MASTER IN TECNOLOGIE BIOINFORMATICHE APPLICATE ALLA MEDICINA PERSONALIZZATA

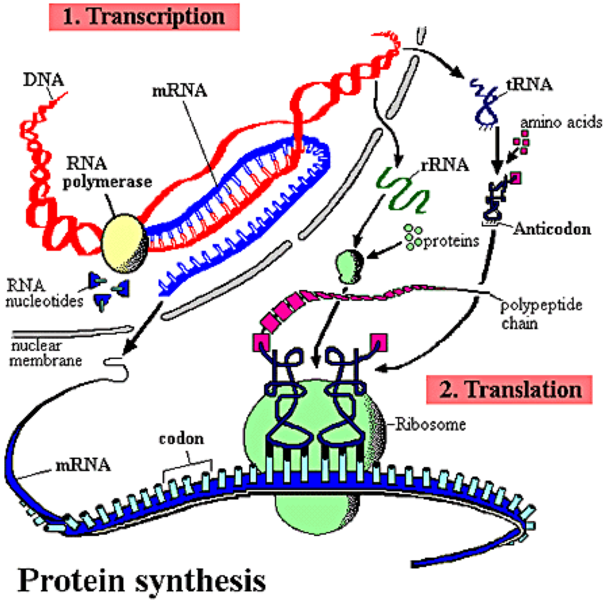
Protein Sequence Analysis

Characteristics of the sequence space and relationships with structure and function spaces

Florencio Pazos (CNB-CSIC)

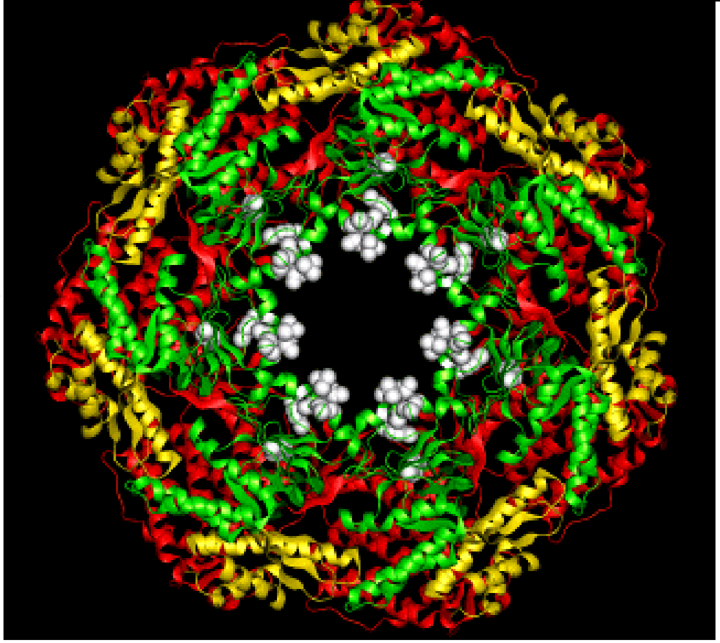
Florencio Pazos Cabaleiro
Protein Design Group (CNB-CSIC)
pazos@cnb.uam.es

