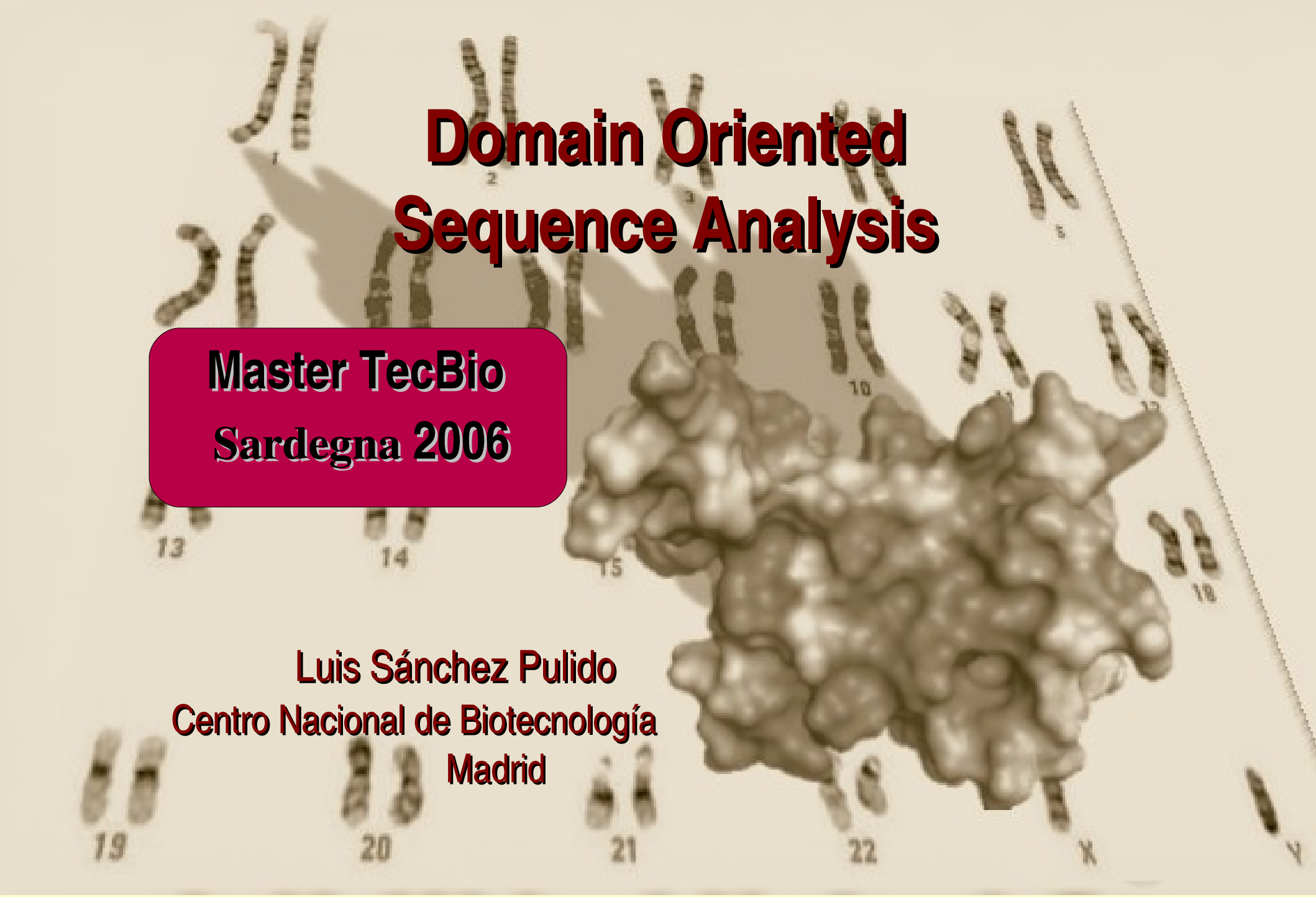
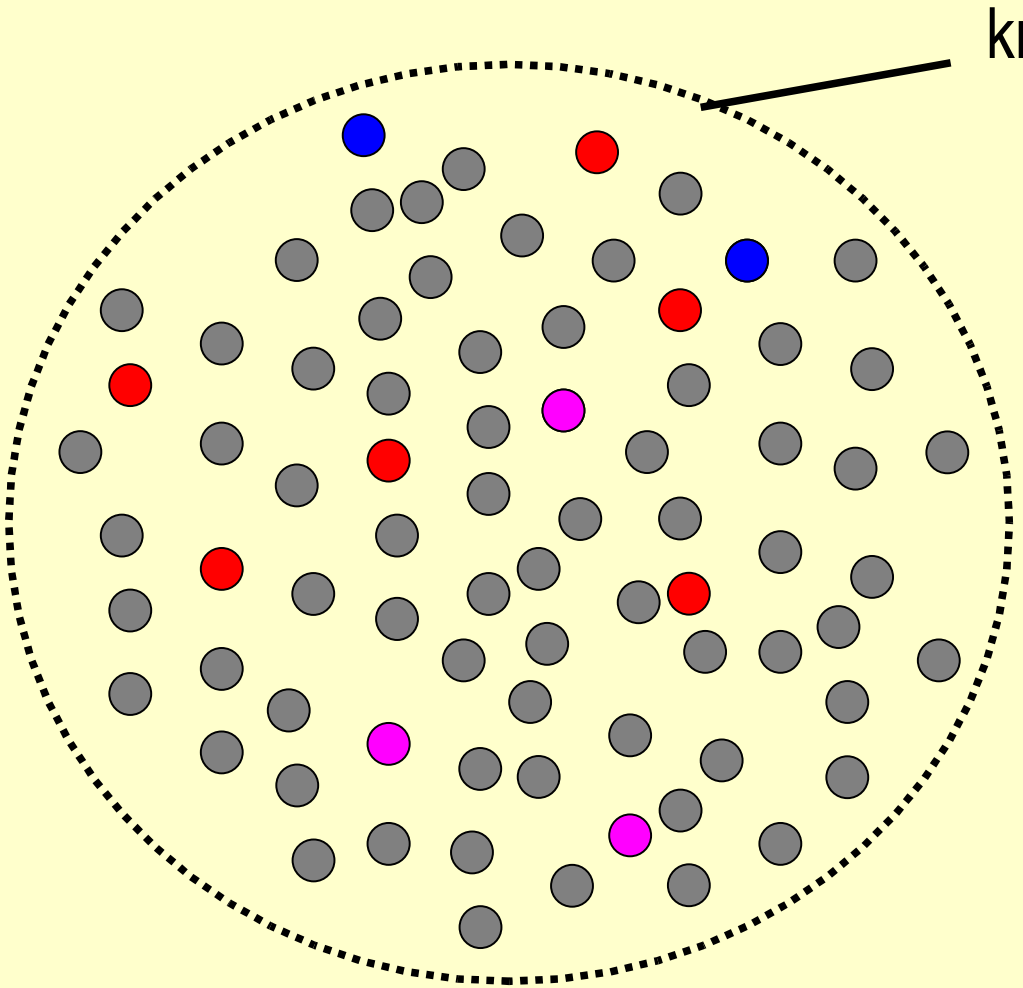
Domain Oriented Sequence Analysis

Master TecBio
Sardegna 2006

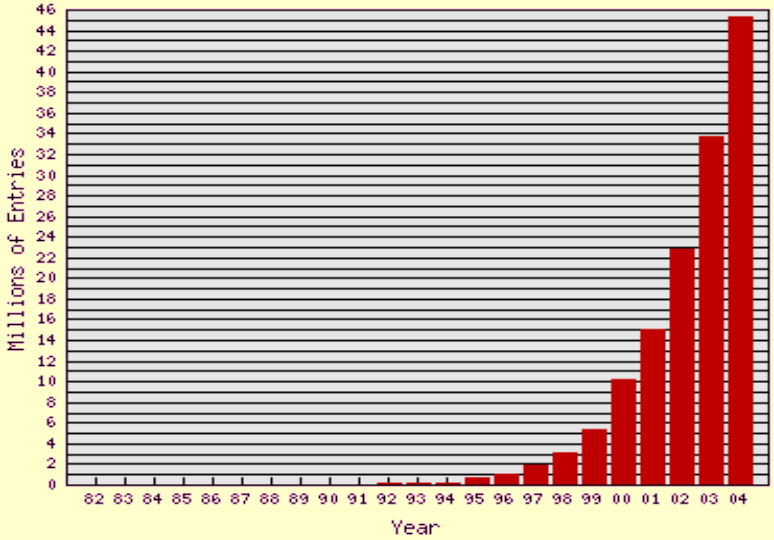
Luis Sánchez Pulido
Centro Nacional de Biotecnología
Madrid

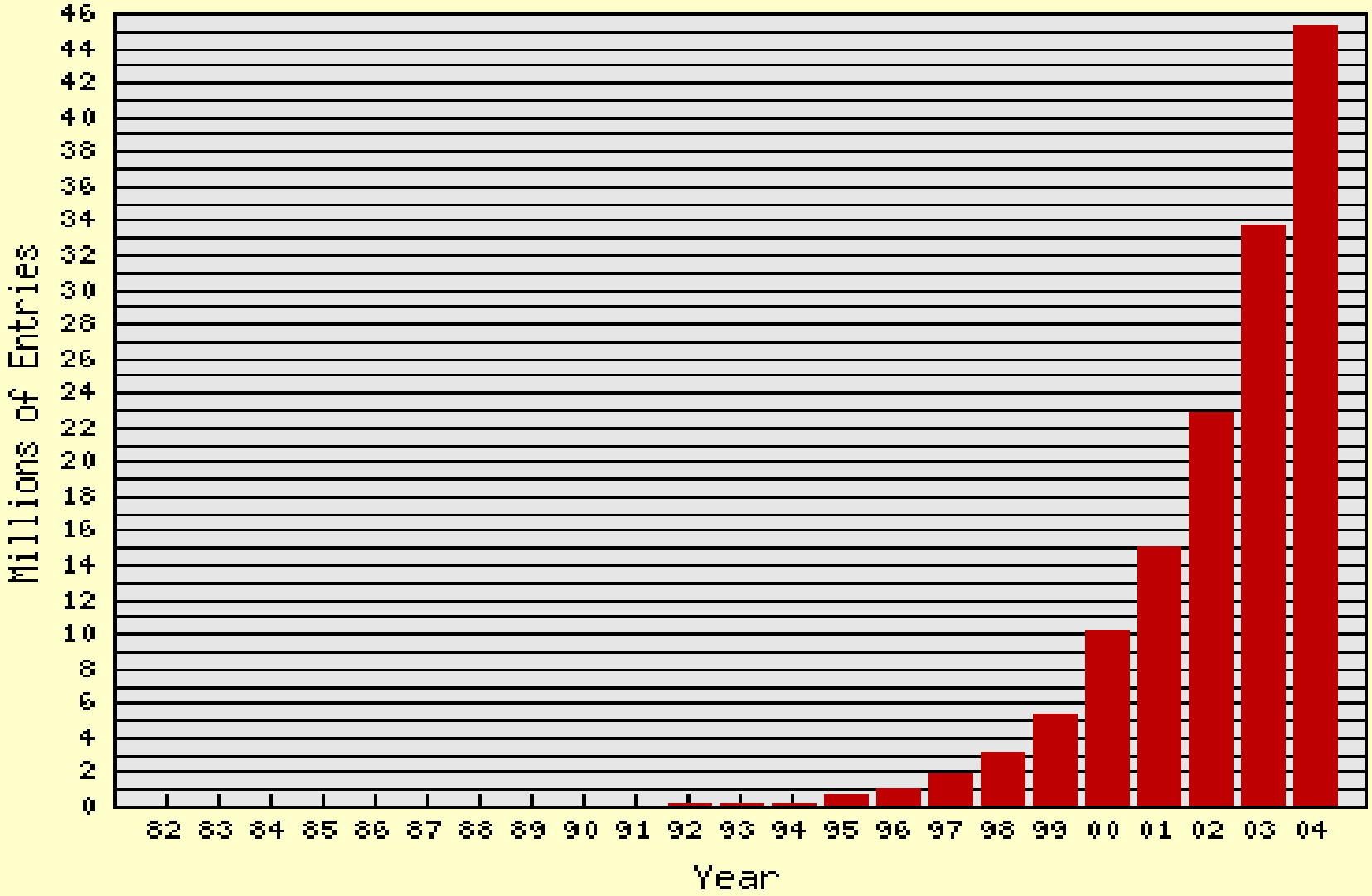


Why to do Protein Sequence Analysis?



- 3D
- Function
- Both
- ?????





Exponential Growth of sequence Databases

source www3.ebi.ac.uk/Services/DBStats/

Grazie all'identificazione del homología fra le proteine, possiamo

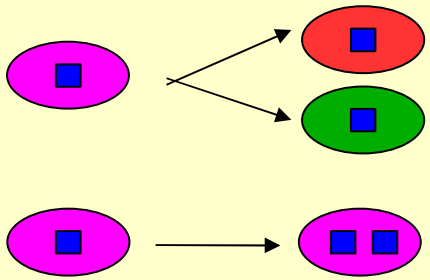
TRASFERIRE LE INFORMAZIONI

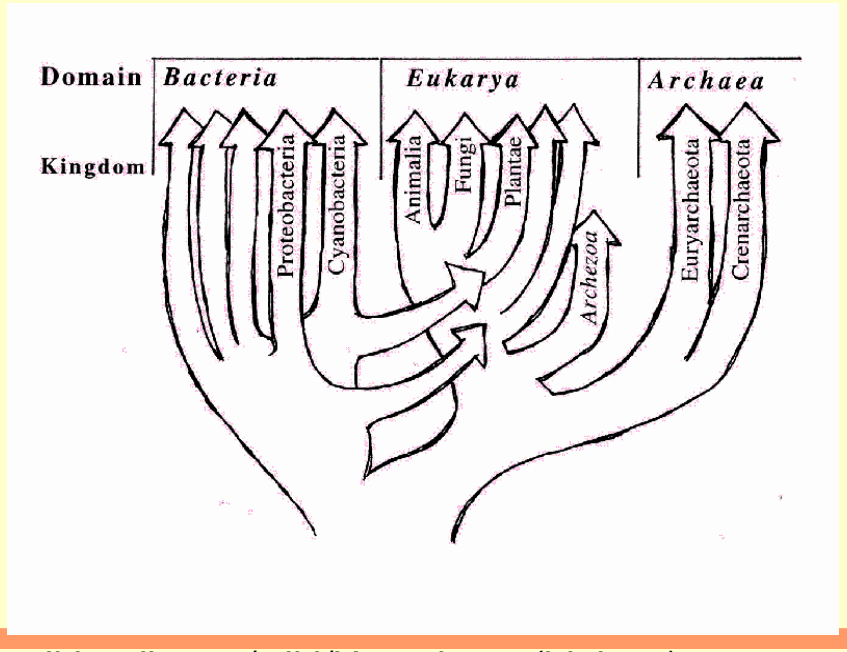
strutturali eo funzionale

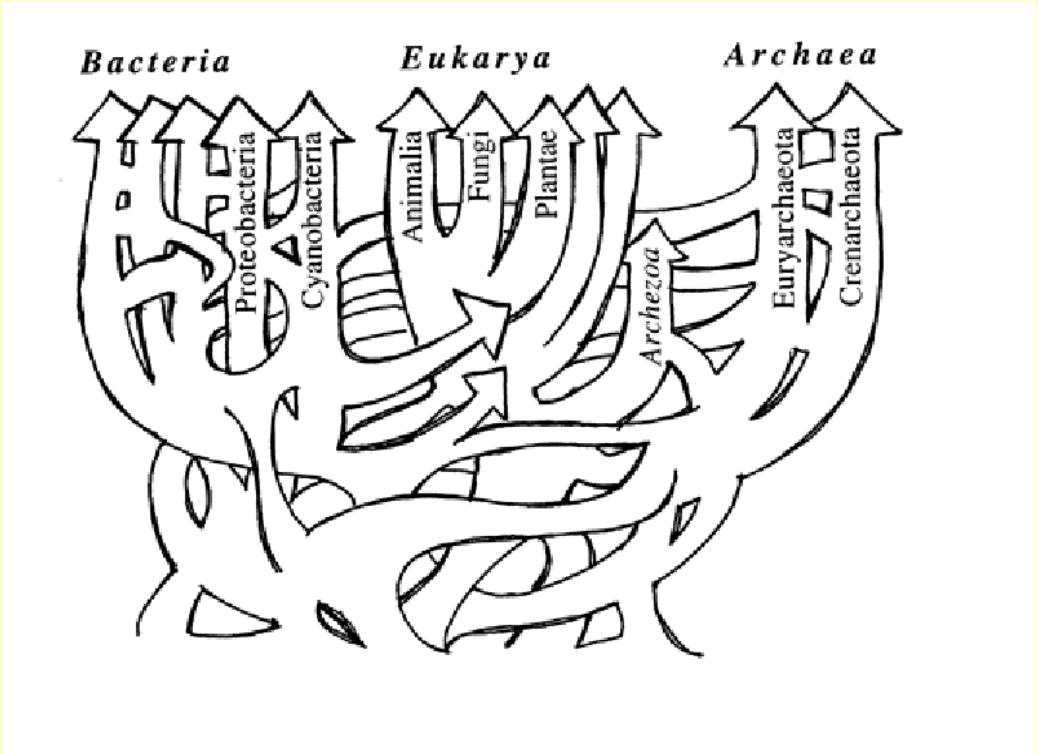
Homologous: pair of proteins with a common ancestor.

