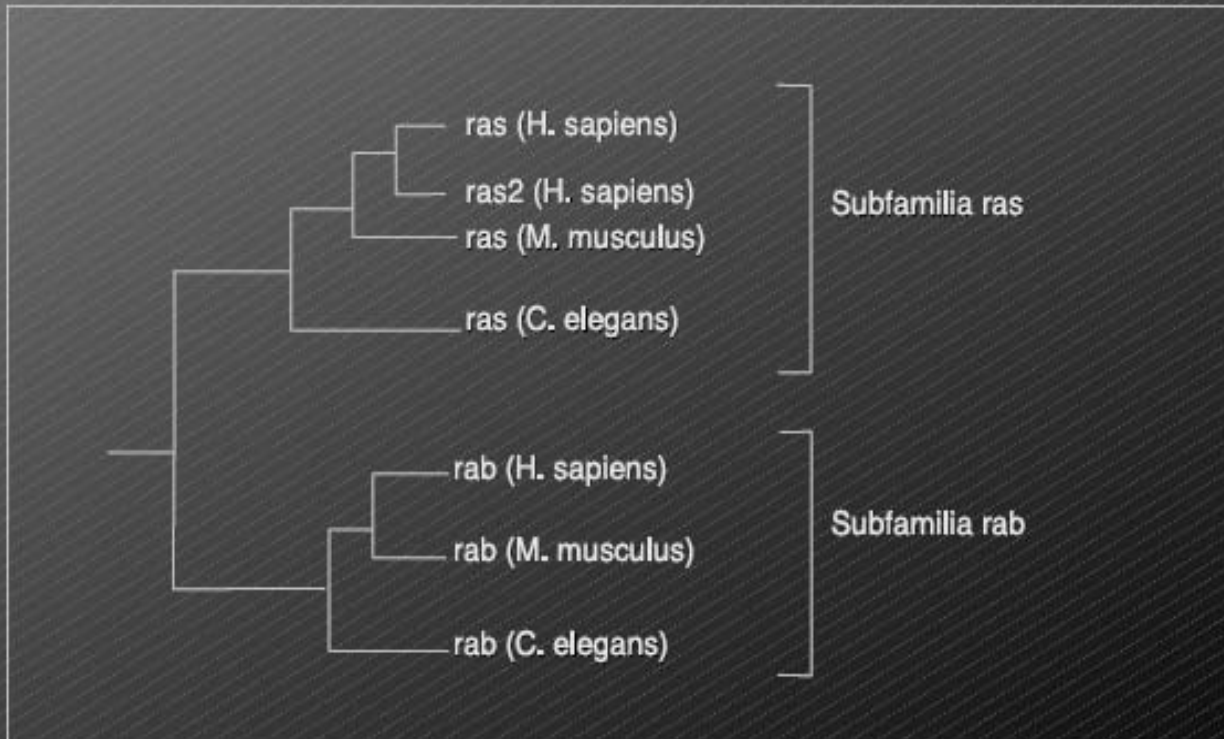
FAMILIAS DE PROTEÍNAS

Lo que encontramos en las bases de datos

Observación: las proteínas homólogas pueden tener funciones distintas.

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

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

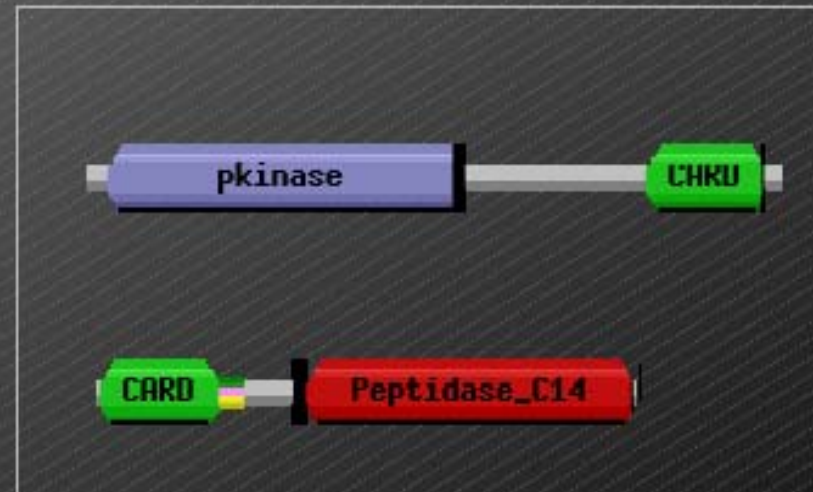


Barajado de dominios (domain-shuffling)

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

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

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



Homólogos: ortólogos y parálogos.