Protein sequences structures and functions



subunit

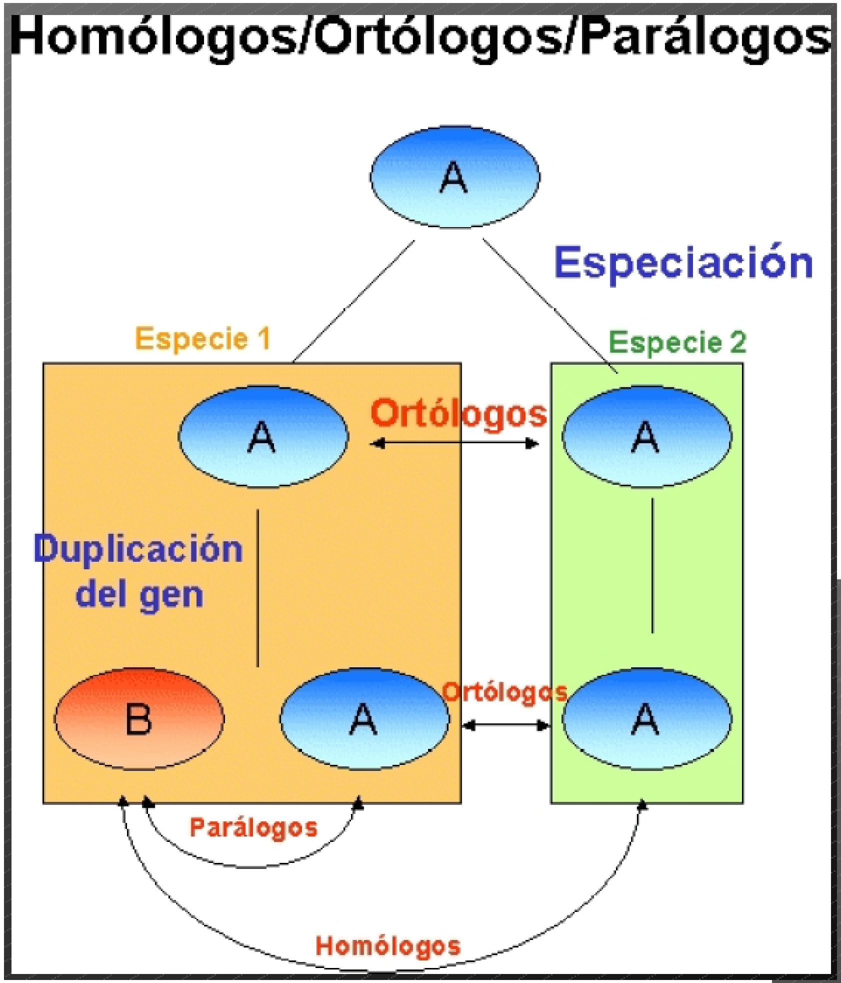
Molecular chaperonin
GroEL



heptamer

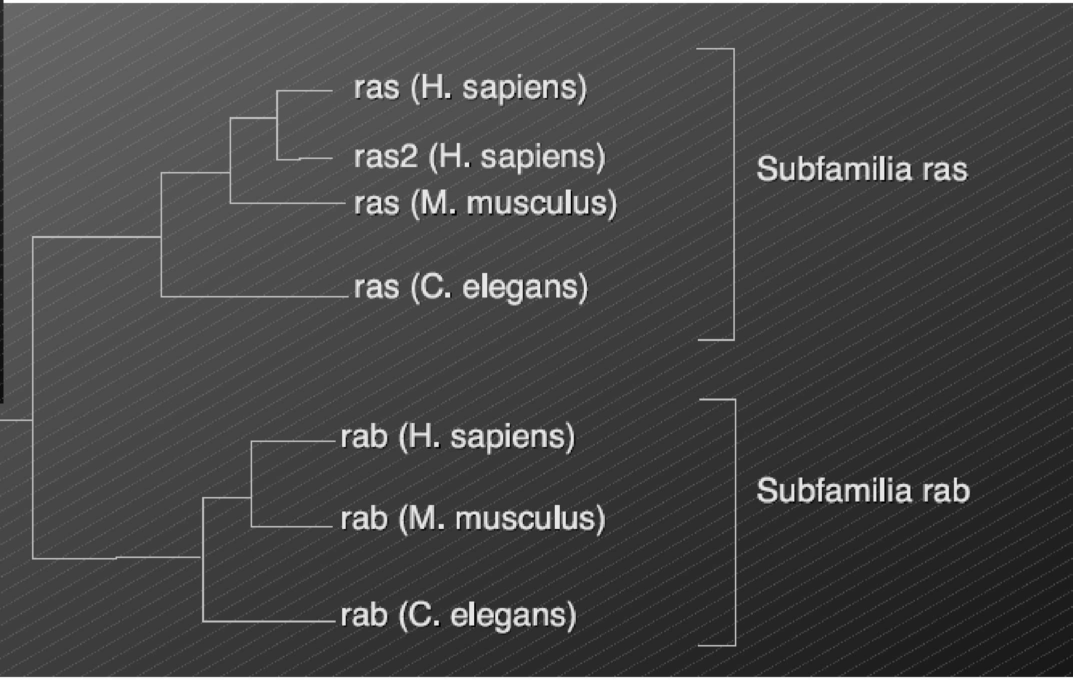
(Dr Jianpeng Ma, Harvard Univ.)

Sequence similarity relatioships

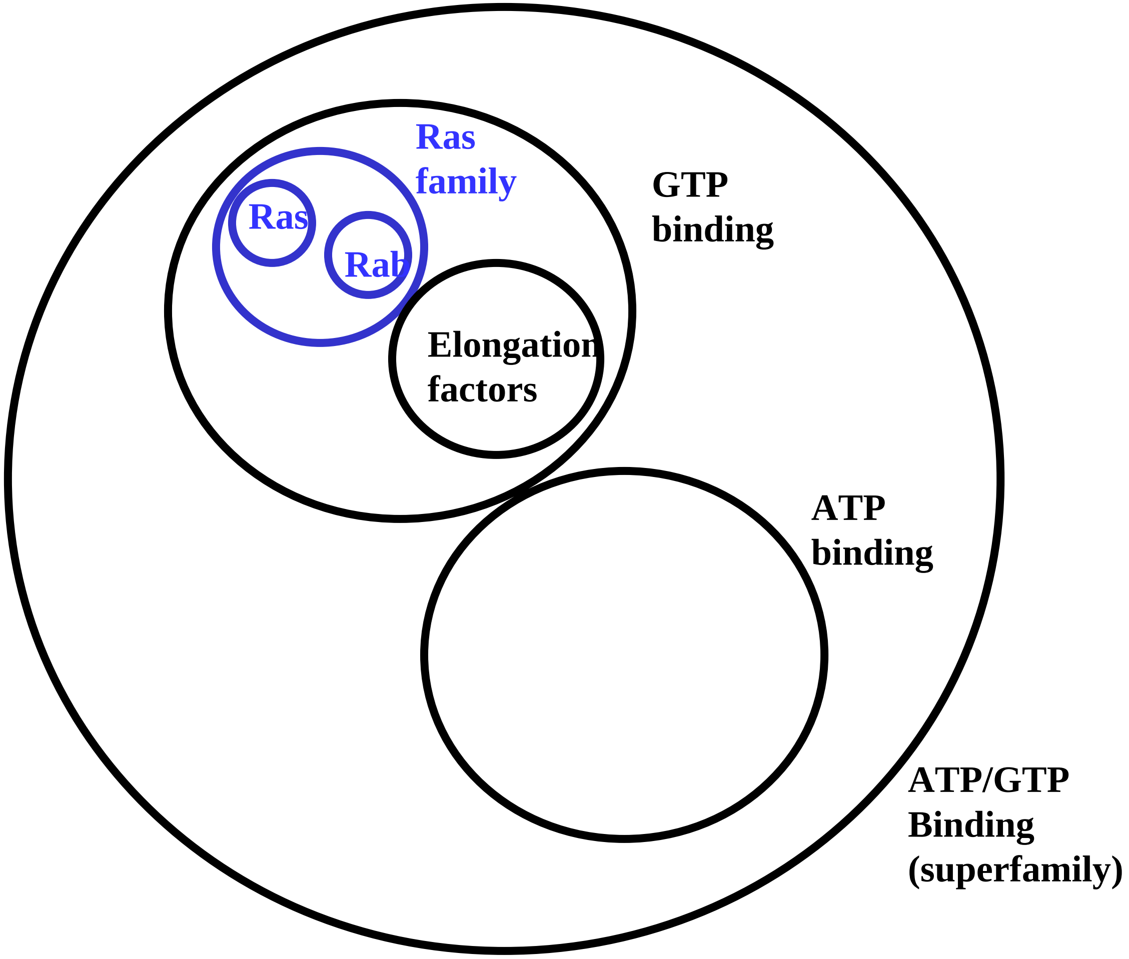


Relationship protein sequence <> function due to the underlying duplication process (divergent evolution)