...and depending on the origin of their divergence:

- **orthologs** - speciation
- **paralogs** – gene duplication
- **xenologs** – horizontal transfer





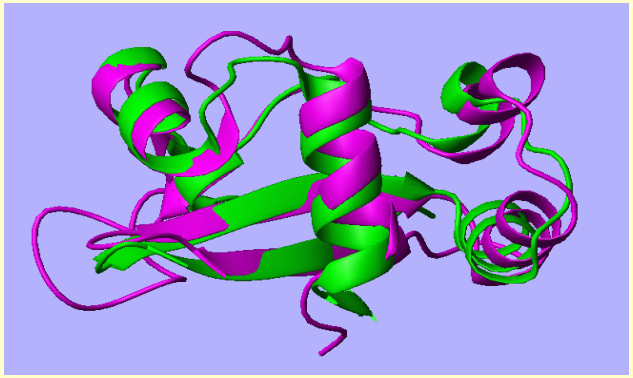


http://en.wikipedia.org/wiki/Homology_(biology)

TRANSFERIRE LE INFORMAZIONI

•*Structural*

From homologous proteins of known structure (by X-Ray, NMR or EM)

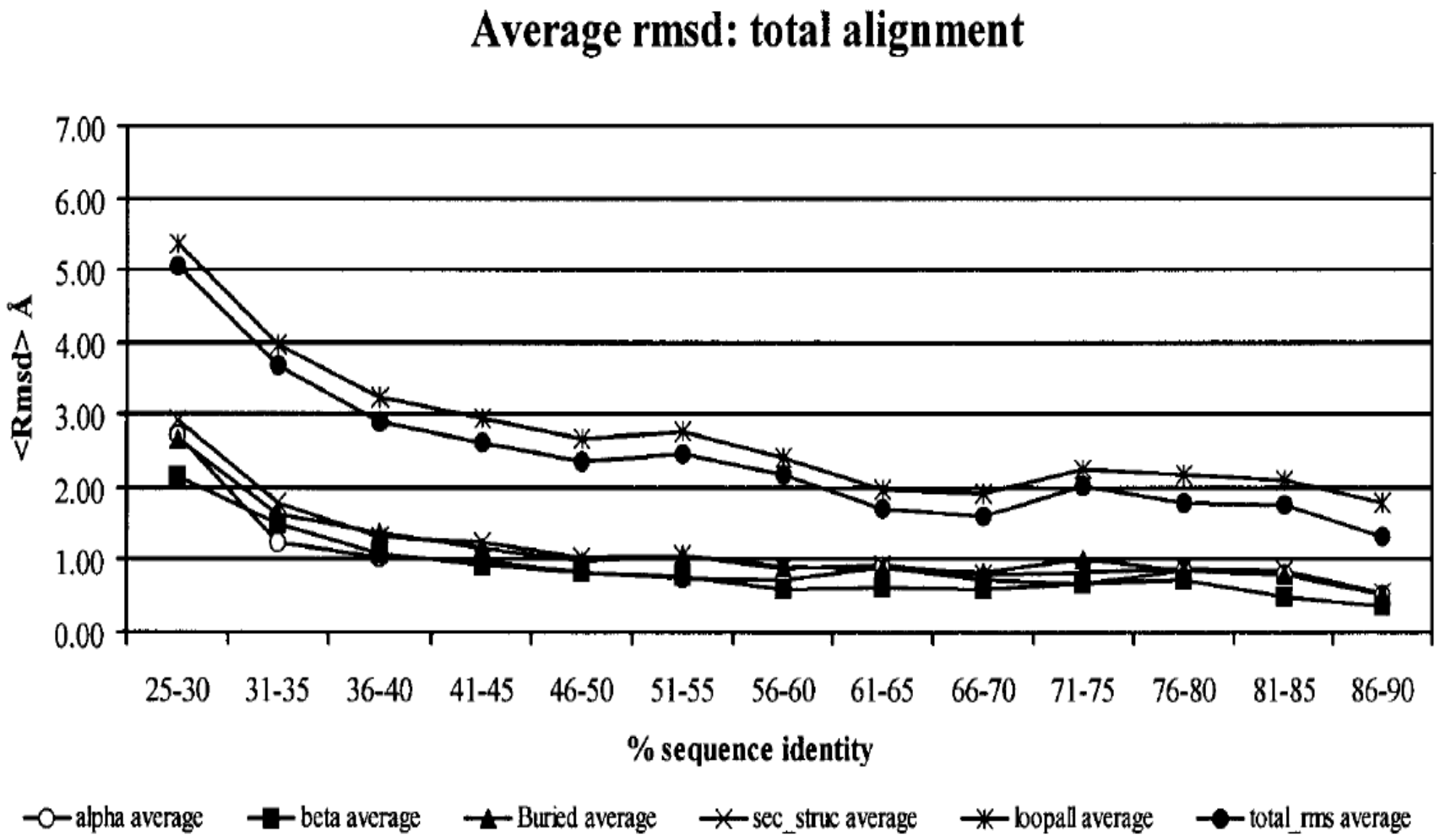


•*Functional*

From homologous proteins of known function... OR

known context... STRING

La struttura si conserva più meglio della sequenza.

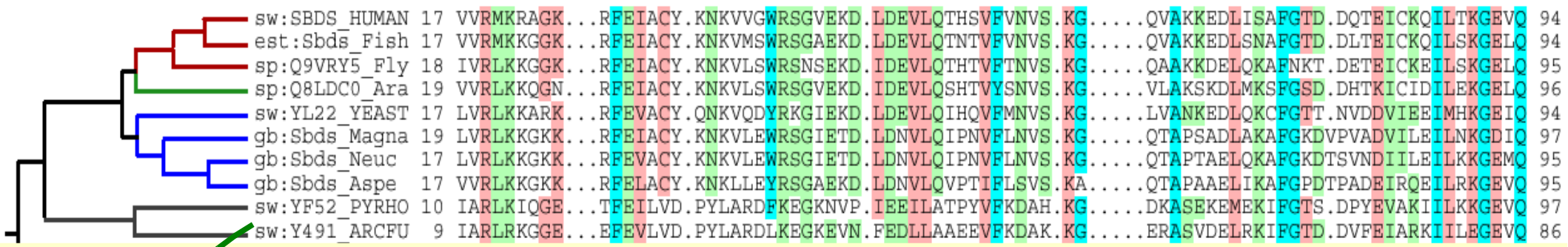


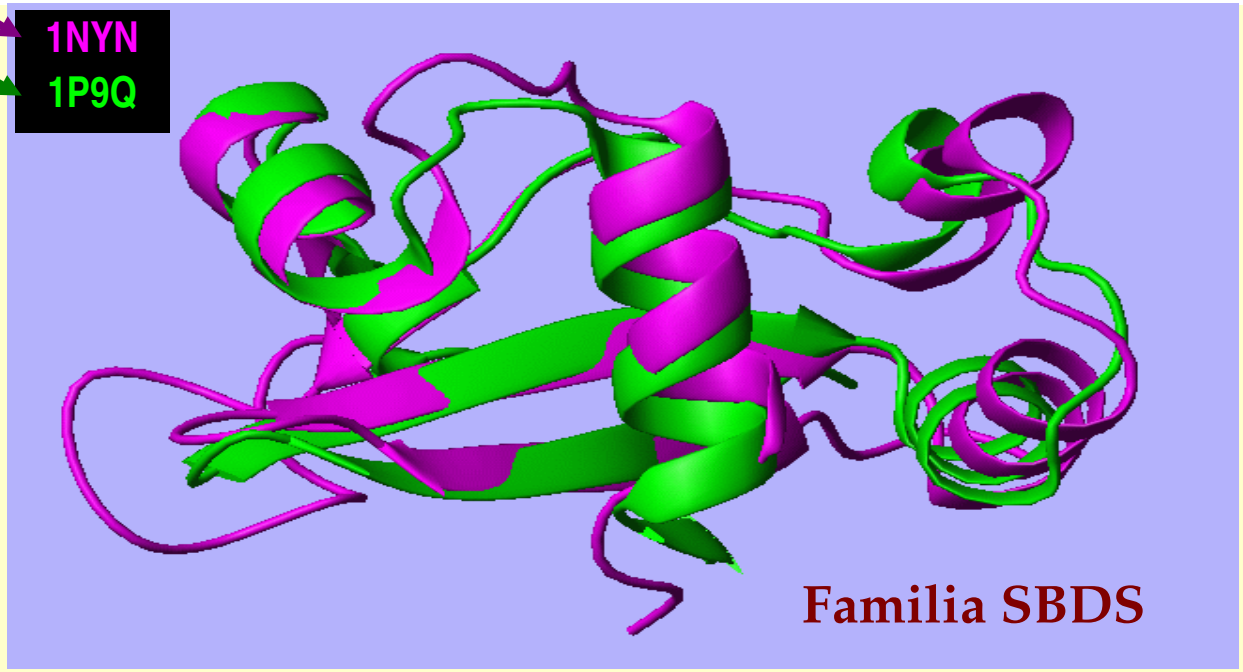
D'Alfonso G, Tramontano A, Lahm A.

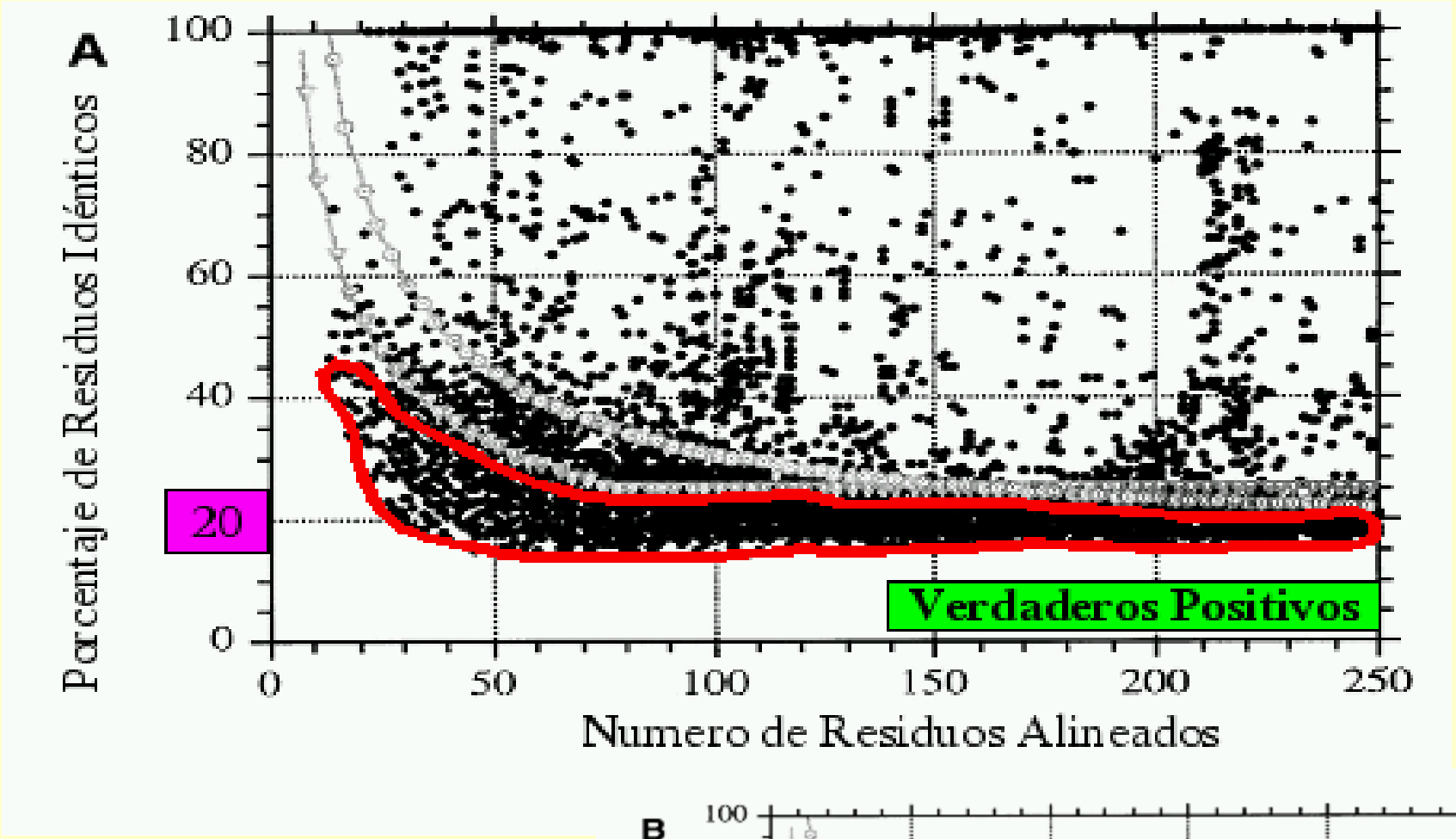
Structural conservation in single-domain proteins: implications for homology modeling.

J Struct Biol. 134, 246-56. (2001)

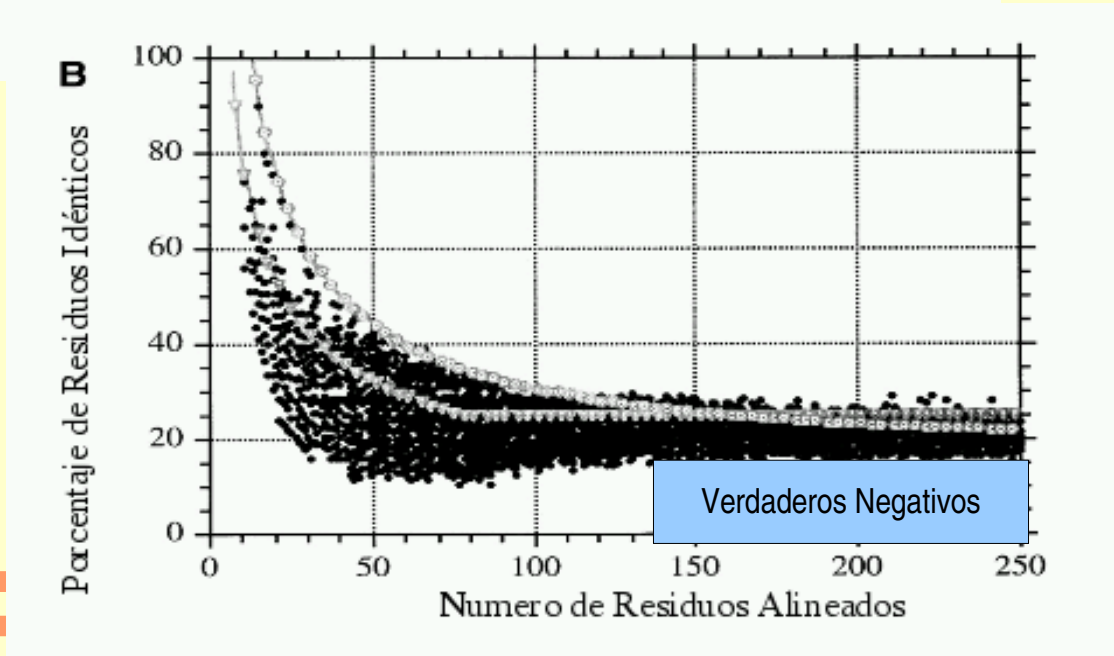
A Remote Homology example:



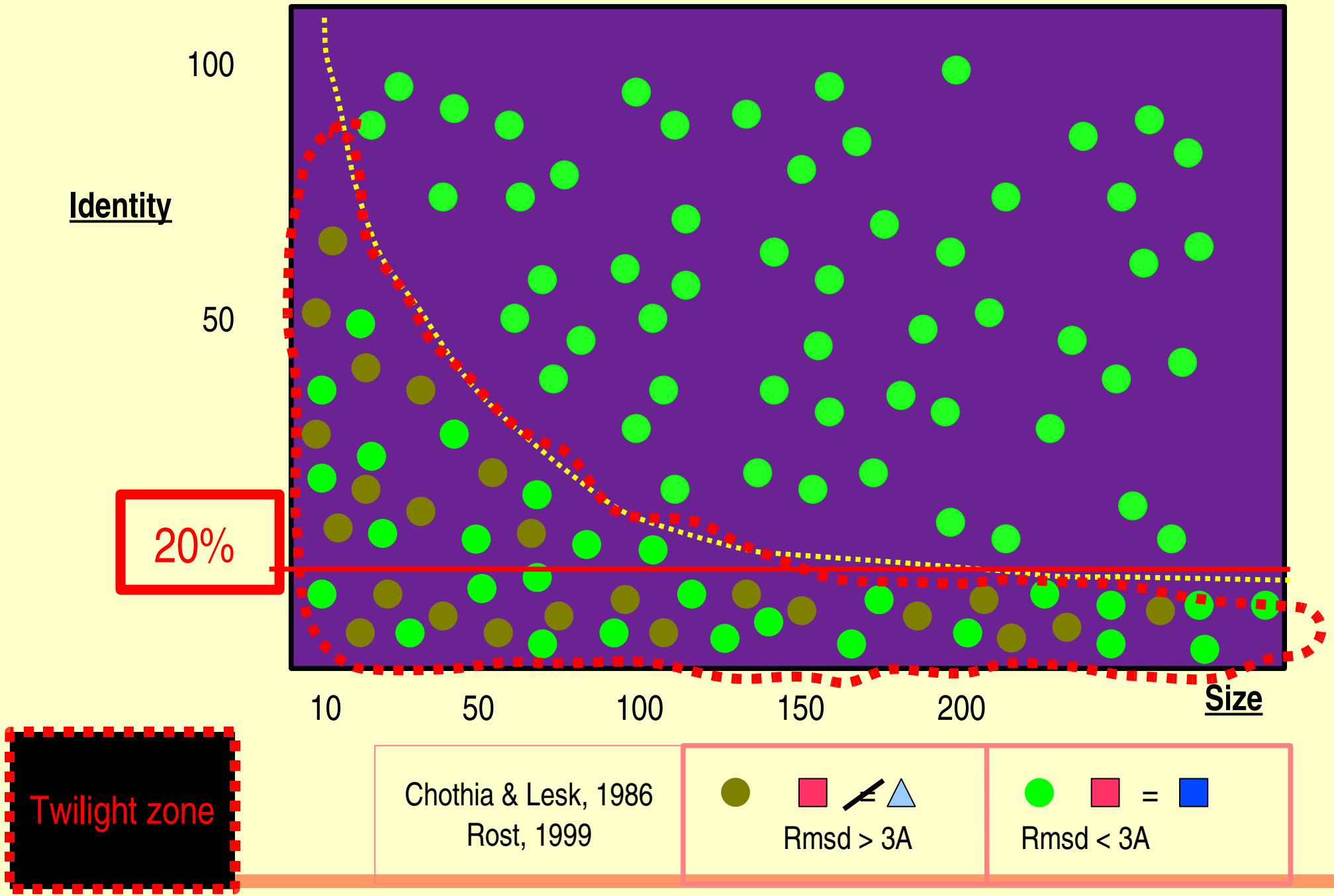




Rost B. (1999)
Twilight zone of
protein sequence alignments.
Protein Eng. 12:85-94.



Comparisons between pairs of sequences with known structure

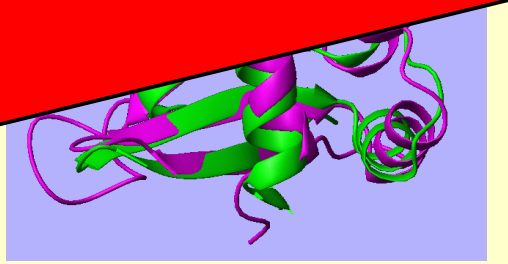


TRANSFERIRE LE INFORMAZIONI

- ***Structural***

From homologous proteins of t

La struttura si conserva più meglio della sequenza.

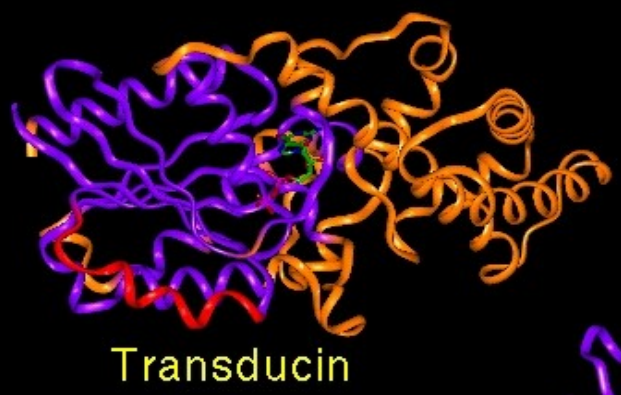


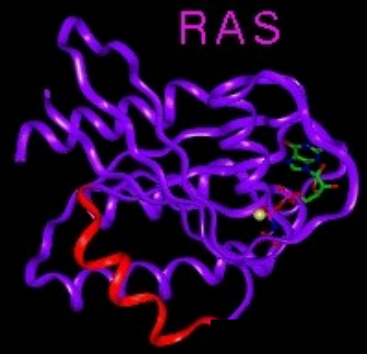
- ***Functional***

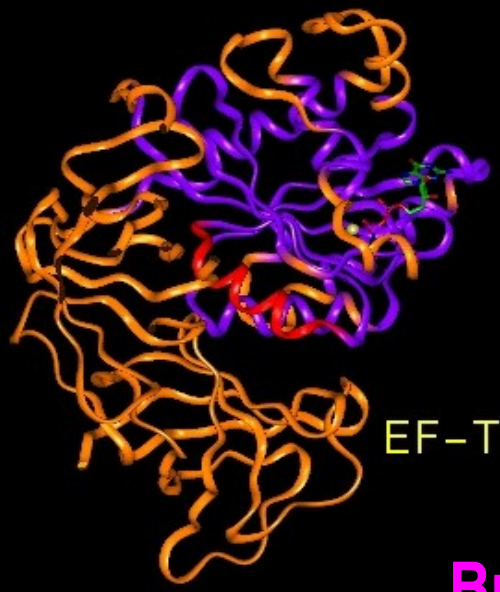
From homologous proteins of known function... OR

known context... STRING

¿FUNCTION?





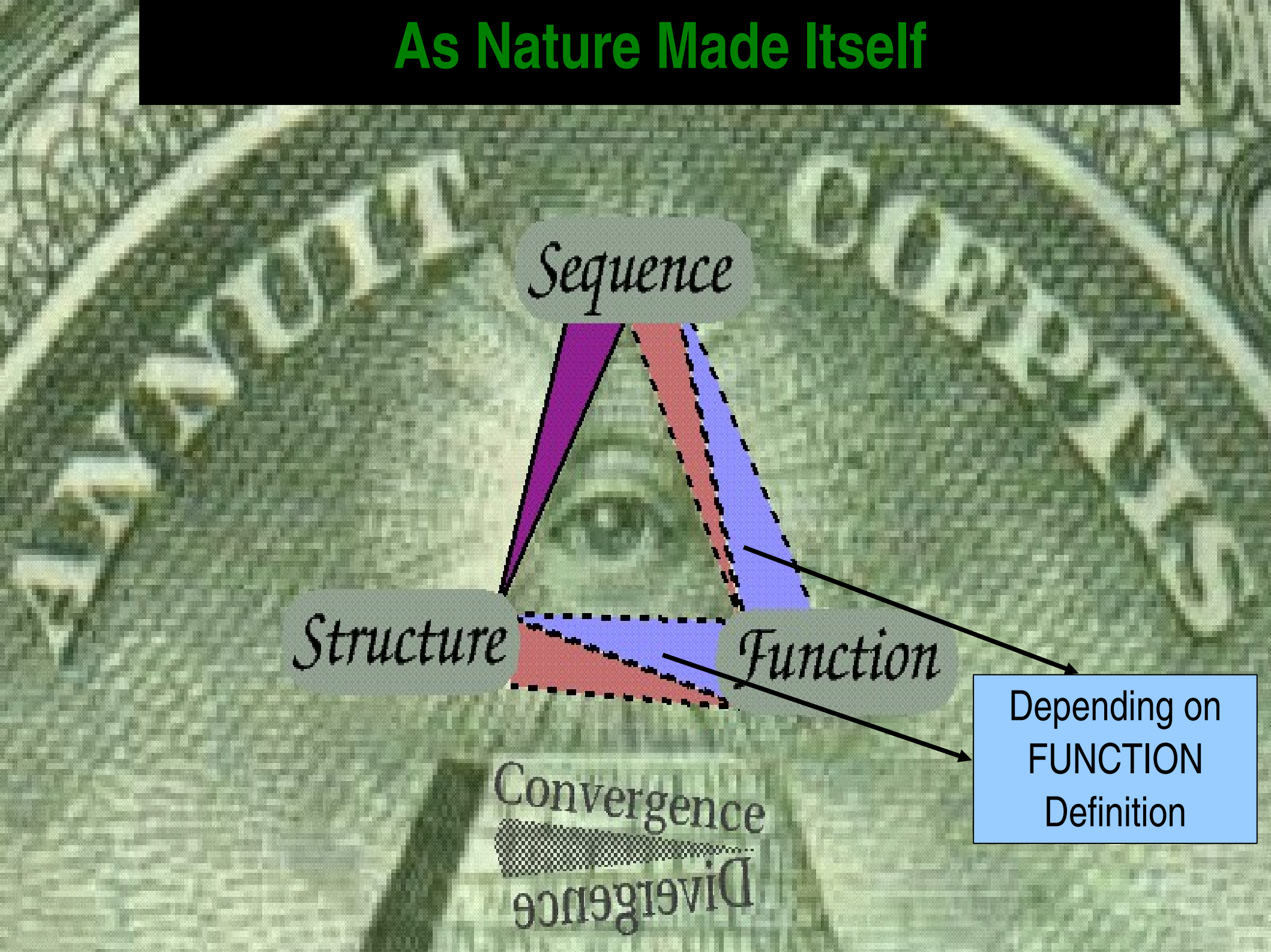


They are homologous Proteins...

The Function could be very divergent

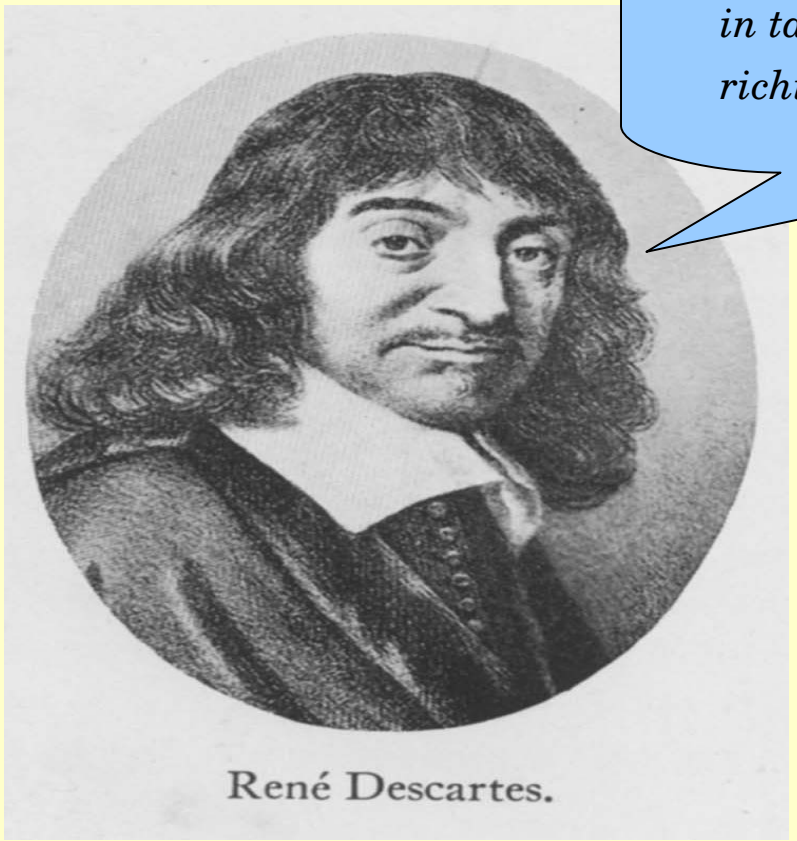
But... All of them bind GTP

As Nature Made Itself



Operazione complessa quella trasferire le informazioni strutturali eo funzionale fra le proteine omologhe.

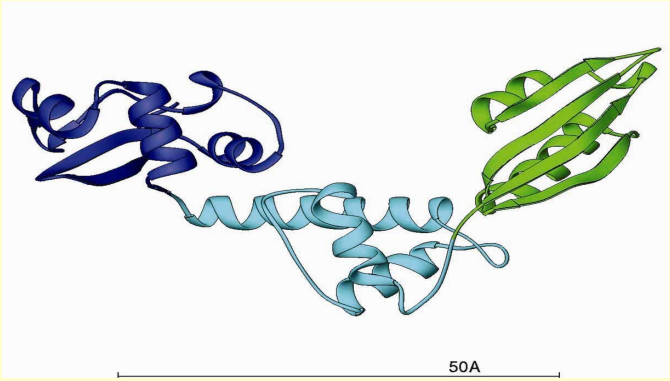
Che possiamo fare???



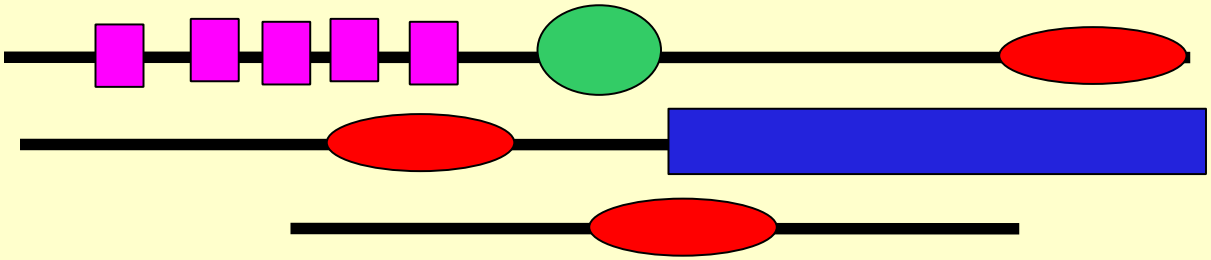
Dividere ogni problema preso in esame in tante parti quanto fosse possibile e richiesto per risolverlo più agevolmente