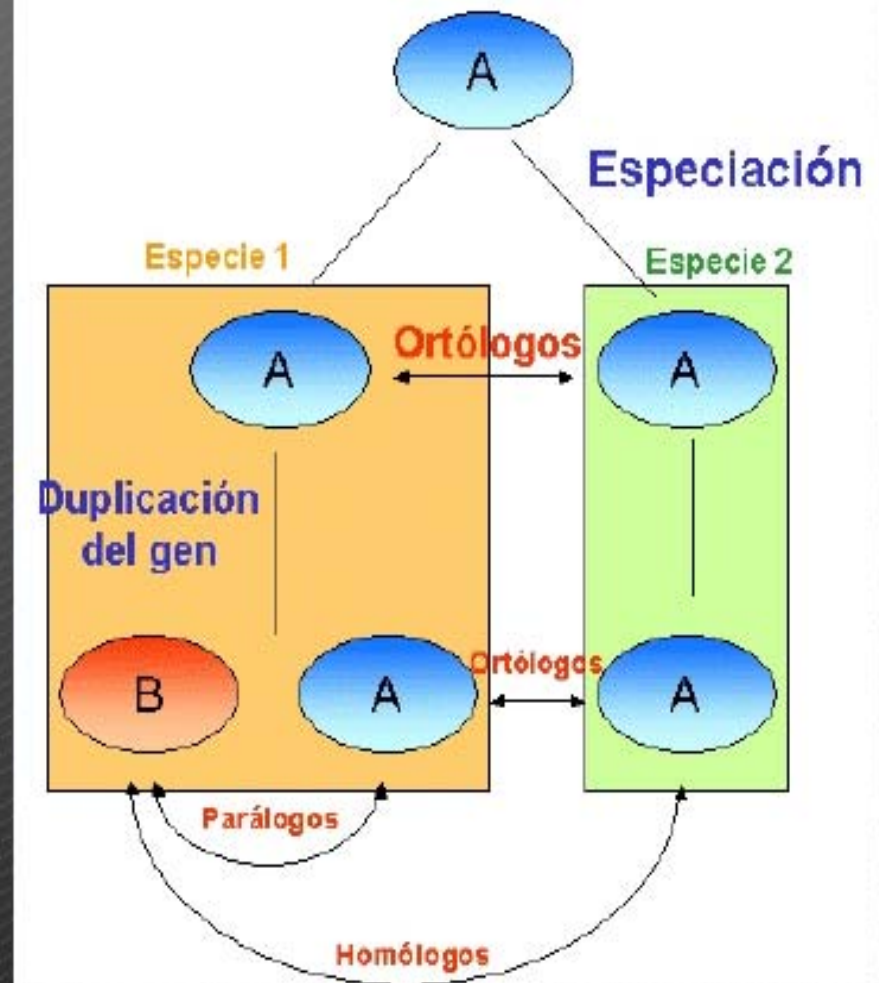
Ortólogos: genes que comparten el último ancestro común y cuya divergencia se debe a la especiación.
Ejemplo: isomerasa de glucosa-6P de *Bacillus subtilis* y de *Escherichia coli*.

Los mismos genes en distintas especies.

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

Ejemplo: tripsina, quimiotripsina, elastasa y trombina.

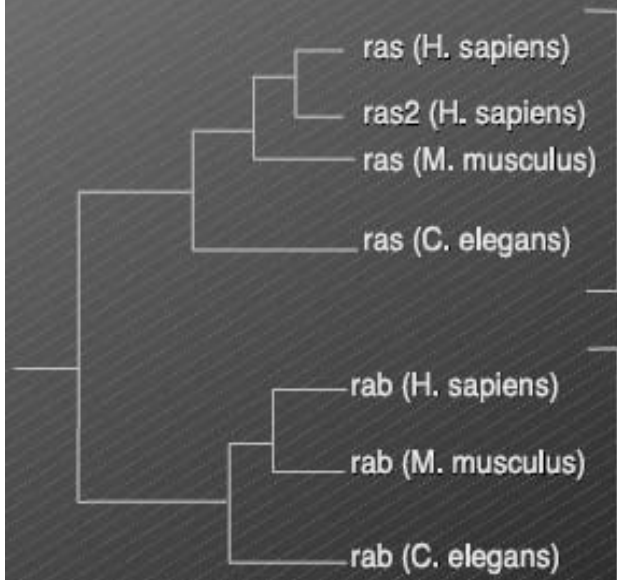
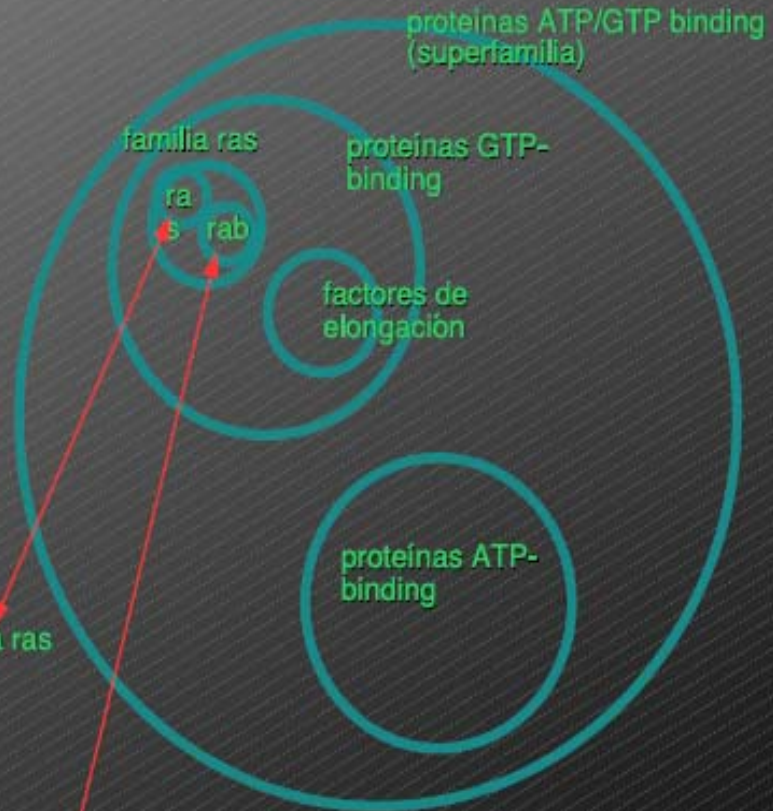
Homólogos/Ortólogos/Parálogos