homologs /orthologs /paralogs



Sequence similarity relatioships



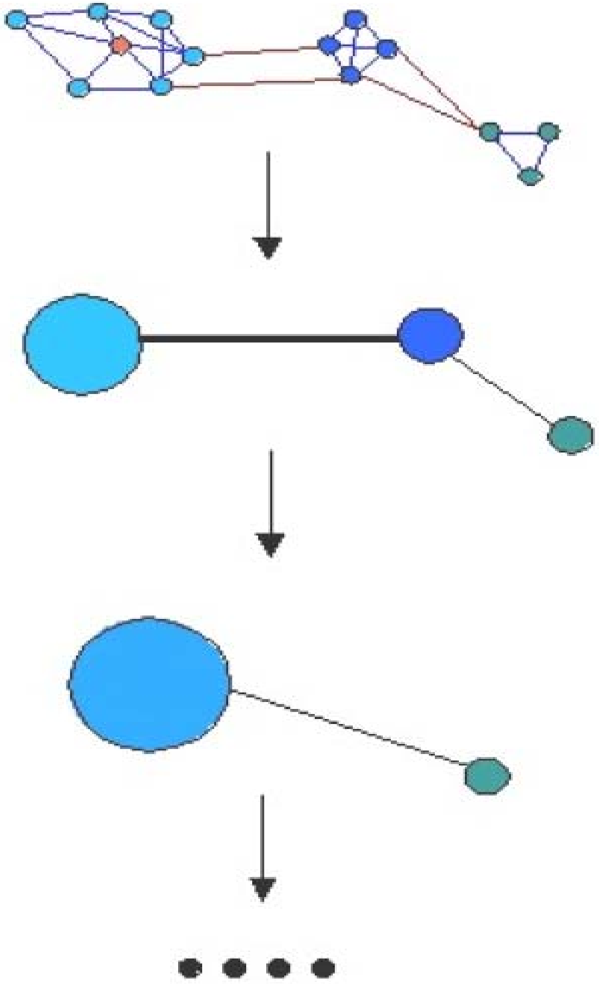
Superfamily: Common origin but maybe not traceable by sequence homology

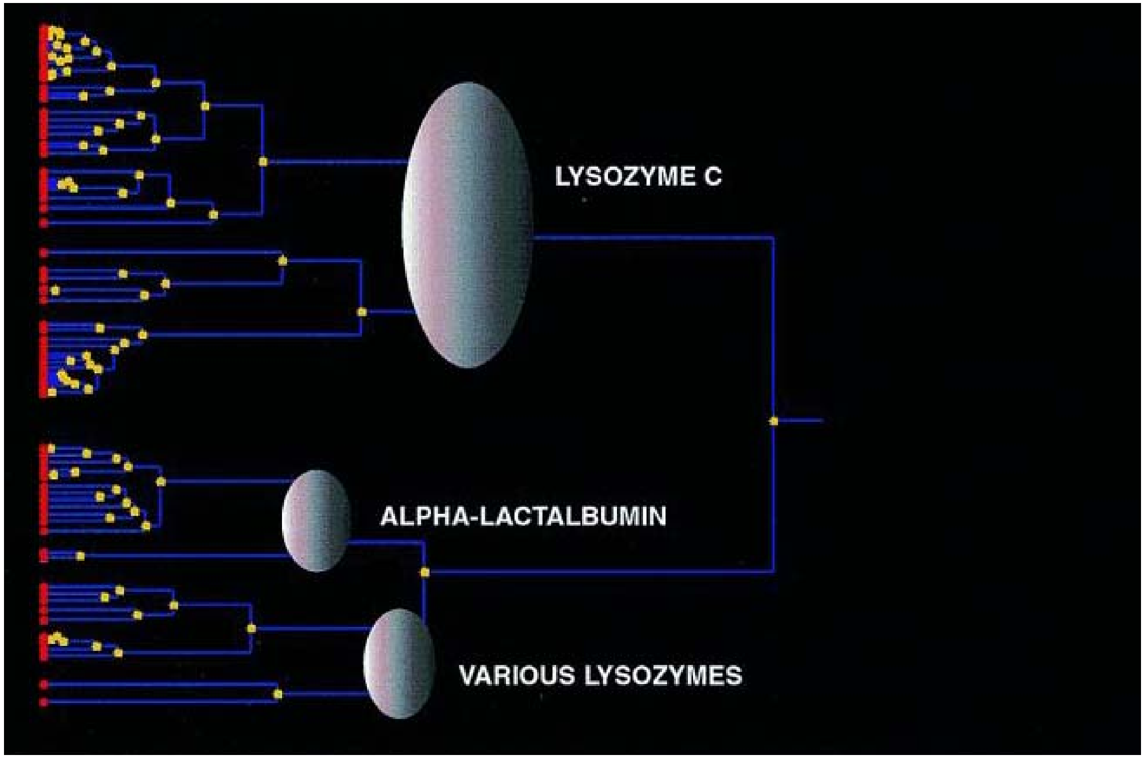
Family: Clear sequence homology. Function can be different.

Subfamily: Clear sequence homology and same function.