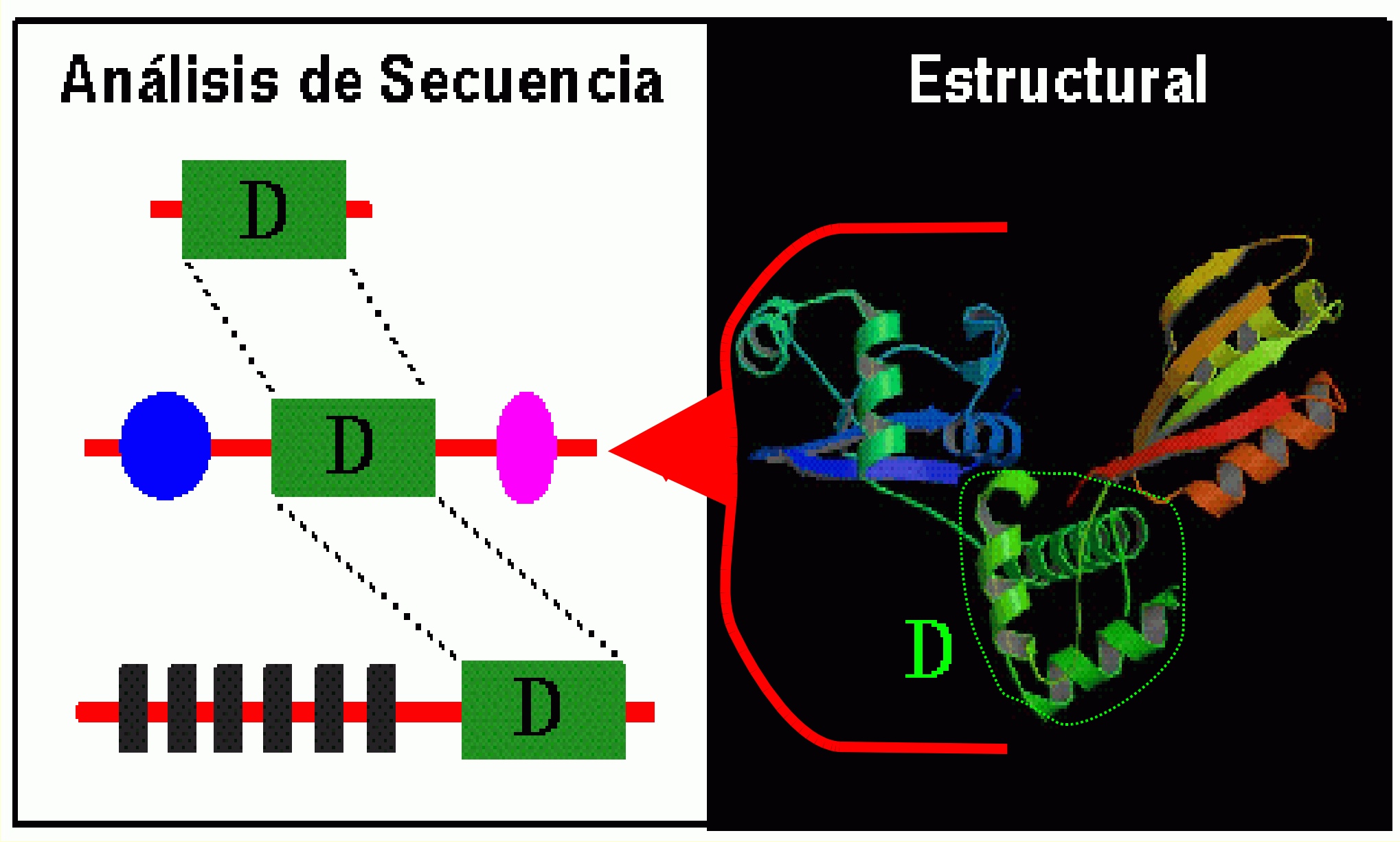
Domain Definition

I dominions della proteina sono stati descritti, da un punto di vista strutturale, come il compatto e localmente le unità strutturali indipendenti, caratterizzato solitamente da un nucleo di hidrofóbico ha definito buon

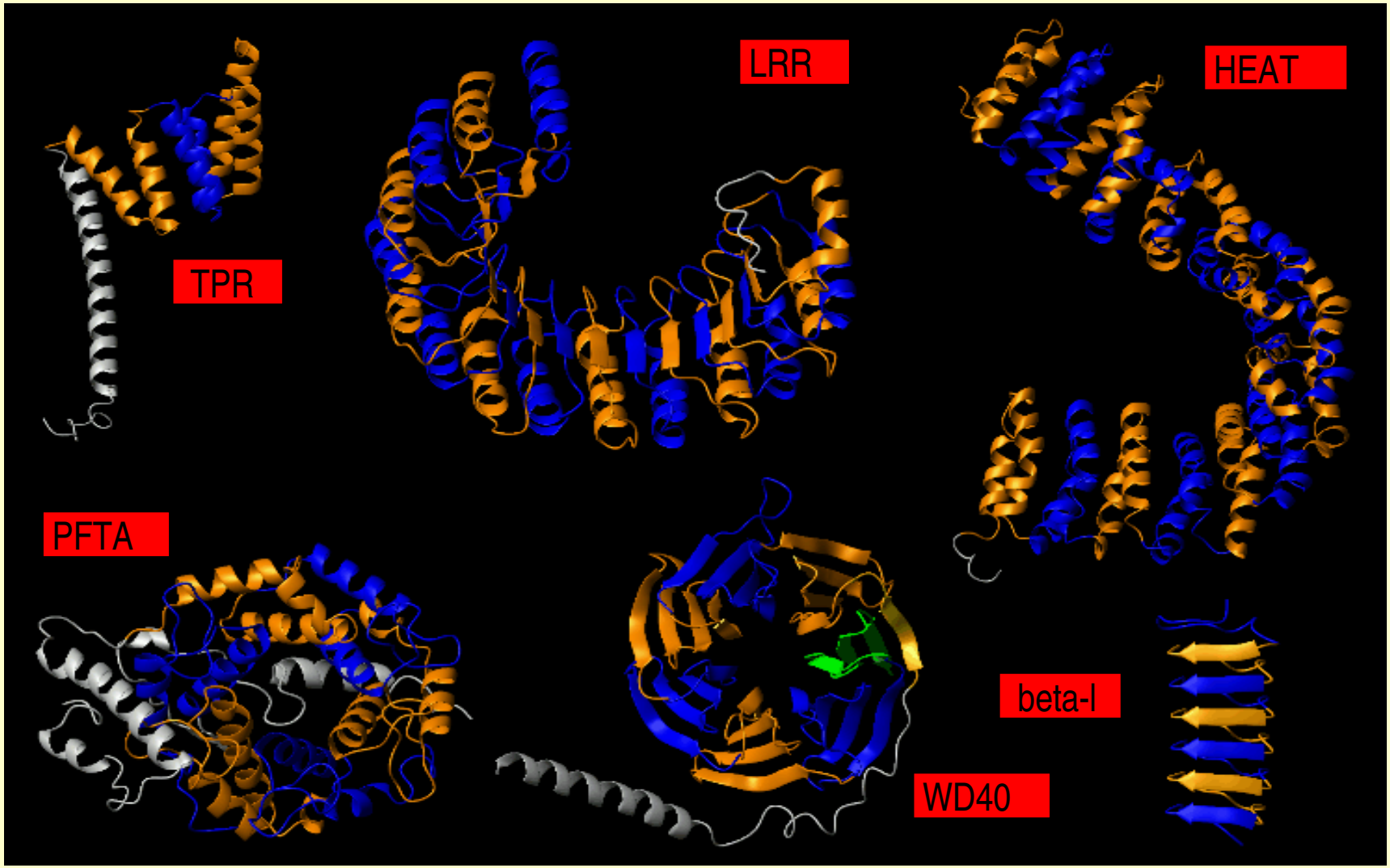


Dal punto di vista dell'analisi di sequenza, i dominions sono definiti come regioni conservate evolutionarily ed acquistano l'attinenza più grande se sono descritti come i moduli mobili, cioè, presenti in famiglie diverse dalle proteine di architettura varia.




Análisis de Secuencia
Estructural
D
D
D
D

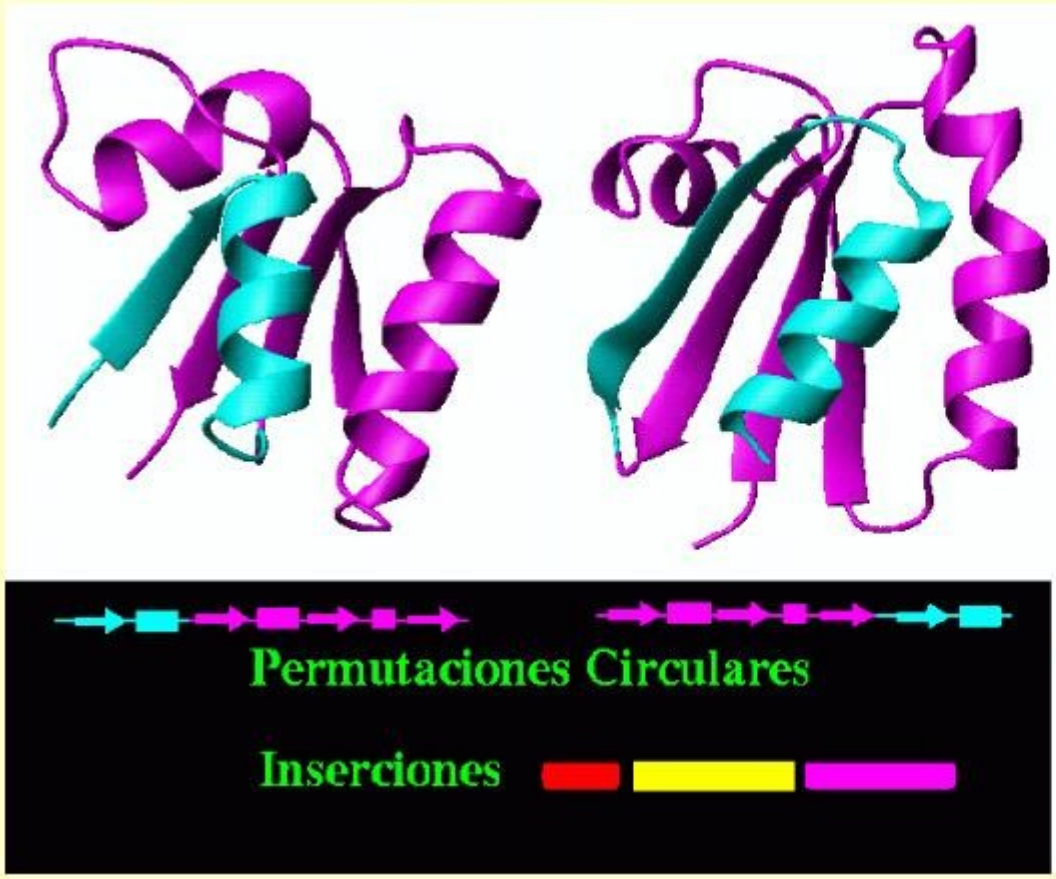
REPEATS – In the limits of Domain Definition

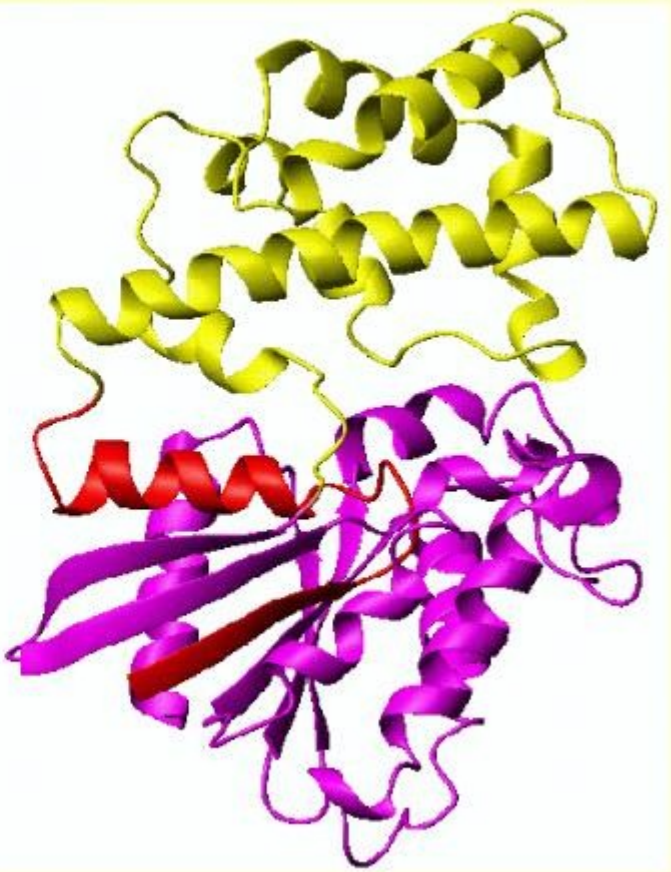


Protein repeats. Short specialist review for the Encyclopedia of Genomics, Proteomics, and Cedida por: Perez-Iratxeta C, Andrade MA (2005) Bioinf. Ed. Wiley and Sons Ltd., UK.

- **Protein irregularities that hinder sequence analysis**
- **Low complexity regions**
- **Repeats, Trans-membrane and Coiled-coil regions (high mutation rates)**
- **and Fold irregularities, such as:**

  Circular Permutations and Insertions

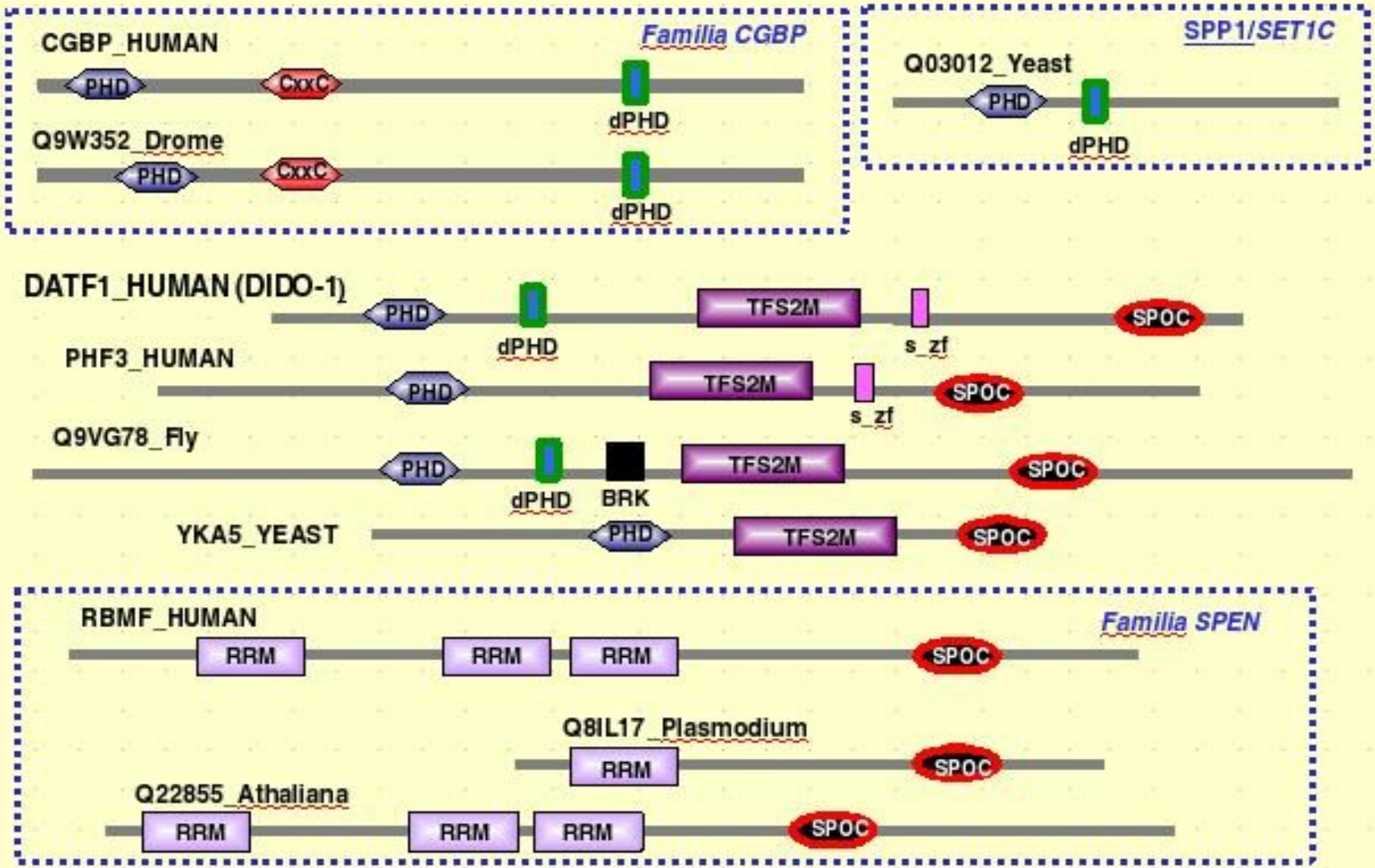




The role of domains in protein evolution

Shuffling, Accretion and Supra-Domains > Christine Vogel *et al.*

(per mescolare)



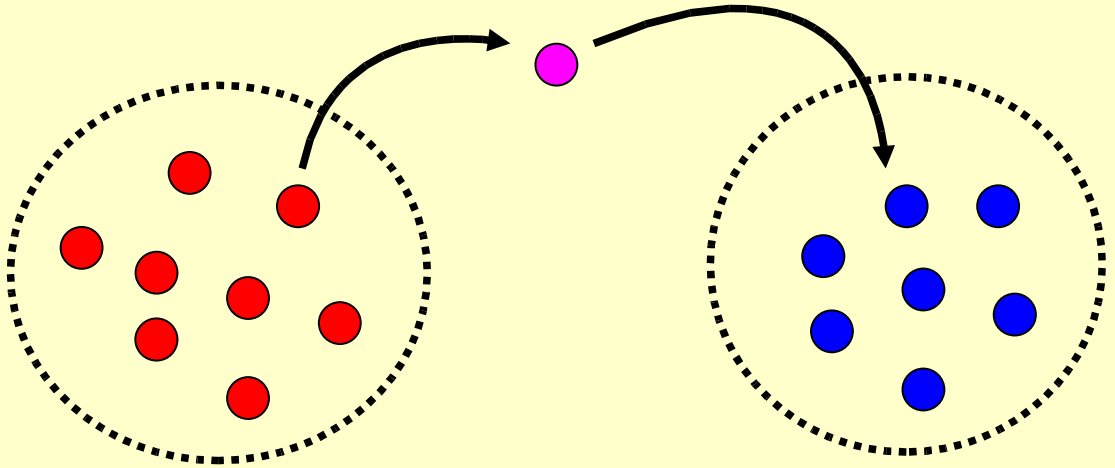
VERSATILITY !!