Homólogos: superfamilias, familias y subfamilias.

Superfamilia: grupo de proteínas con un origen común.

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



Subfamilia ras

Subfamilia rab

Dos formas de representarlo

Interés de analizar la organización en familias de las proteínas

Predicción de función.

| File | Edit | Colour | Sort | Picked |
|--|------|--------|---|------------------------|
| (36x635) -----300-----310-----320-----330-----340-----350-----360----- | | | | |
| 1ba1 | 4 | 376 | QATKDAGT.IAG.....LNHLRIINEPTAAAIAYGLDKKVGAEARNVLI | FDLGGGTFDWSILTIEDG... |
| HS7C_BOVIN | 4 | 377 | QATKDAGT.IAG.....LNHLRIINEPTAAAIAYGLDKKVGAEARNVLI | FDLGGGTFDWSILTIEDG... |
| HS7C_BOVIN | 4 | 377 | QATKDAGT.IAG.....LNHLRIINEPTAAAIAYGLDKKVGAEARNVLI | FDLGGGTFDWSILTIEDG... |
| HS7C_MOUSE | 4 | 377 | QATKDAGT.IAG.....LNHLRIINEPTAAAIAYGLDKKVGAEARNVLI | FDLGGGTFDWSILTIEDG... |
| HS7D_DROME | 4 | 377 | QATKDAGT.IAG.....LNHLRIINEPTAAAIAYGLDKKVGAEARNVLI | FDLGGGTFDWSILTIEDG... |
| HS7Q_XENLA | 5 | 378 | QATKDAGV.LAG.....LNHLRIINEPTAAAIAYGLDKGARGEQNVLI | FDLGGGTFDWSILTIDDG... |
| 1dkgD | 4 | 375 | QATKDAGR.IAG.....LEVKRIINEPTAALAYGLDK...TGNRTIAY | DLGGGTFDISIIEIDEK... |
| DNAK_PASMU | 2 | 378 | QATKDAGR.IAG.....LEVKRIINEPTAALAYGLDKGGG.NRTIAY | DLGGGTFDISIIEIDEV... |
| DNAK_SALTY | 1 | 377 | QATKDAGR.IAG.....LEVKRIINEPTAALAYGLDKEVG.NRTIAY | DLGGGTFDISIIEIDEV... |
| DNAK_VIBCH | 2 | 377 | QATKDAGR.IAG.....LEVKRIINEPTAALAYGLDKGGG.DRTIAY | DLGGGTFDISIIEIDEV... |
| DNAK_BURPS | 2 | 379 | QATKDAGR.IAG.....LEVKRIINEPTAALAFGLDKAEKGRKIAY | DLGGGTFDWSIIEIADVDG... |
| DNAK_BURCE | 2 | 380 | QATKDAGR.IAG.....LEVKRIINEPTAALAFGLDKAEKGRKIAY | DLGGGTFDWSIIEIADVDG... |
| 1jseA | 4 | 322 | RAILDAGL.EAG.....ASKVFLTEEPKAAIGSNLN...VEEPSGNKVV | DIIGGTTTEVAVISL... |
| MREB_067013 | 11 | 328 | RAVVDAAK.SAG.....AREVYLVAEPMAAIGALP...VEEPIGNMIV | DIIGGTTTDAIVISLA... |
| MREB_BACSU | 6 | 325 | RAVIDATR.QAG.....ARDAYPIEEPFAAIGANLP...VWEPTGSMVY | DIIGGTTTEVAVISL... |
| MREB_09K8H5 | 6 | 325 | RAVEDATK.QAG.....AKYAYTLEEPFAAIGADLP...VWEPTGSMVY | DIIGGTTTEVAVISL... |
| MREB_092B66 | 6 | 324 | RAVIDATR.QAG.....AKDAFTIEEPFAAIGALP...VGEPTGSMVY | DIIGGTTTEVAVISL... |
| MREB_09L1G6 | 9 | 326 | RAVIEASS.QAG.....ARQVHIIEEPMAAIGCSLP...VHEATGSMVY | DIIGGTTTEVAVISL... |
| 1e4FT | 8 | 384 | EMFYNFLQDTVK.....S.PFQLKBSLSVSTAEGVLT...PEKDRGVVYV | HLGYNFTGLIAYKN... |
| FTSA_ENTHR | 5 | 379 | HNIRKCVENAGL.....V.VNELVITPLALTETILSD...GEKDFGTIV | IDMGGGTTTAVMHD... |
| FTSA_ENTFA | 1 | 375 | HNIRKCVKAGL.....G.INELVITPLALTETILT...GEKDFGTIV | IDMGGGTTTAVMHD... |
| FTSA_BACSU | 5 | 379 | HNLRCVERAGI.....E.ITDICLPLAAGSALS...DEKNLGVALL | DIIGGTTTAVMFD... |
| FTSA_BORBU | 5 | 378 | QNLVRCVHRAGF.....A.VDEWLGSLASSYATLSK...EEREMGVLF | IDMGGGTTDIIYID... |
| FTSA_ECOLI | 8 | 383 | KNIVKAVERCGL.....K.VDQLIFAGLASSYSVLTE...DERELGVCVY | DIIGGTTMDIAYTG... |
| 1yagA | 5 | 346 | EKMTQIMFETFN.....VPAFYVSIQAVLSLYASRT.....TGIVLIS | SGDGVTHVYPIYA... |
| ACT_BOTCI | 5 | 346 | EKMTQIVFETFN.....APAFYVSIQAVLSLYASRT.....TGIVLIS | SGDGVTHVYPIYE... |
| ACT_LNEUCR | 5 | 346 | EKMTQIVFETFN.....APAFYVSIQAVLSLYASRT.....TGIVLIS | SGDGVTHVYPIYE... |
| ACT4_CAEL | 6 | 347 | EKMTQIMFETFN.....TPAMYVAIQAVLSLYASRT.....TGIVLIS | SGDGVTHVYPIYE... |
| ACT5_HUMAN | 5 | 346 | EKMTQIMFETFN.....TPAMYVAIQAVLSLYASRT.....TGIVLIS | SGDGVTHVYPIYE... |
| ACT5_CHICK | 6 | 347 | EKMTQIMFETFN.....TPAMYVAIQAVLSLYASRT.....TGIVLIS | SGDGVTHVYPIYE... |
| 1qhaA | 78 | 456 | ADVVKLLN.KAIKKRGDYDANIVAVVNDYGTMMTCGYD...DQHCEVGLII | GTG.TNACYMEELRHIDLV |
| HXK1_HUMAN | 78 | 456 | ADVVKLLN.KAIKKRGDYDANIVAVVNDYGTMMTCGYD...DQHCEVGLII | GTG.TNACYMEELRHIDLV |
| HXK1_BOVIN | 78 | 456 | HYVVKLLD.KAIKKRGDYDANIVAVVNDYGTMMTCGYD...DQHCEVGLII | GTG.TNACYMEELRQIDFG |
| HXK_SCHMA | 68 | 443 | HNVAELLD.TELDKRE.LNVKCAVWVNDYGTLASCALE...DPKCAVGLIV | GTG.TNWAYIEDSSKVELM |
| HXK2_DROME | 128 | 505 | KNVVSLLQ.EAIDRRGDLKINTVAIINDYGTLMSCAFY...HPNCRIGLIV | GTG.SNACYVEKTVHAECE |
| HXK1_SPIOL | 95 | 485 | EDVVAELT.KAHLRKG.VDMRWALVNDYGTLAGGRYY...KEDVIARVIL | GTG.TNARYVERASAIHKM |