| *Sometimes arbitrary*

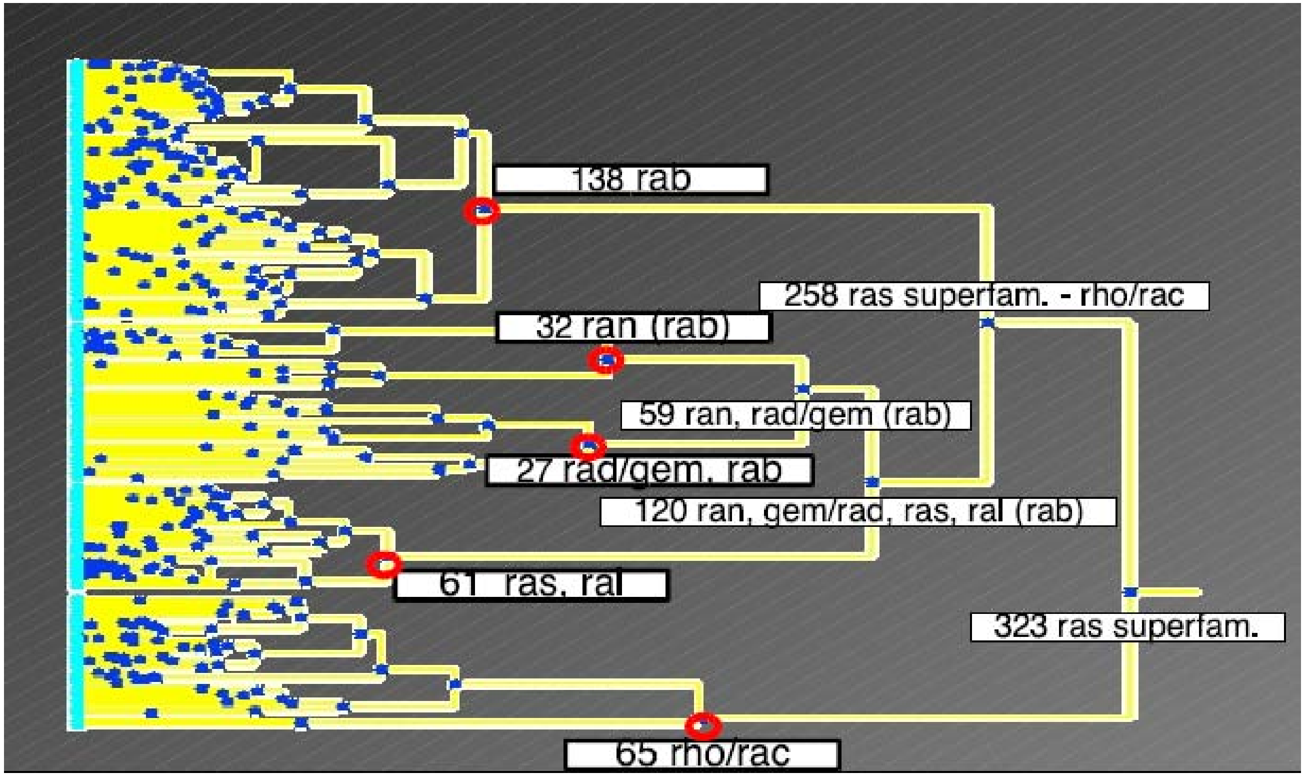
Clustering the whole sequence space





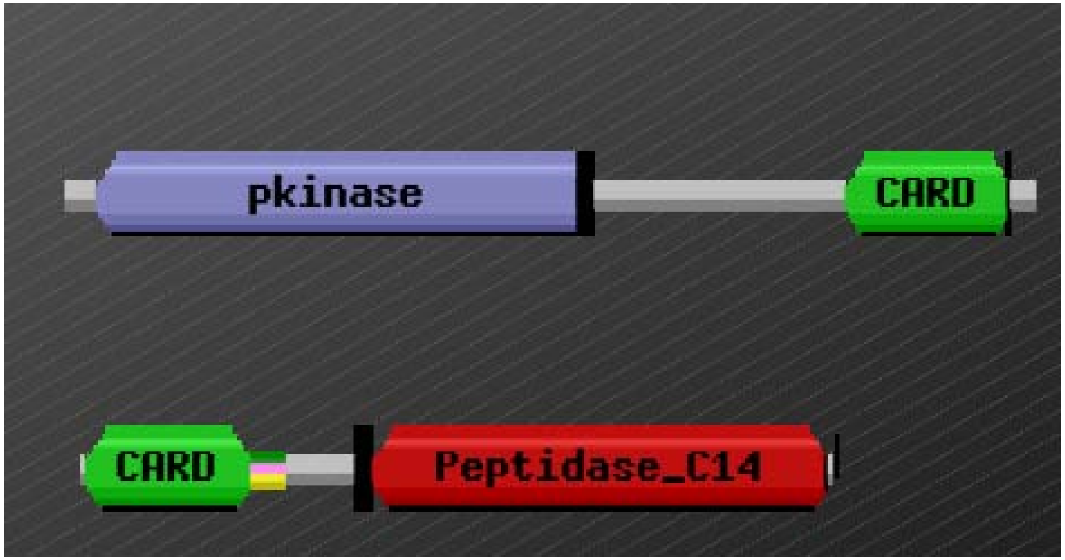
- Yona, G., Linial, N. and Linial, M. (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.*, **28**, 49-55.
- Kaplan, N., Sasson, O., Inbar, U., Friedlich, M., Fromer, M., Fleischer, H., Portugaly, E., Linial, N. and Linial, M. (2005) ProtoNet 4.0: a hierarchical classification of one million protein sequences. *Nucleic Acids Res.*, **33**, D216-218.
- **http://www.protonet.cs.huji.ac.il/**

Clustering the whole sequence space

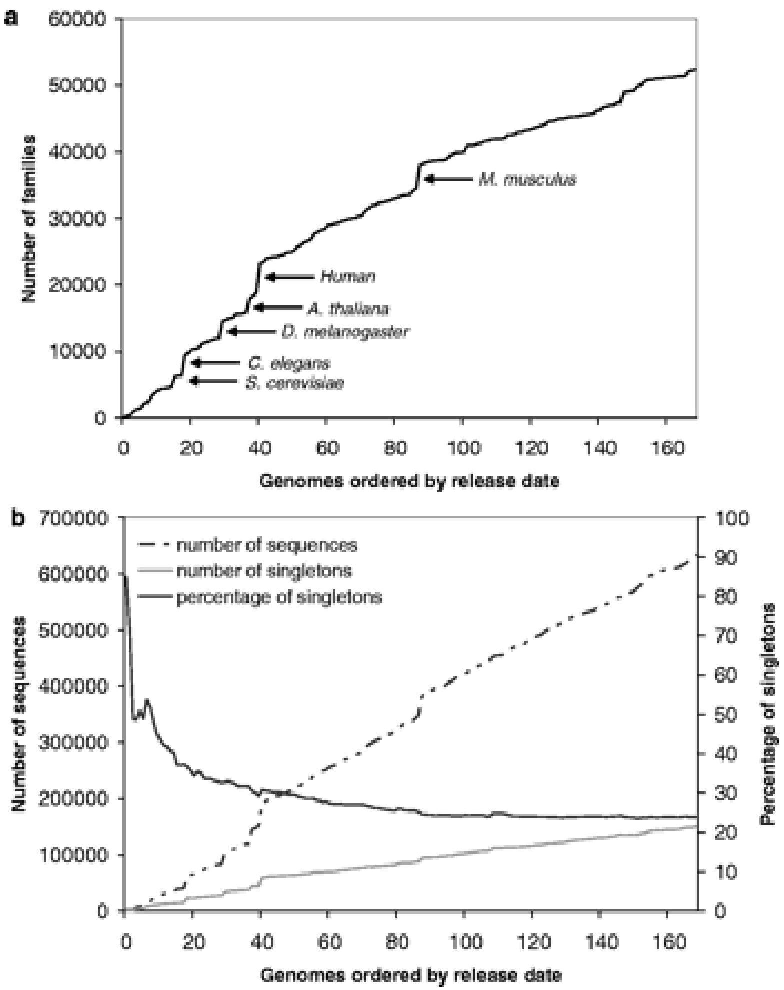


- Yona, G., Linial, N. and Linial, M. (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.*, **28**, 49-55.
- Kaplan, N., Sasson, O., Inbar, U., Friedlich, M., Fromer, M., Fleischer, H., Portugaly, E., Linial, N. and Linial, M. (2005) ProtoNet 4.0: a hierarchical classification of one million protein sequences. *Nucleic Acids Res.*, **33**, D216-218.
- **http://www.protonet.cs.huji.ac.il/**