METHODS
ON
DOMAIN
ORIENTED
SEQUENCE ANALYSIS

Detection of homologous protein sequences

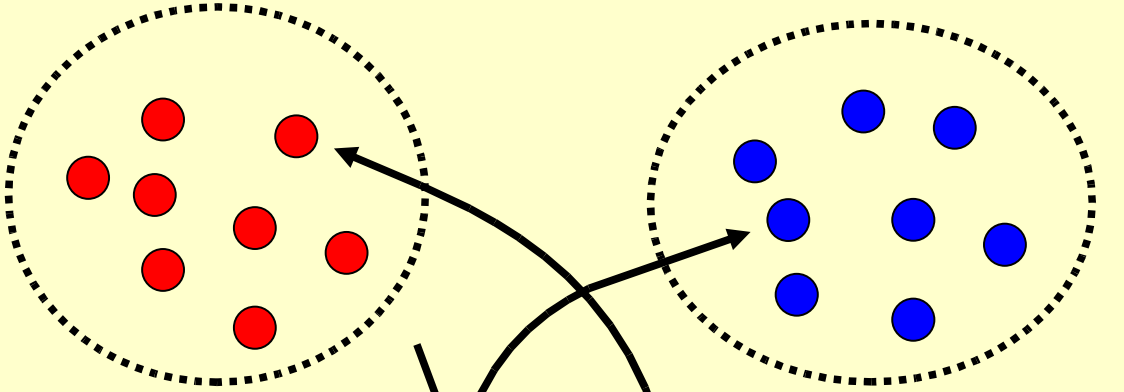
Two strategies

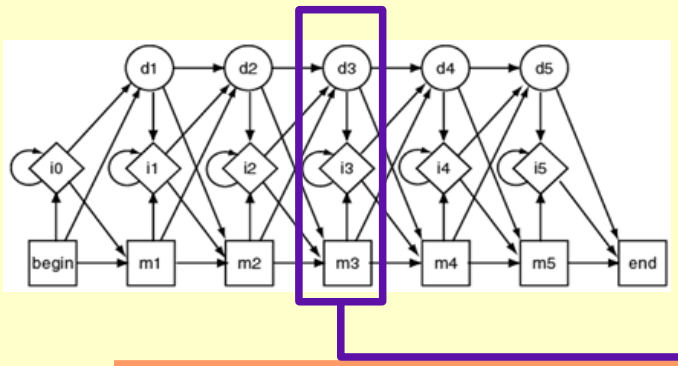
ISS (Blast, FASTA)

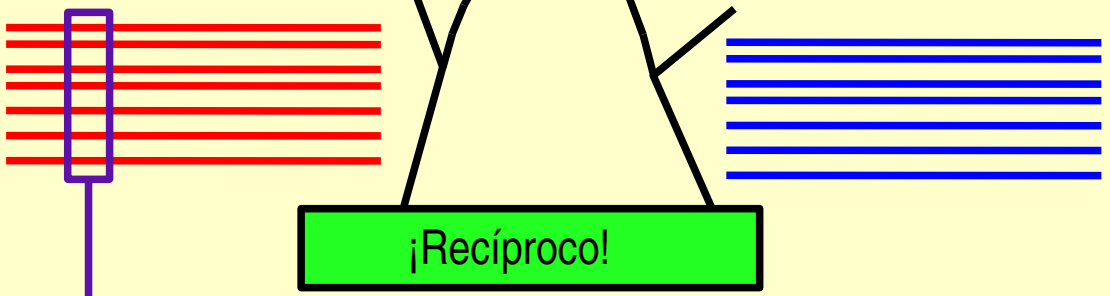


Profiles

(PSSMs: PsiBlast, HMMs)







Pfam: Pfam Home Page - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

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

Getting Started Latest Headlines CNB Index of file:///home/... cartoon Traducción en línea ... CNB SBDS

wellcome trust sanger institute

Pfam

Search Pfam

RSS | Pfam Home | Search by | Browse by | FTP | iPfam | Help | About

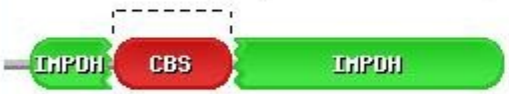
Search by:
- Protein name or sequence
- Keyword
- Domain query
- DNA sequence
- Taxonomy query

Pfam is a large collection of multiple seque s and hidden Markov models covering many common protein domains and famil mily in Pfam you can:

- Look at multiple alignments
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures

For more information on Pfam, on using this site, or on the changes between Pfam releases 19.0 and 20.0, click here.

Pfam can be used to view the domain organisation of proteins. A typical example is shown below. Notice that a single protein can belong to several Pfam families.



74% of protein sequences have at least one match to Pfam. This number is called the sequence coverage and is shown in the pie chart on the right.

Pfam is a database of two parts, the first is the curated part of Pfam containing over 8296 protein families. To give Pfam a more comprehensive coverage of known proteins we automatically generate a supplement called Pfam-B. This contains a large number of small families taken from the PRODOM database that do not overlap with Pfam-A. Although of lower quality Pfam-B families can be useful when no Pfam-A families are found.

Version 20.0

May 2006, **8296** families

Sequence coverage Pfam-A : 74%
Sequence coverage Pfam-B : 11%
Other

Web feed

You can use the RSS feed to keep updated about Pfam releases

XML RSS

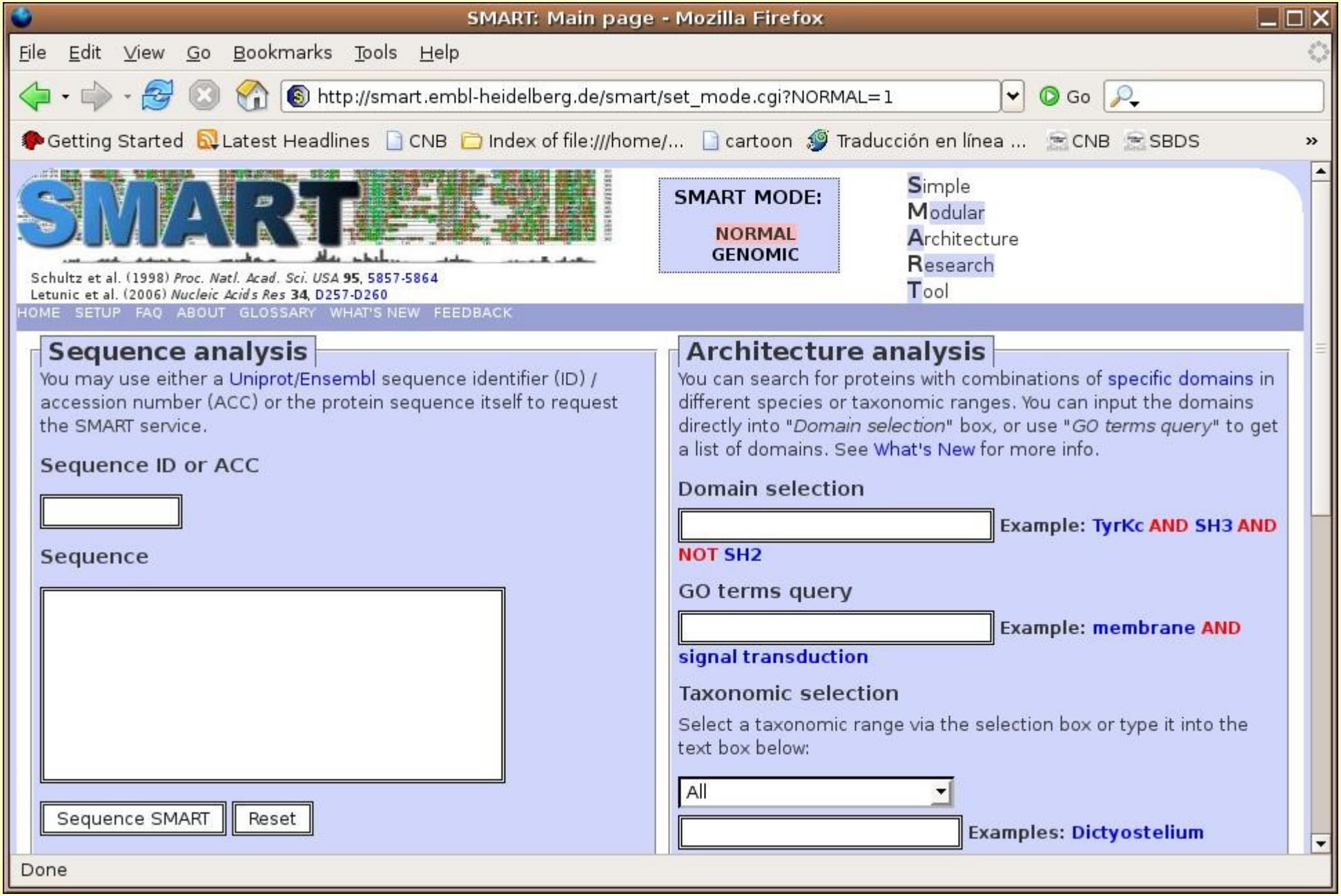
Enter your keyword(s) here

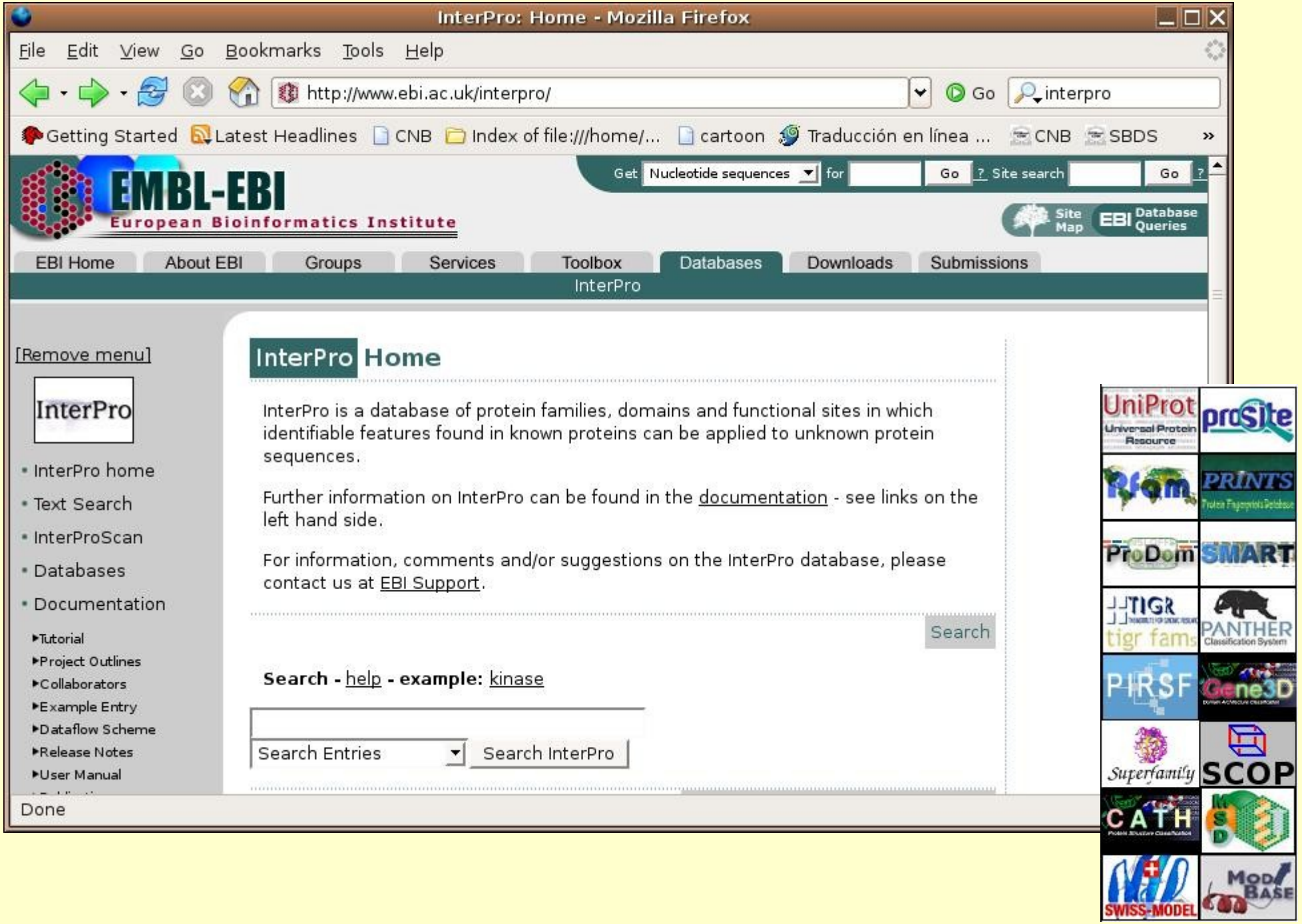
Go Example

Enter a SWISS-PROT 48.1 or TrEMBL 31.1 name or accession number

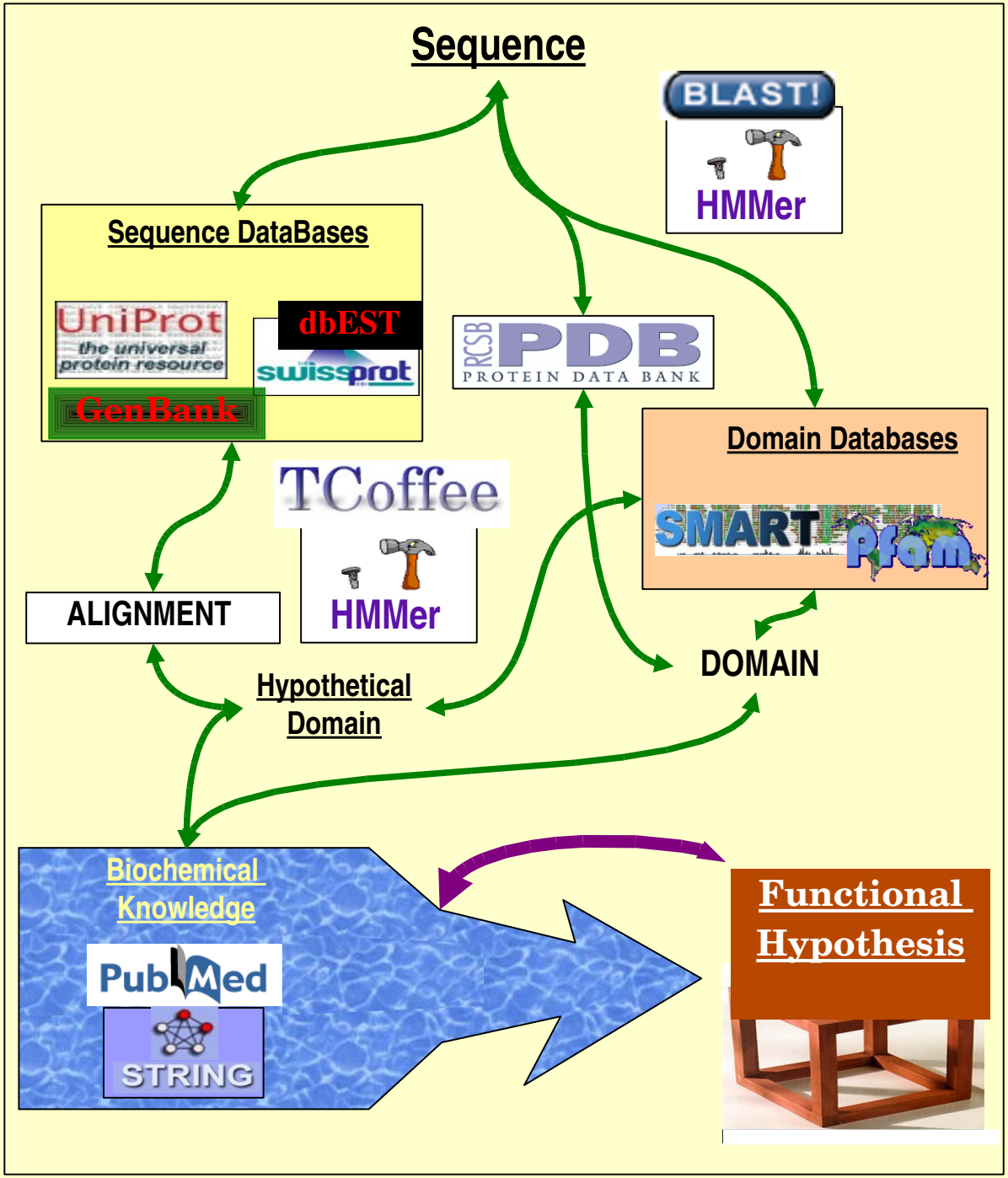
Go Example

Looking up www.sanger.ac.uk...