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

Cómo analizar la organización en familias de las proteínas

Arboles filogenéticos: lo más fiable, pero es laborioso y hay que hacerlo manualmente
(lo veréis el próximo día)

Bases de datos construidas por expertos:

Pfam

Prosite

InterPro

...

Métodos automáticos:

ProtoMap

COGs

...

Prosite

PROSITE:

<http://us.expasy.org/prosite/>

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

-gran cantidad de información para cada familia de proteínas.

-baja cobertura: sólo 1.245 familias

```
ID MOLYBDOPTERIN_EUK; PATTERN.
AC PS00559;
DT DEC-1991 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE Eukaryotic molybdopterin oxidoreductases signature.
PA [GA]-x(3)-[KRNQHT]-x(11,14)-[LIVMPYWS]-x(8)-[LIVMP]-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;
//
```

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



[Home](#) [ScanProsite](#) [ProRule](#) [Documents](#) [Downloads](#) [Links](#) [Funding](#)

Database of protein domains, families and functional sites

PROSITE consists of [documentation entries](#) describing protein domains, families and functional sites as well as associated [patterns](#) and [profiles](#) to identify them [[More details](#) / [References](#) / [Disclaimer](#) / [Commercial users](#)]. PROSITE is complemented by [ProRule](#), a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [[More details](#)].