Protein domains

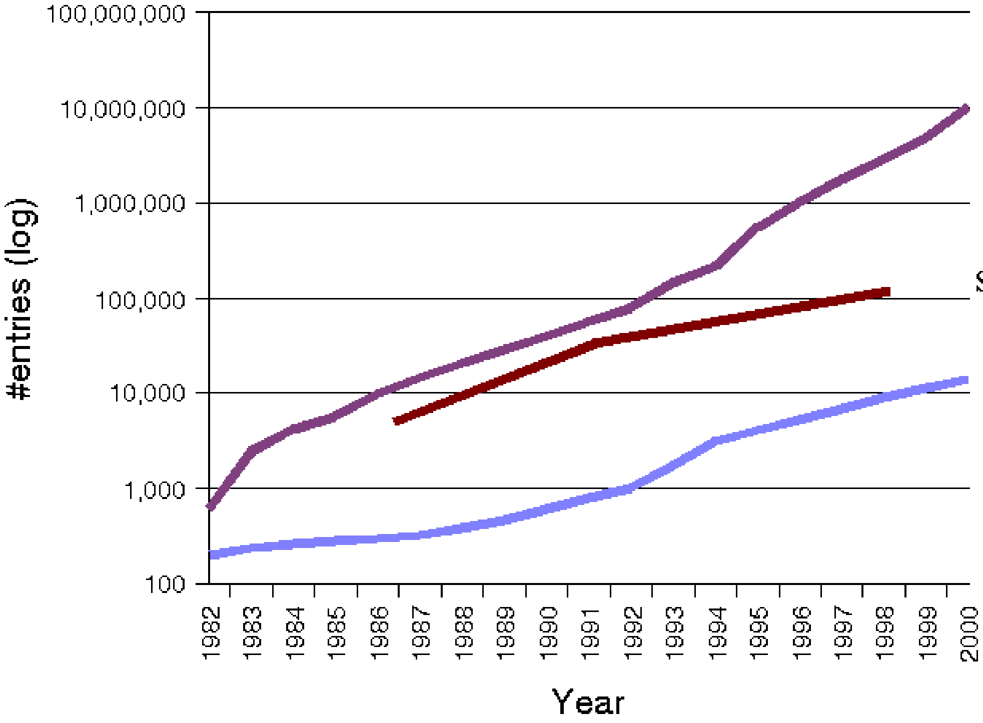


How many protein families?



Marsden, R.L., Lee, D., Maibaum, M., Yeats, C. and Orengo, C.A. (2006) Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. *Nucleic Acids Res.*, **34**, 1066-1080.

Protein Structure



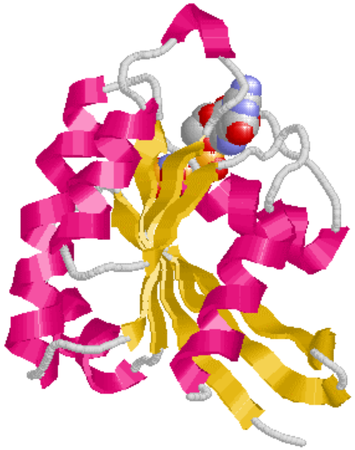
genbank

MREYKLVVLGSGGVGKSALTVQFVQGIFVDEYDPTIEDSY
RKQVEVDCQQCMLEILDTAGTEQFTAMRDLYMKNGQGFAL
VYSITAQSTFNDLQDLREQILRVKDTEDVPMILVGNKCDL
EDERVVGKEQGQNLARQWCNCAFLESSAKSKINVNEIFYD
LVRQINR