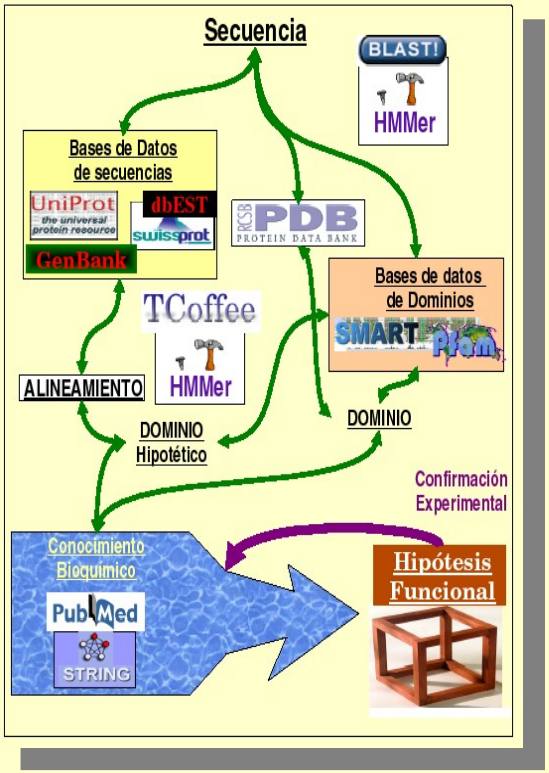

SMART: Main page - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://smart.embl-heidelberg.de/smart/set_mode.cgi?NORMAL=1
Go
Getting Started Latest Headlines CNB Index of file:///home/... cartoon Traducción en línea ... CNB SBDS
SMART
Schultz et al. (1998) Proc. Natl. Acad. Sci. USA 95, 5857-5864
Letunic et al. (2006) Nucleic Acids Res 34, D257-D260
SMART MODE:
NORMAL
GENOMIC
Simple
Modular
Architecture
Research
Tool
HOME SETUP FAQ ABOUT GLOSSARY WHAT'S NEW FEEDBACK
Sequence analysis
You may use either a Uniprot/Ensembl sequence identifier (ID) / accession number (ACC) or the protein sequence itself to request the SMART service.
Sequence ID or ACC
Sequence
Sequence SMART
Reset
Architecture analysis
You can search for proteins with combinations of specific domains in different species or taxonomic ranges. You can input the domains directly into "Domain selection" box, or use "GO terms query" to get a list of domains. See What's New for more info.
Domain selection
Example: TyrKc AND SH3 AND NOT SH2
GO terms query
Example: membrane AND signal transduction
Taxonomic selection
Select a taxonomic range via the selection box or type it into the text box below:
All
Examples: Dictyostelium
Done

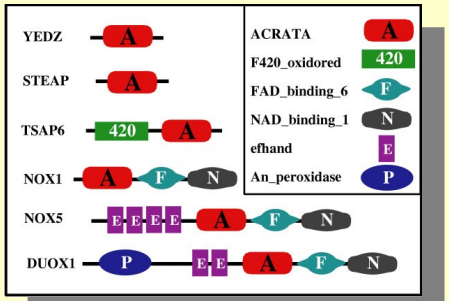

InterPro: Home - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.ebi.ac.uk/interpro/
Go
interpro
Getting Started
Latest Headlines
CNB
Index of file:///home/...
cartoon
Traducción en línea ...
CNB
SBDS
EMBL-EBI
European Bioinformatics Institute
Get Nucleotide sequences for Go ? Site search Go ?
Site Map
EBI Database Queries
EBI Home
About EBI
Groups
Services
Toolbox
Databases
Downloads
Submissions
InterPro
[Remove menu]
InterPro
InterPro home
Text Search
InterProScan
Databases
Documentation
Tutorial
Project Outlines
Collaborators
Example Entry
Dataflow Scheme
Release Notes
User Manual
InterPro Home
InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.
Further information on InterPro can be found in the documentation - see links on the left hand side.
For information, comments and/or suggestions on the InterPro database, please contact us at EBI Support.
Search
Search - help - example: kinase
Search Entries
Search InterPro
Done
UniProt
Universal Protein Resource
proSite
Pfam
PRINTS
ProDom
SMART
TIGR
tigr fams
PANTHER
Classification System
PIRSF
Gene3D
Superfamily
SCOP
CATH
SWISS-MODEL
MODBASE

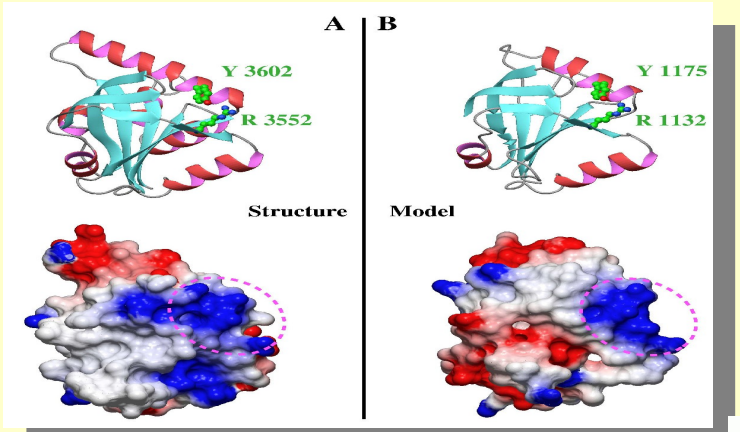
Domain Oriented Sequence Analysis Flow-Chart

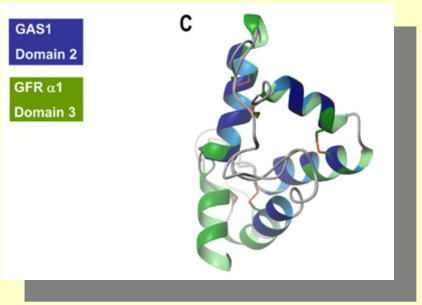


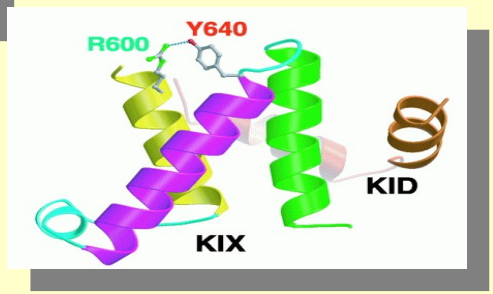
REAL-LIFE EXAMPLES

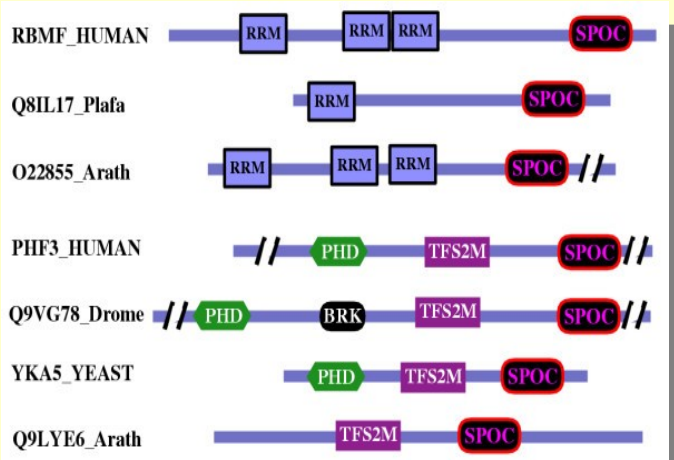


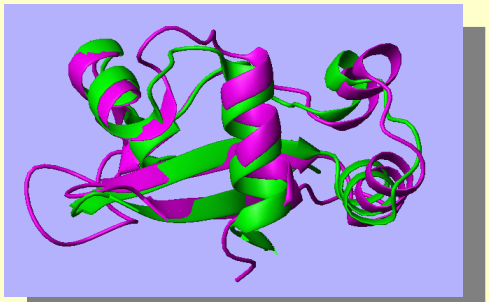








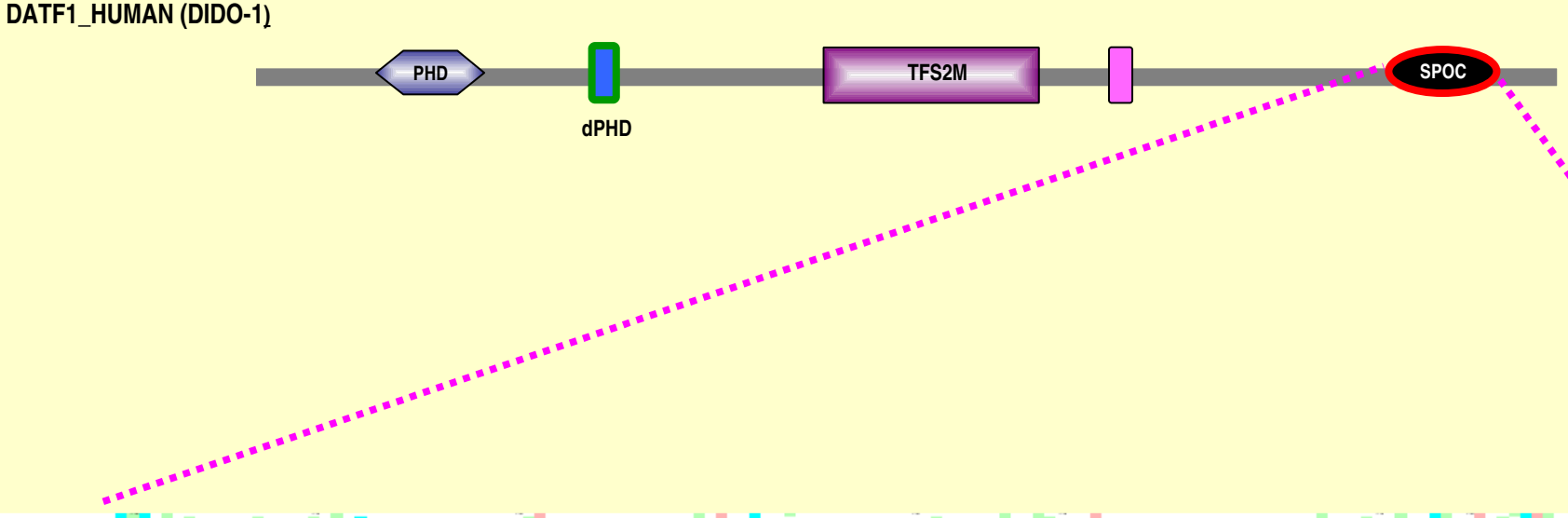


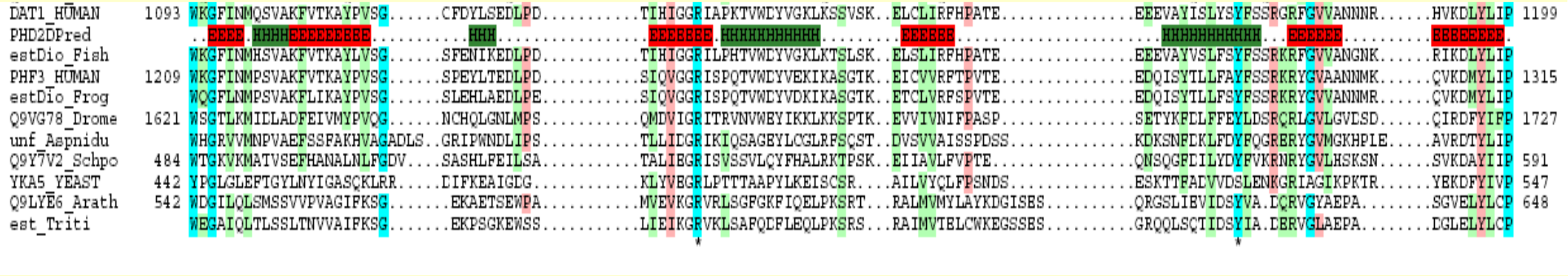


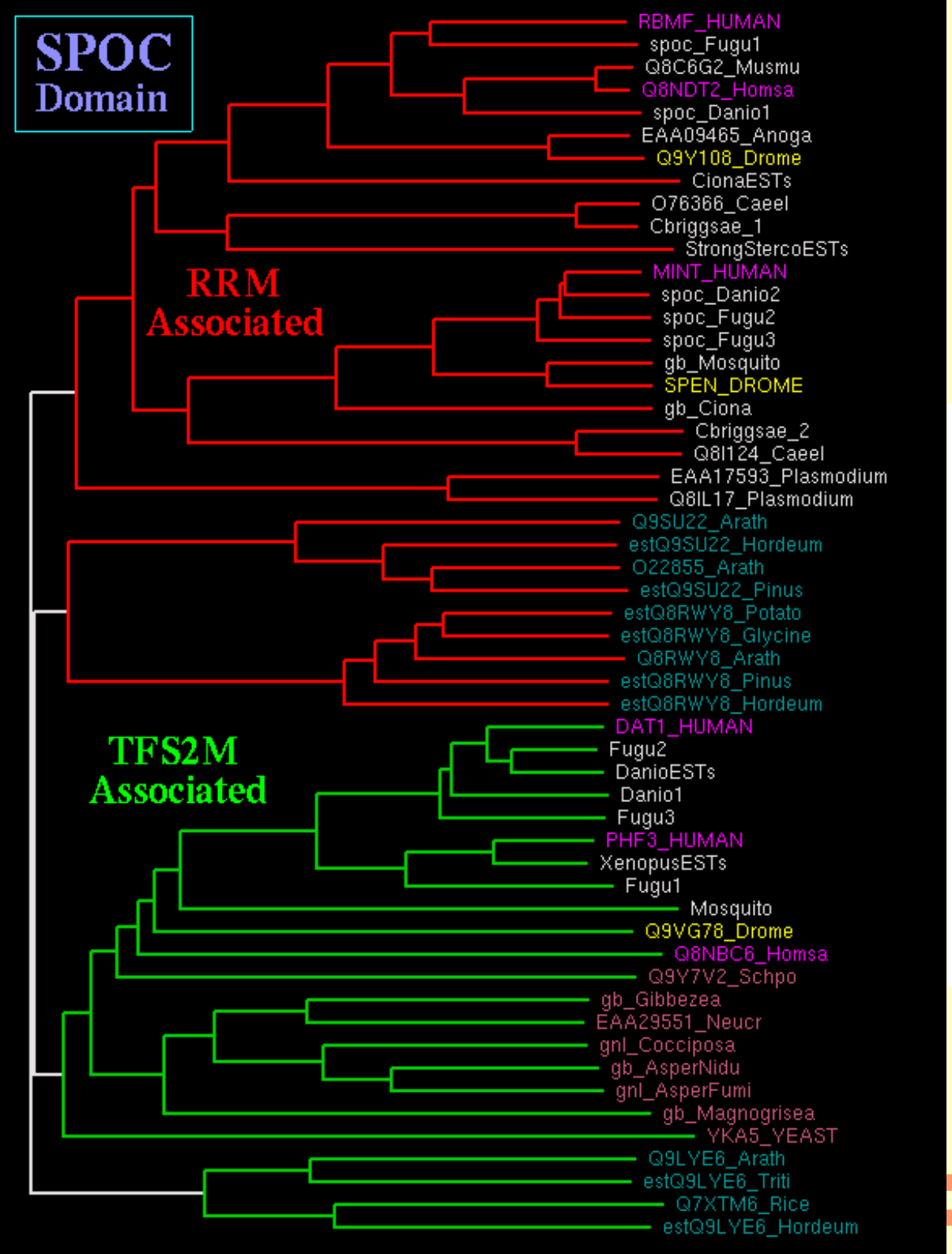
SPOC: A widely distributed domain associated with cancer, apoptosis and transcription.