Release 20.56, of 16-Oct-2009 (1561 documentation entries, 1308 patterns, 867 profiles and 871 ProRule)

PROSITE access

e.g: PDOC00022, PS50089, SH3, zinc finger

add wildcard '*'

Browse:

- [by documentation entry](#)
- [by ProRule description](#)
- [by taxonomic scope](#)
- [by number of positive hit](#)

PROSITE tools

Scan a sequence against PROSITE patterns and profiles - quick scan

(Output includes graphical view and feature detection)



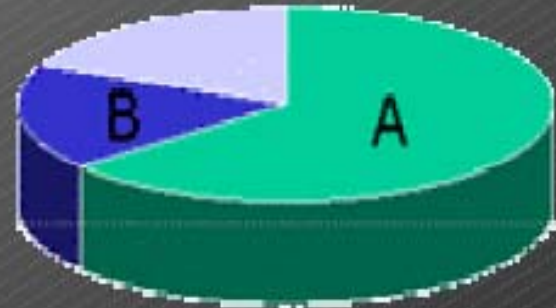
- [ScanProsite](#) - advanced scan
- [PRATT](#) - allows to interactively generate conserved patterns from a series of unaligned proteins.
- [MyDomains - Image Creator](#) ^{new} - allows to generate custom domain figures.



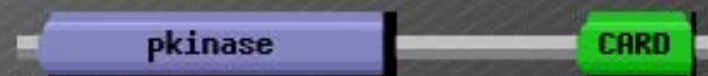
Pfam

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

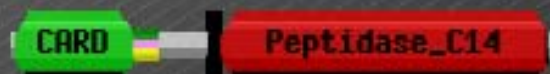
- caracterizan dominios de proteínas con perfiles HMM.
- gran cantidad de información.
- alta cobertura (7.316 familias, 73% swiss-prot y TrEMBL)



Rick:



Caspasa 9:



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

-Interfaz web muy útil:

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

http://pfam.sanger.ac.uk/

Pfam 24.0 (October 2009, 11912 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

QUICK LINKS YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

- SEQUENCE SEARCH** Analyze your protein sequence for Pfam matches
- VIEW A PFAM FAMILY** View Pfam family annotation and alignments
- VIEW A CLAN** See groups of related families
- VIEW A SEQUENCE** Look at the domain organisation of a protein sequence
- VIEW A STRUCTURE** Find the domains on a PDB structure
- KEYWORD SEARCH** Query Pfam by keywords

JUMP TO

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the [help](#) pages for more information

InterPro (1)

Interpro:

<http://www.ebi.ac.uk/Interpro/>

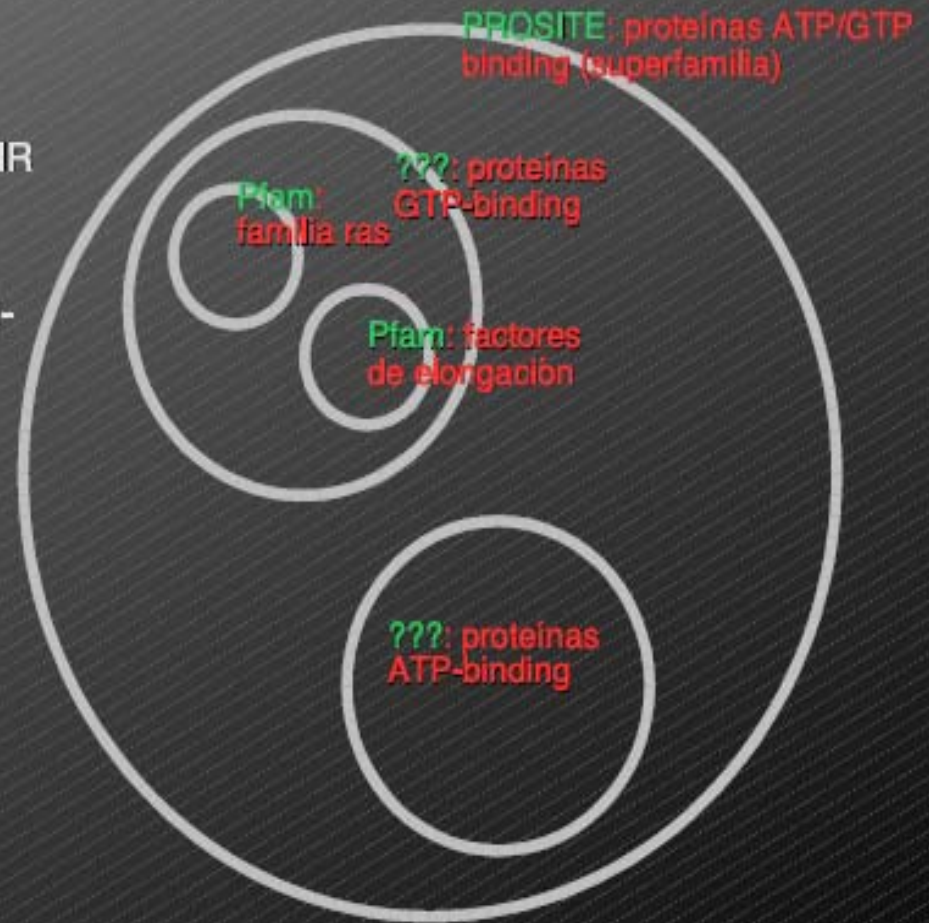
-para poner un poco de orden en el maremagnum de las bases de datos: PROSITE, Pfam, Prints, PRODOM, Smart, PIR

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

-introduce jerarquía

-gran cantidad de información.