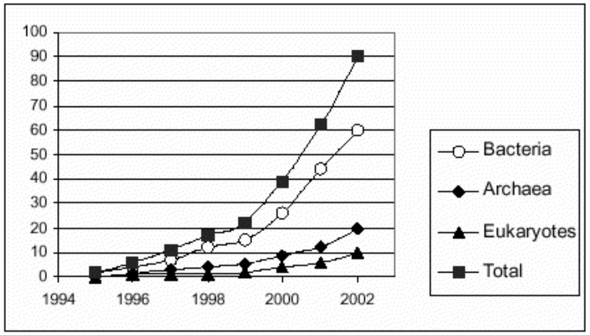
swissprot

MLEILDTA**GTEQFT**AMRDLYMKNGQGFAL
VYSITAQSTFND**LQDLREQIL**RVKDTEDVPMIL
VGNKCDLEDERV

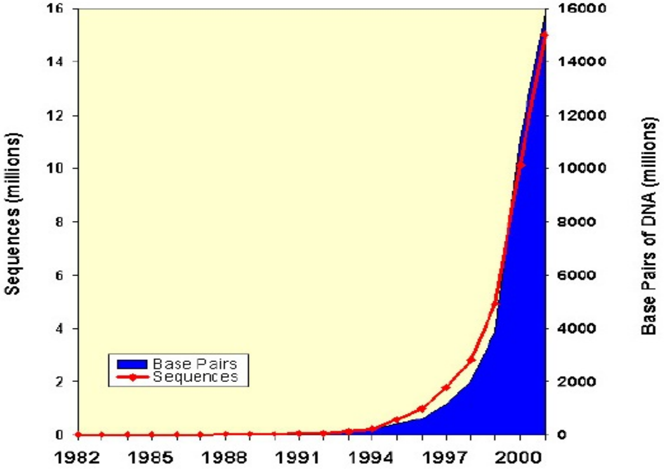
PDB



Genome sequencing

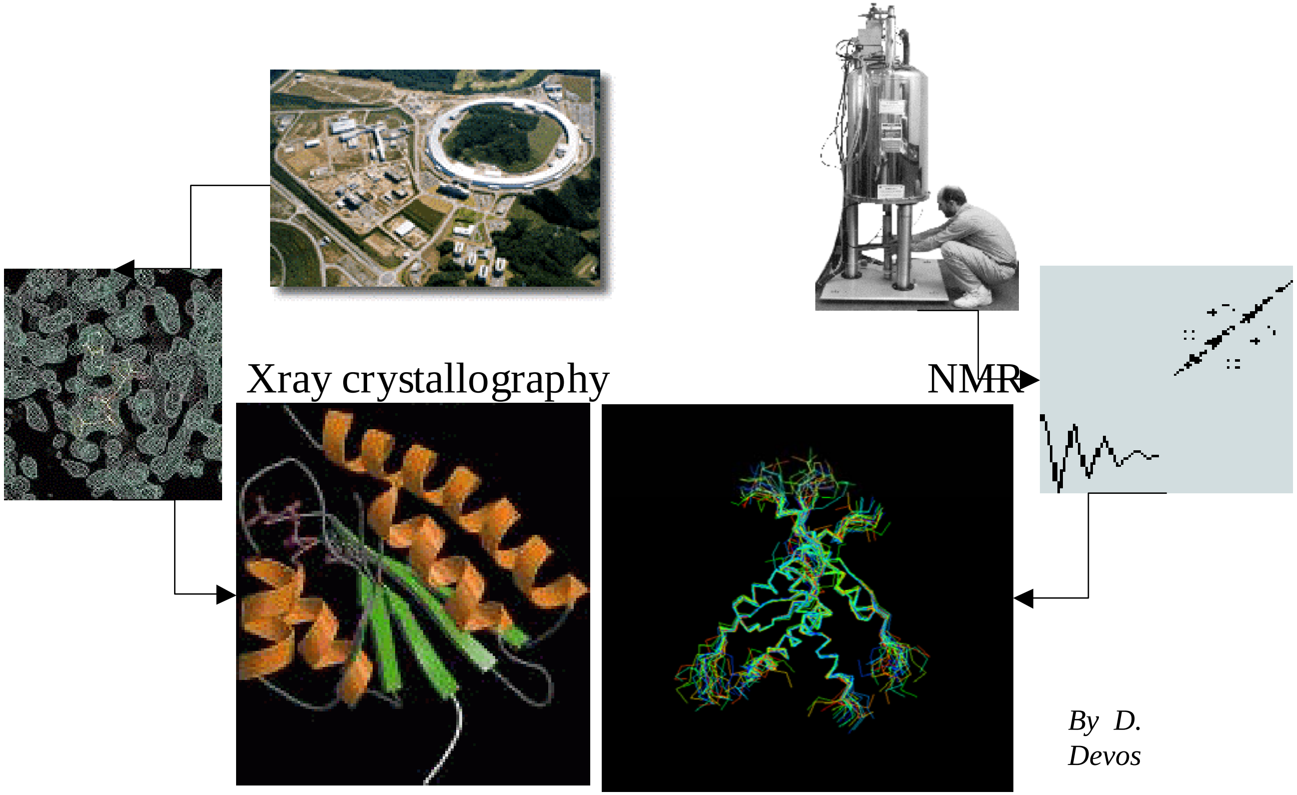


Growth of GenBank



- Collins, F.S., Green, E.D., Guttmacher, A.E. & Guyer, M.S. (2003) A vision for the future of genomic research. *Nature*, **422**, 835-847.
- Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. (2004). *Science* **304,** 66-74.