Sanchez-Pulido L, Rojas AM, Van Wely K, Martinez-A C, Valencia A.

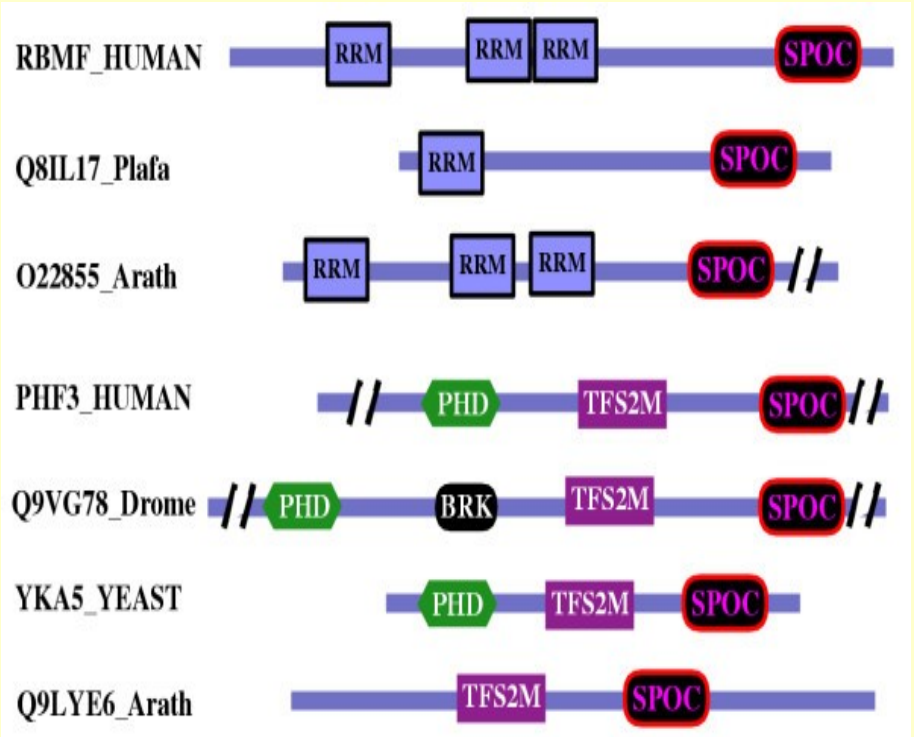
CNB-CSIC

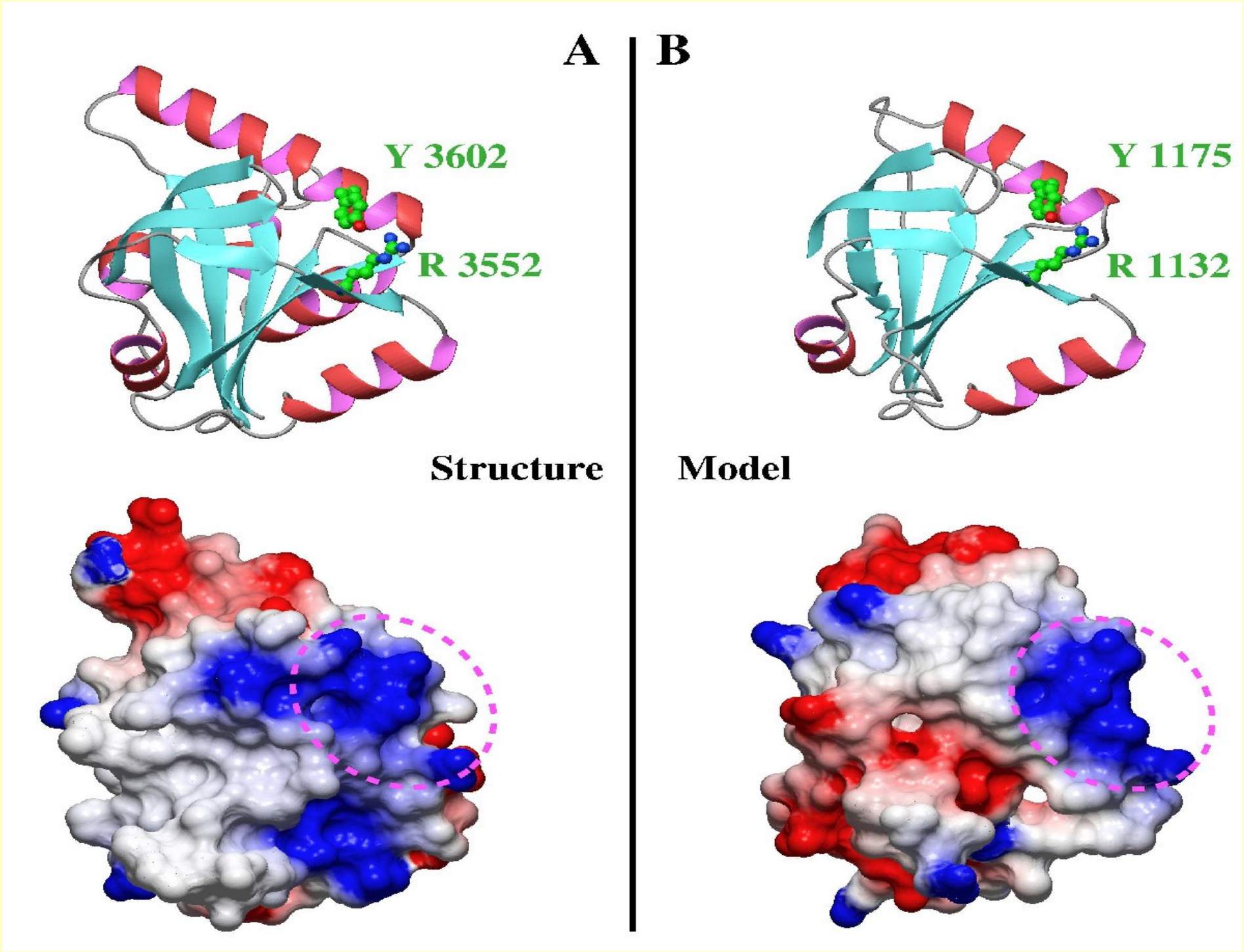






Domain located in, at least, Two Architectures




A
B
Y 3602
R 3552
Y 1175
R 1132
Structure
Model

Gas1 is related to the GFRα family and regulates Ret signaling

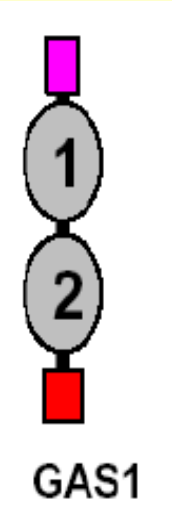
Cabrera J.R., Sánchez-Pulido L., Rojas A.M., Valencia A.,

Mañes S., Naranjo J.R. & Mellstrom B. (2005)

CNB – CSIC

PROTEÍNA INICIAL: GAS1 (Growth Arrest Specific 1)

FUNCIÓN: Regulación de procesos apoptóticos.

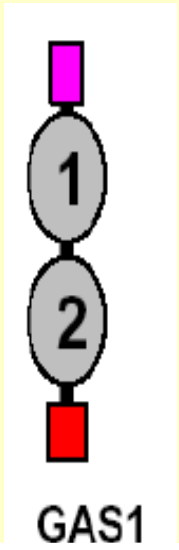


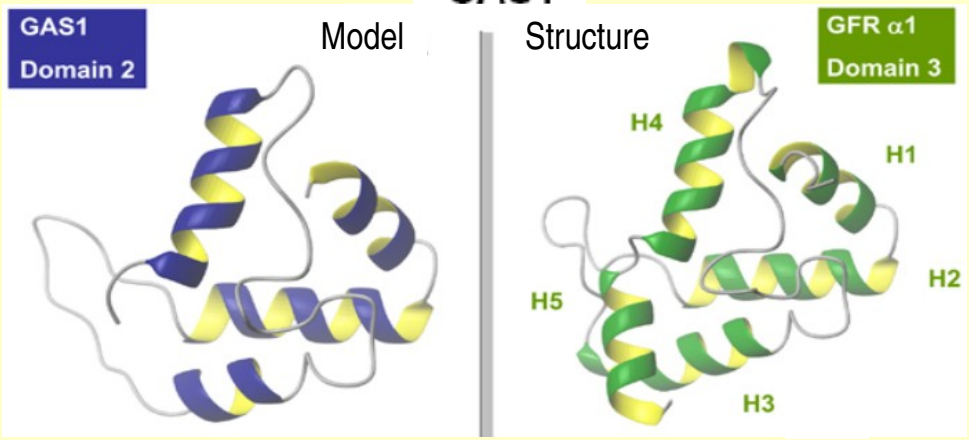
En un primer abordaje:

- Péptido señal
- Duplicación Interna
- GPI-Anclaje

GAS1 &

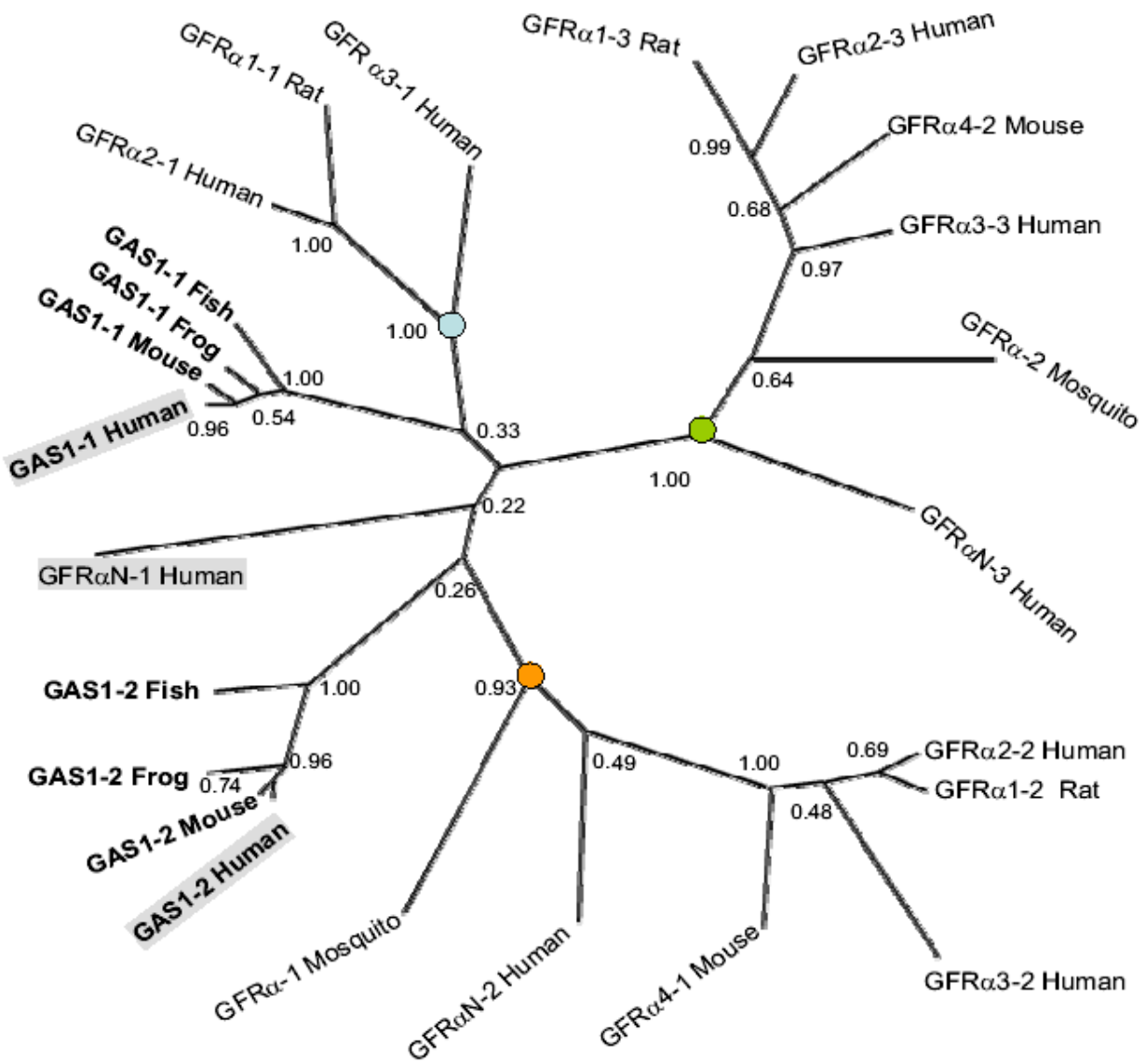
```
GAS1_HUMAN_1     48 CWQALLQCCGE.PECSYAYNQYAEACAPVLAQHGGGDAPGAAAAAFPASAASFSSRWRCP....SHCISALIQLNHTRRGP..........ALEDCDCA.....CDENCKSTKRAIE..PC 146
GAS1_HUMAN_2    166 CTEARRRCDRD.SRCNLALSRYLTYCGKV.....................FNGLRCT....DECRTVIEDMLAMPKVA..........LLNDCVCD...GLERPICESVKENMAR.LC 243
GAS1_MOUSE_1     47 CWQALLQCCGE.PDCSYAYSQYAEACAPVLAQRGGADAPG..PAGAFPASAASSPRWRCP....SHCISALIQLNHTRRGP..........ALEDCDCA.....CDEHCRSTKRAIE..PC 143
GAS1_MOUSE_2    162 CTEARRRCDRD.SRCNLALSRYLAYCGKL.....................FNGLRCT....DECRAVIEDMLAVPKAA..........LLNDCVCD...GLERPICESVKENMAR.LC 239
estGas1_Frog_1   25 CWQAMMRCQEE.AECSYAYRQYVDACSSVLPRPGGEA......ASSSSSSSSSSRRRCP....SHCISALIQLNHTRWGP..........ALEDCDCA.....MDETCRATKRAIE..PC 116
estGas1_Frog_2  138 CMEARKLCEGD.WRCGMSLSRYLTKCGRL.....................FDGLRCT....DECKEVIEDMMRVPKAM..........LLSECECD...GHERPICESIKENMAR.LC 215
Gas1_Fish_1      31 CWKAILKCHGD.PDCHYAYDQYLYACASVI....................SGEHQKCP....SHCISSLIQLNRTQSGP..........ALEDCDCA.....LDPVCRSAKQAIE..PC 107
Gas1_Fish_2     116 CTEARLECEAD.PTCSSAMKDYLFHCRKL.....................FGGERCT....EECRRVIADMRSIPKAQ..........QLDTCVCD...GAERNICEYIKASMKT.FC 193
```