-alta cobertura.



- InterPro:Home
 - Advanced Search
 - InterProScan
 - InterPro BioMart
 - Databases
 - Documentation
 - Release Notes
 - User Manual
 - BioMart Manual
 - FAQ
 - Tutorial
 - Example Entry
 - Project Outline
 - People
 - Database
 - Contributors
 - Publications
 - Web Services
 - FTP site
 - Protein Focus
 - Collagen

EBI > Databases > InterPro

Search InterPro: >

InterPro: Home

InterPro is a database of protein families, domains, regions, repeats and sites in which identifiable features found in known proteins can be applied to new protein sequences.

Release News

We are pleased to announce the release of InterPro 23.0. It contains 19150 entries and 434 new methods, which include new methods from PANTHER, HAMAP and the new releases from PROSITE and SMART and covers 76.4% of UniProt Knowledgebase release 15.8.

For full details of the release please see [Release Notes](#).

The InterPro BioMart

The [InterPro BioMart](#) allows you to retrieve InterPro data from a query-optimised data warehouse that is synchronised with the main InterPro database shortly after each InterPro release.

The BioMart interface allows you to build simple or complex queries, with control over:

- how the data is filtered, to restrict which records are included
- the attributes (columns) that are included in the results



For further details, please see the [InterPro BioMart Manual](#).

User support and feedback

We welcome feedback, particularly if you find errors or omissions please let us know. If you need information or help, have any comments and/or suggestions on the InterPro database, please contact us at [EBI Support](#).

InterPro Funding



- [InterPro home](#)
- [Text Search](#)
- [InterProScan](#)
- [Databases](#)
- [Documentation](#)
- [FTP Site](#)

■ **InterProScan Help**

- [Help](#)
- [FAQ](#)
- [README](#)

■ [InterProScan Programmatic Access](#)

■ **Database Information**

- [UniProt](#)
- [UniParc](#)

InterProScan Sequence Search

This form allows you to query your sequence against InterPro. For more detailed information see the documentation for the perl stand-alone InterProScan package ([Readme file](#) or [FAQ's](#)), or the InterPro [user manual](#) or [help pages](#).

Please Note: PatternScan is a new version of the PROSITE pattern search software which uses new code developed by the PROSITE team. The ScanRegExp program was internally developed by the InterPro team to be equivalent to the PROSITE code and depends on data which is no longer generated (confirm.patterns from Emotif). It was therefore deemed necessary to move over to using the same program as PROSITE, ps_scan.pl which uses evaluator mini-profiles to confirm whether or not a match is true positive. The outcome of this is a more sensitive predictor of True matches and an effective increase in the coverage of True PROSITE matches. ScanRegExp program will be phased out of the External Services InterProScan by the end of 2008 and the data will no longer be provided in the InterPro data updates from our FTP site. [Help](#) for more information.

Please Note: Due to resource limitations the InterProScan service will not accept nucleotide sequence submissions until further notice. Please see the [Help](#) for more information.

 [Download Software](#)