Experimental determination of protein structures

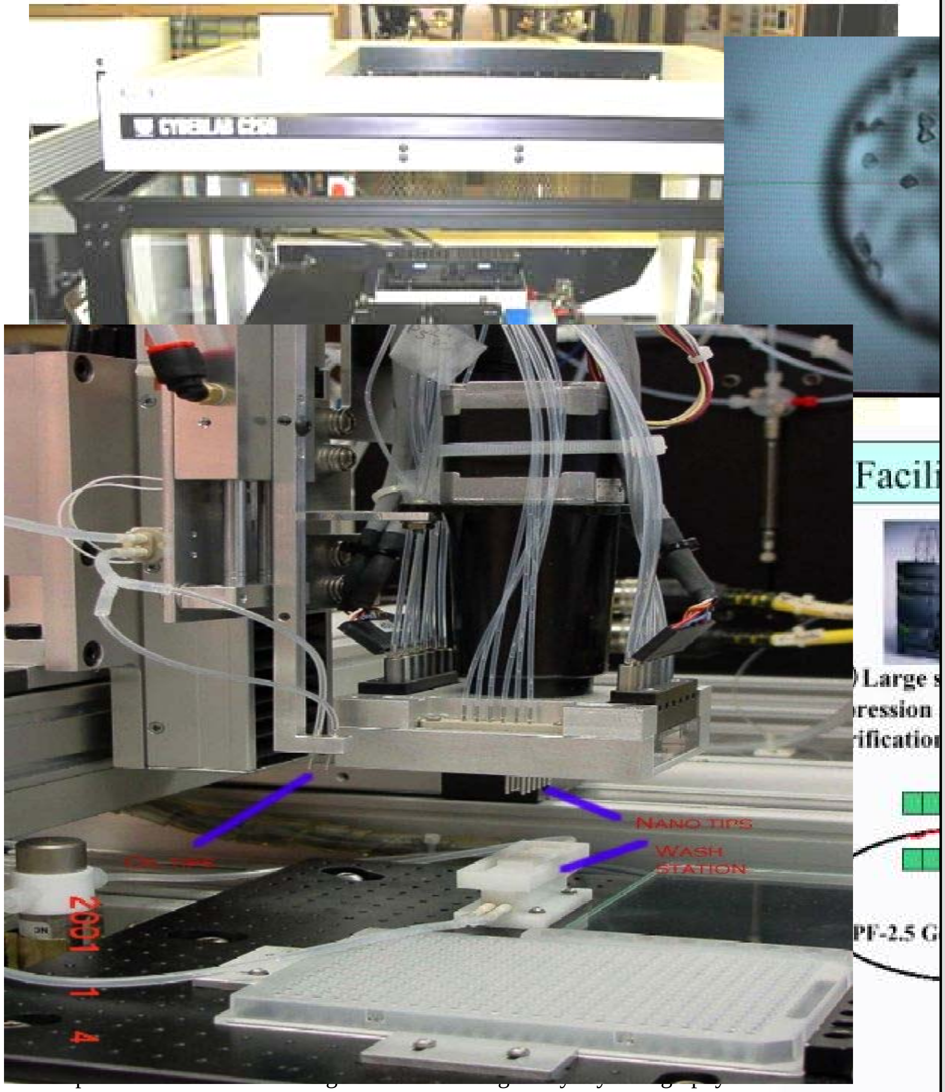


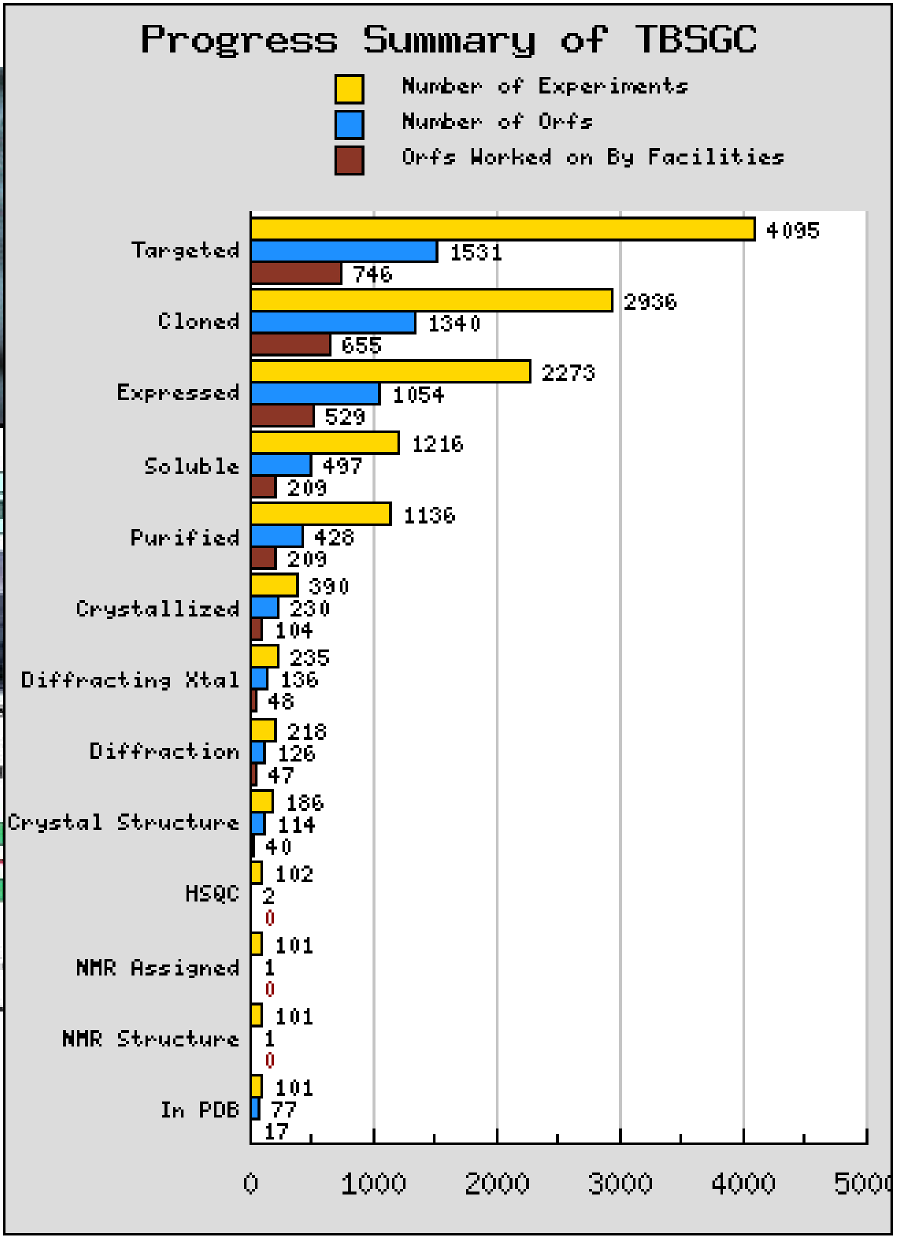
Xray crystallography

NMR

By D. Devos

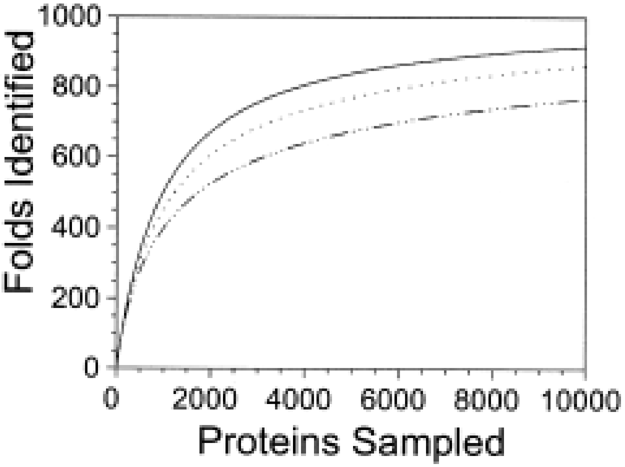
High-throughput determination of protein structures – *Structural Genomics*



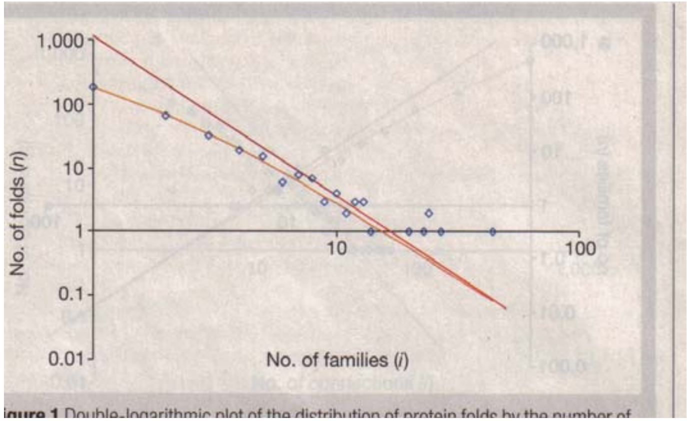


Goldsmith-Fischman, S. and Honig, B. (2003) Structural genomics: Computational methods for structure analysis. *Protein Sci*, **12**, 1813-1821.

Characteristics of the structure space



Leonov, H., Mitchell, J.S. & Arkin, I.T. (2003) Monte Carlo estimation of the number of possible protein folds: effects of sampling bias and folds distributions. *Proteins*, **51**, 352-359.



igure 1 Double-logarithmic plot of the distribution of protein folds by the number of

Koonin, E.V., Wolf, Y.I. & Karev, G.P. (2002) The structure of the protein universe and genome evolution. *Nature*, **420**, 218-223.