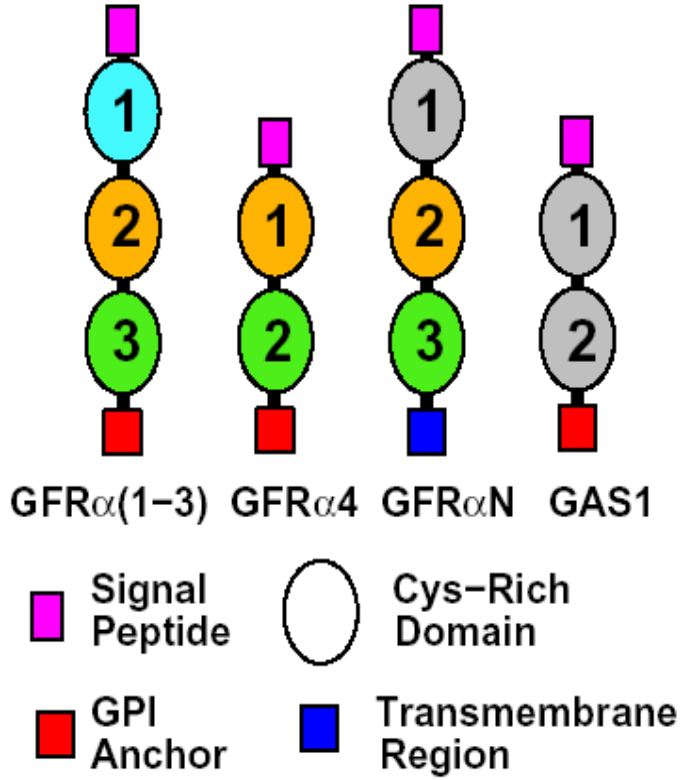


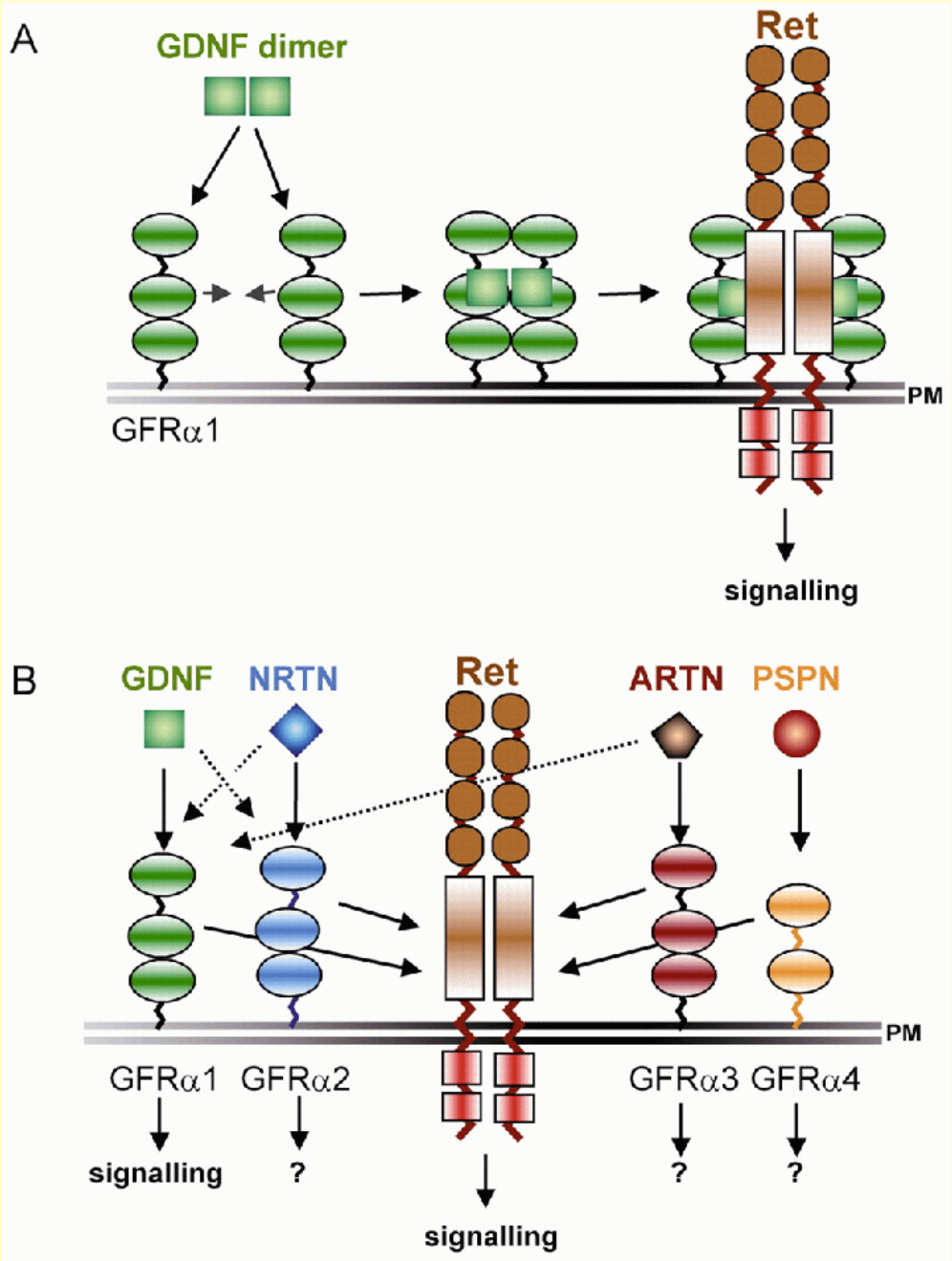
GAS1
Domain 2

GFR α1
Domain 3



- Arquitectura similar





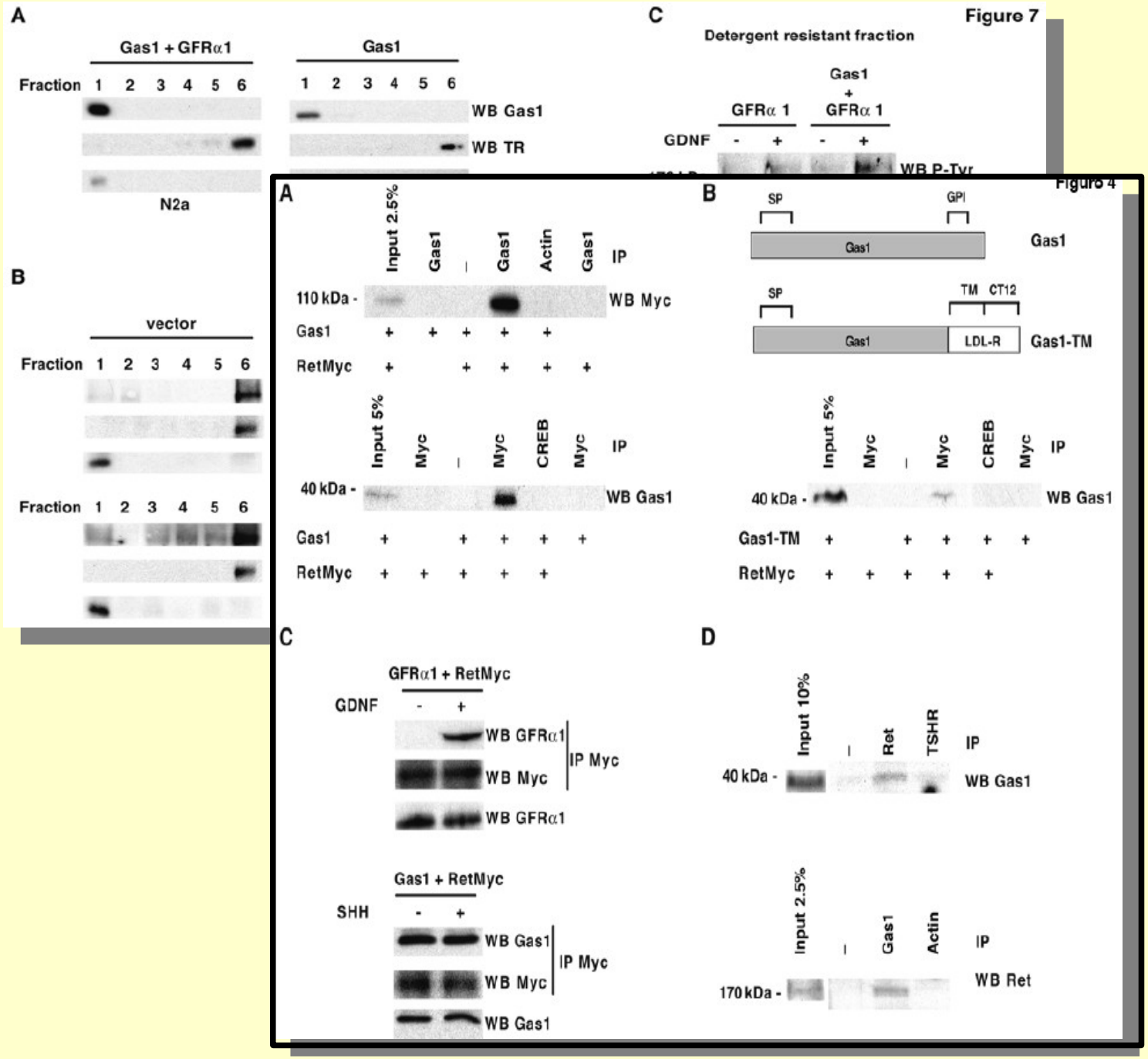
JR Cabrera
CNB-CSIC

BIBLIOGRAFÍA

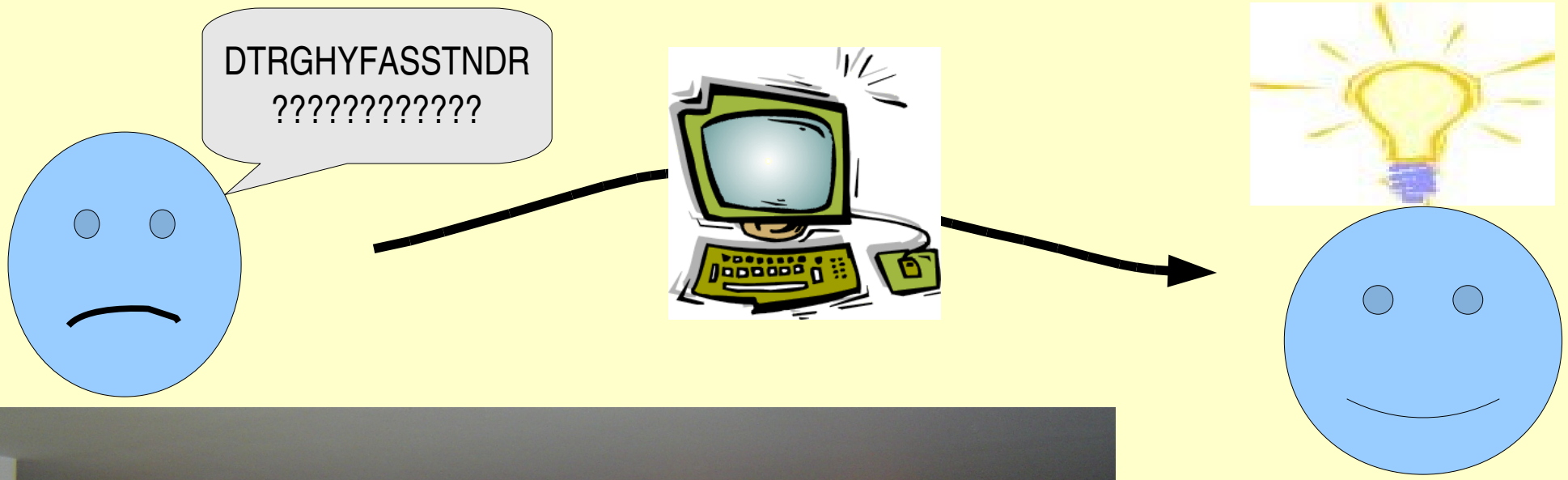
Ligand diversity::

- GDNF Glial cell Derived Neurotrophic Factor
- NRTN Neurturin
- ARTN Artemin
- PSPN Persephin

All GFRalpha interact with Ret -----> and GAS1...?



Computational predictions supported by experimental analysis.



Jorge Ruben,

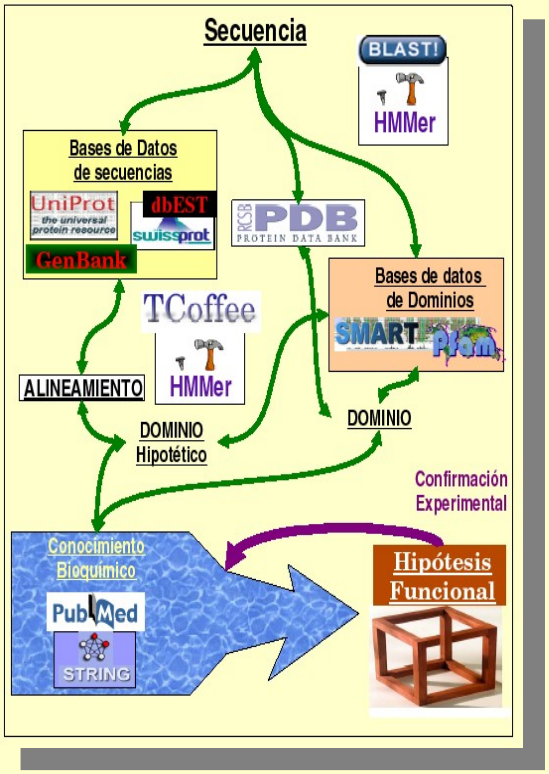
Santos

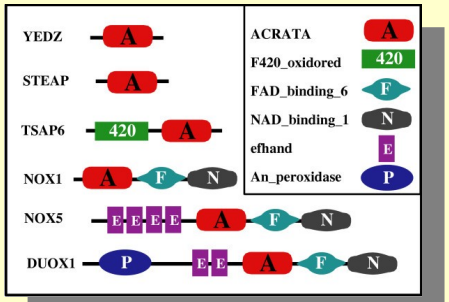
&

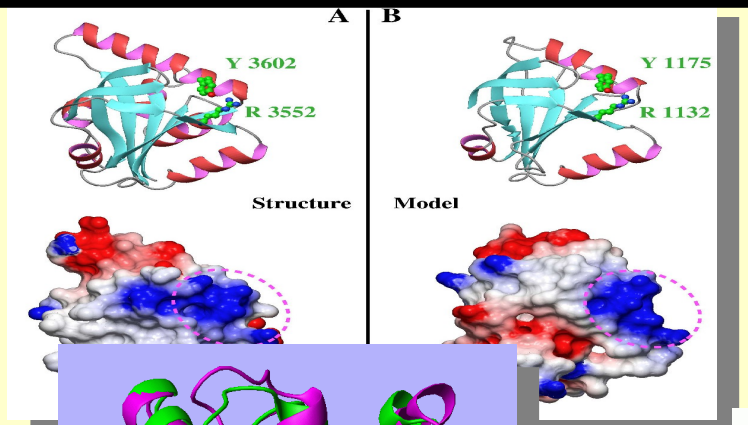
Ana

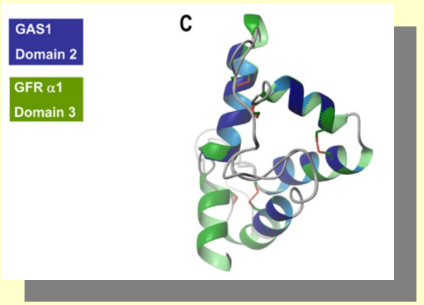
CONCLUSIONE

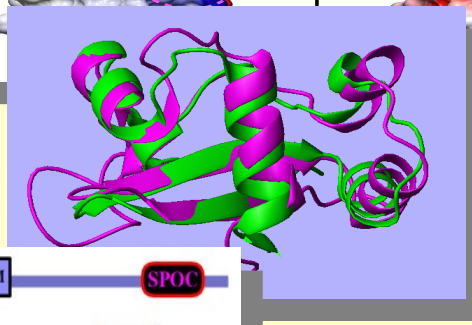
¡Ogni Proteina `e un Mondo!

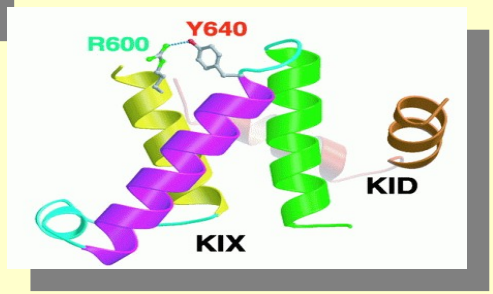


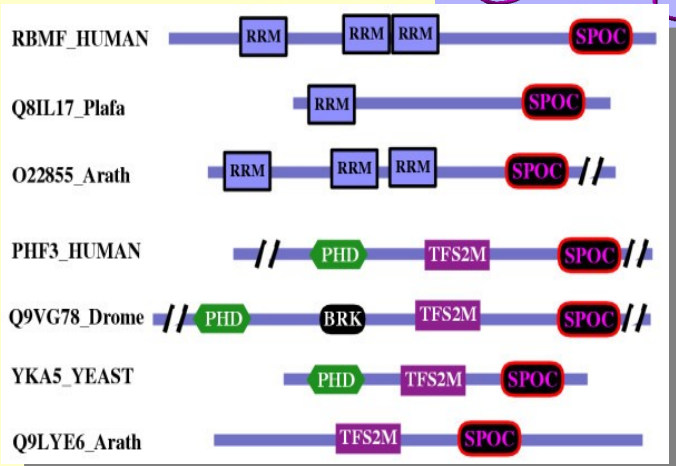












"As a general guide to functional annotation, it should be kept in mind that current methods for genome analysis, even the most powerful and sophisticated of them, facilitate, but do not supplant the work of a human expert."

Eugene Koonin.

Basic References:

Zuckerkandl E, y Pauling L. (1965)
Evolutionary divergence and convergence in proteins.
Evolving Genes and Proteins,
Academic Press, New York, 97-166.

Bork P, Gibson TJ. (1996)
Applying motif and profile searches.
Methods Enzymol. 266:162-184.

Iyer LM, Aravind L, Bork P, Hofmann K, Mushegian AR, Zhulin IB, Koonin EV. (2001)
Quoderat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences.
Genome Biol. 2:RESEARCH0051.

Questions:
sanchez@cnb.uam.es & arojas@cnio.es