RESULTS YOUR EMAIL

APPLICATIONS TO RUN Clear all Check all

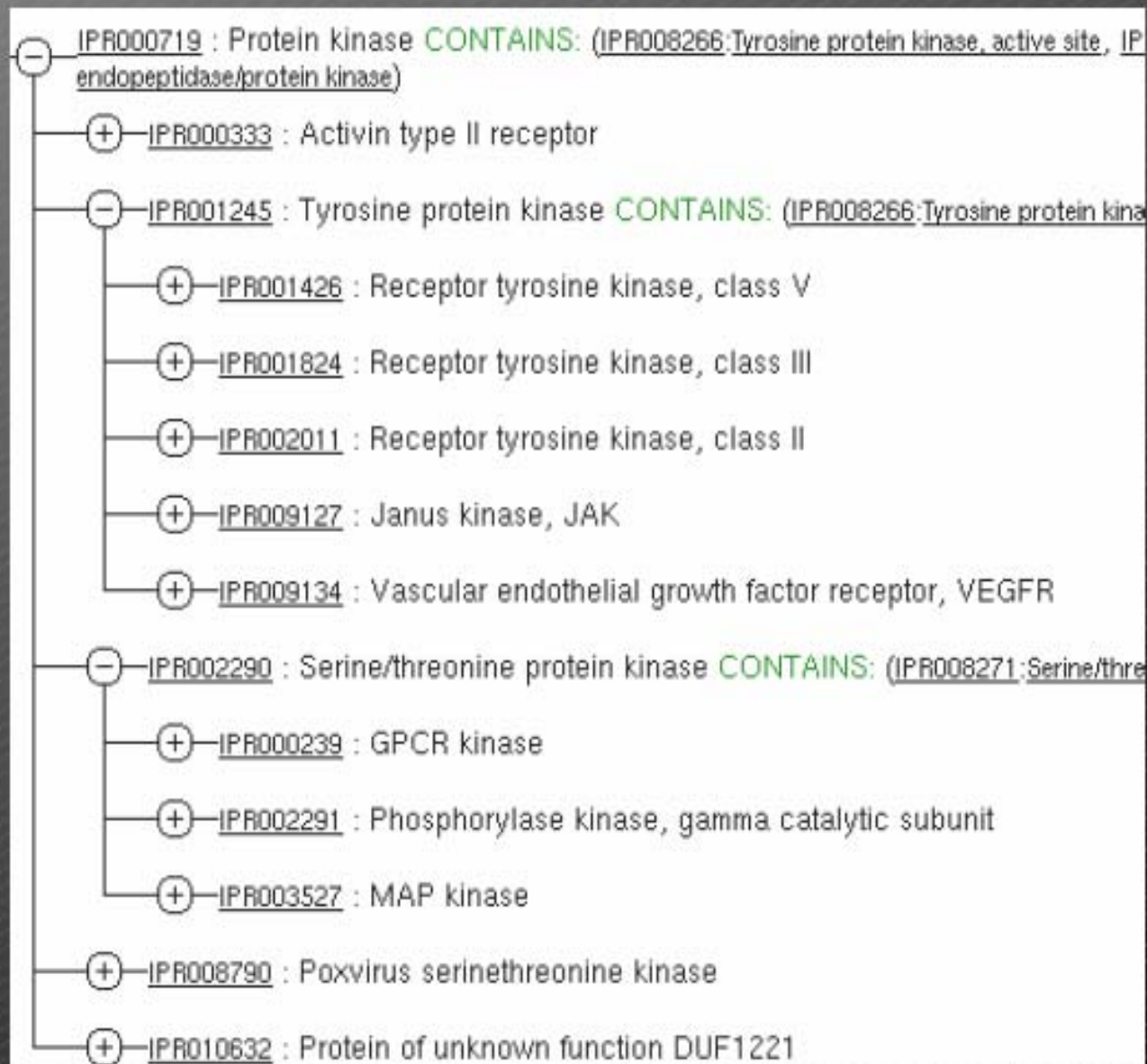
| | | | | |
|---|---|--|---|---|
| <input checked="" type="checkbox"/> BlastProDom | <input checked="" type="checkbox"/> FPrintScan | <input checked="" type="checkbox"/> HMMPIR | <input checked="" type="checkbox"/> HMMPfam | <input checked="" type="checkbox"/> HMMSmart |
| <input checked="" type="checkbox"/> HMMTigr | <input checked="" type="checkbox"/> ProfileScan | <input checked="" type="checkbox"/> ScanRegExp | <input type="checkbox"/> patternScan | <input checked="" type="checkbox"/> SuperFamily |
| <input checked="" type="checkbox"/> TMHMM | <input checked="" type="checkbox"/> HMMPanther | <input checked="" type="checkbox"/> Gene3D | | <input checked="" type="checkbox"/> SignalPHMM |

Enter or Paste a PROTEIN Sequence in any format:

Help

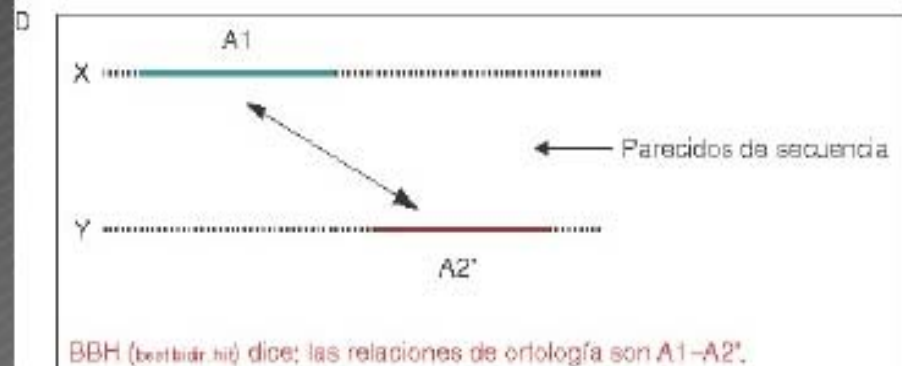
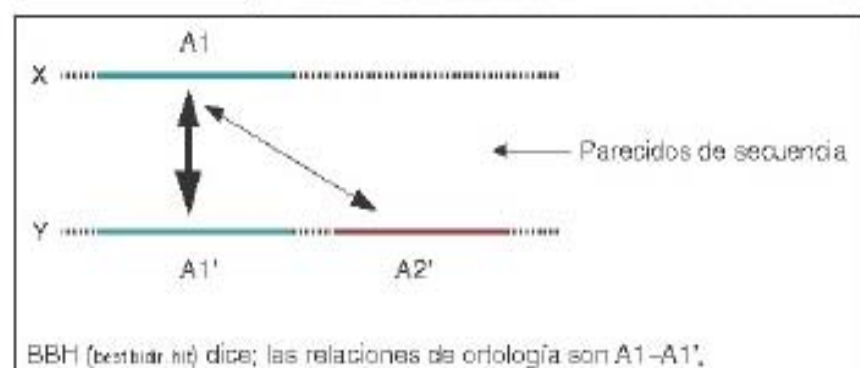
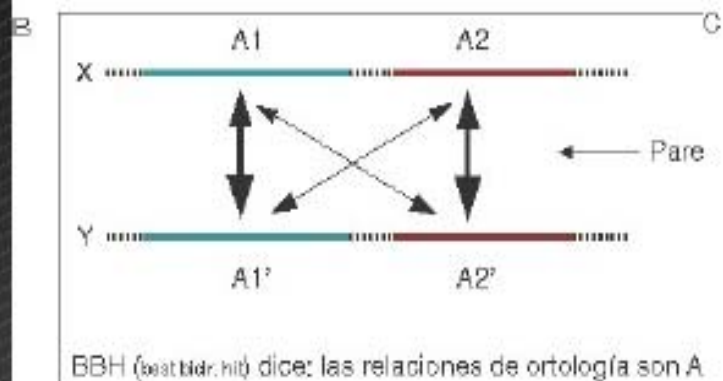
La jerarquía en InterPro:

ejemplo de las kinasas de proteínas.



COGs: clasificación en grupos de ortólogos

Identificación de ortólogos basada en "Best Bidirectional Hits"



El BBH sólo es aplicable con genomas completos.

COGs: clasificación en grupos de ortólogos

¿Qué se puede hacer con COGs?

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

Versión previa de COGs: 44 genomas de microorganismos

Actualmente: 66 genomas de microorganismos y 7 de eucariotas



COGs

Phylogenetic classification of proteins encoded in complete genomes



Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

66 genomes
38 orders
28 classes
14 phyla

Unicellular clusters FTP

[Science 1997 Oct 24;278\(5338\):631-7,](#)
[BMC Bioinformatics 2003 Sep 11;4\(1\):41.](#)

[Initial](#)

[version](#)

| | | |
|--|---|--|
| <u>Euryarchaeota</u> | <u>Aquificae</u> | <u>Actinobacteria</u> |
| <u>Methanobacteriales</u> <u>Mth</u> | <u>Aquificales</u> <u>Aae</u> | <u>Actinomycetales</u> <u>Cgl</u> <u>Mtu</u> <u>MtC</u> <u>Mle</u> |
| <u>Methanococcales</u> <u>Mja</u> | <u>Thermotogae</u> | <u>Firmicutes</u> |
| <u>Halobacteriales</u> <u>Hbs</u> | <u>Thermotogales</u> <u>Tma</u> | <u>Clostridiales</u> <u>Cac</u> |
| <u>Thermoplasmatales</u> <u>Tac</u> <u>Tvo</u> | <u>Cyanobacteria</u> | <u>Bacillales</u> <u>Sau</u> <u>Lin</u> <u>Bsu</u> <u>Bha</u> |
| <u>Thermococcales</u> <u>Pho</u> <u>Pab</u> | <u>Nostocales</u> <u>Nos</u> | <u>Lactobacillales</u> <u>Lla</u> <u>Spy</u> <u>Spn</u> |
| <u>Archaeoglobales</u> <u>Afu</u> | <u>Chroococcales</u> <u>Syn</u> | <u>Mycoplasmatales</u> <u>Uur</u> <u>Mpu</u> <u>Mpn</u> <u>Mge</u> |
| <u>Methanopyrales</u> <u>Mka</u> | <u>Deinococcus-Thermus</u> | <u>Proteobacteria</u> |
| <u>Methanosarcinales</u> <u>Mac</u> | <u>Deinococcales</u> <u>Dra</u> | <u>Pseudomonadales</u> <u>Pae</u> |
| <u>Crenarchaeota</u> | <u>Fusobacteria</u> | <u>Enterobacteriales</u> <u>Eco</u> <u>EcZ</u> <u>Ecs</u> <u>Ype</u> <u>Sty</u> <u>Buc</u> |
| <u>Thermoproteales</u> <u>Pya</u> | <u>Fusobacteriales</u> <u>Fnu</u> | <u>Xanthomonadales</u> <u>Xfa</u> |
| <u>Sulfolobales</u> <u>Sso</u> | | <u>Vibrionales</u> <u>Vch</u> |
| <u>Desulfurococcales</u> <u>Ape</u> | | |

Eukaryotic Clusters FTP

| Code | Name | Abbreviation |
|----------|---|--------------|
| A | <i>Arabidopsis thaliana</i> (thale cress) | <i>ath</i> |
| C | <i>Caenorhabditis elegans</i> (worm) | <i>cel</i> |
| D | <i>Drosophila melanogaster</i> (fruit fly) | <i>dme</i> |
| H | <i>Homo sapiens</i> (human) | <i>hsa</i> |
| Y | <i>Saccharomyces cerevisiae</i> (baker yeast) | <i>sce</i> |
| P | <i>Schizosaccharomyces pombe</i> (fission yeast) | <i>spo</i> |
| E | <i>Encephalitozoon cuniculi</i> | <i>ecu</i> |