~50.000 structural domains

↓

~2.000 unique folds

globin

trefoil

up-down

immunoglobulin

α/β sandwich

jelly roll

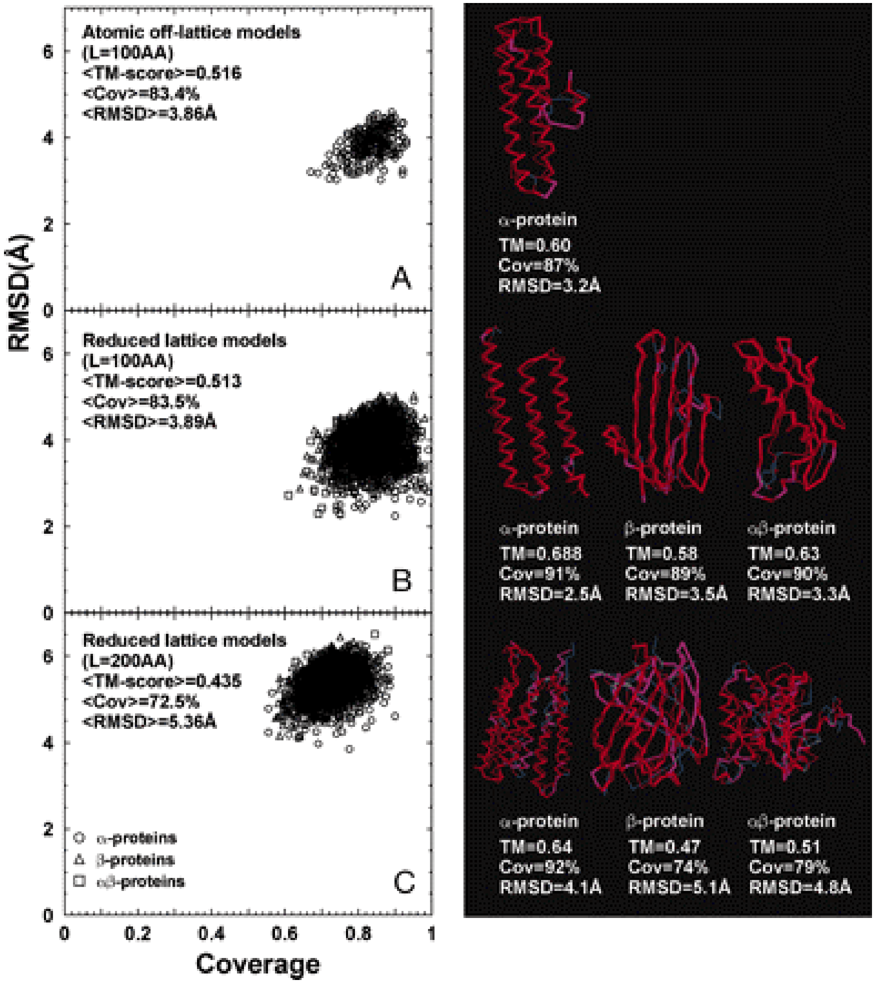
doubly wound

UB α/β roll

TIM barrell

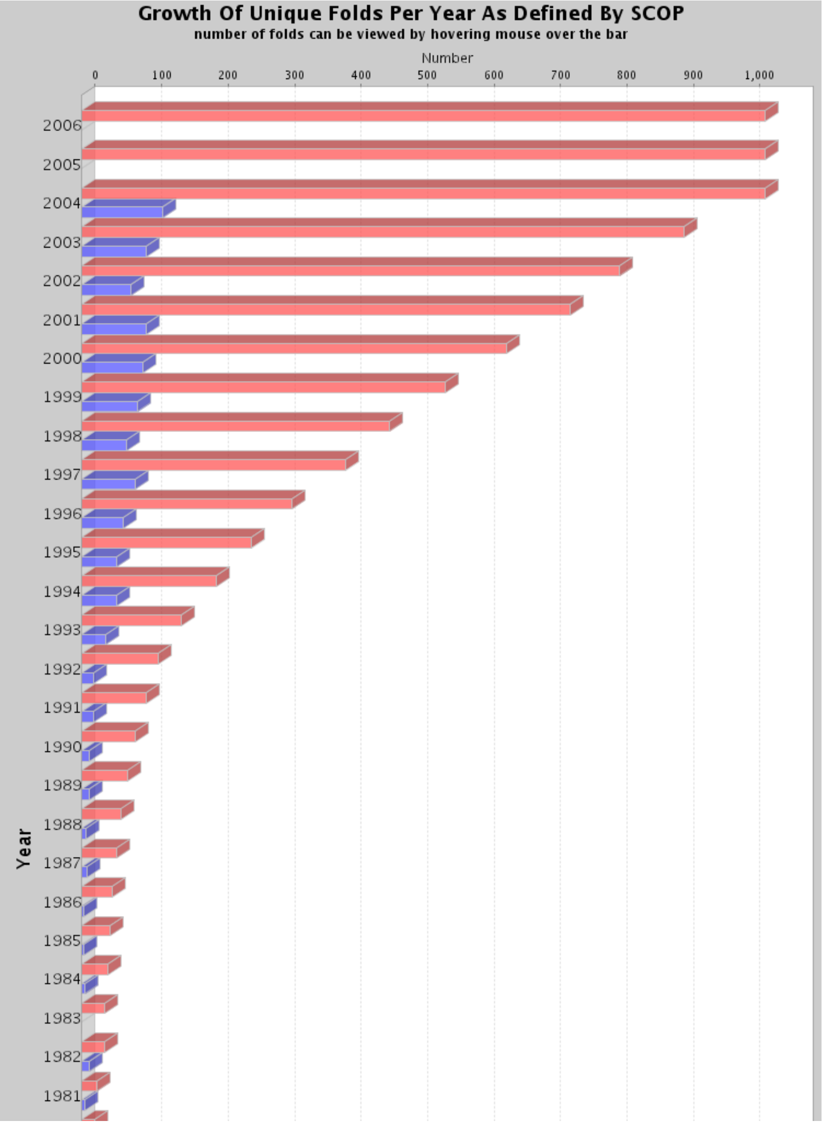
Orengo, C.A., Jones, D.T. & Thornton, J.M. (1994) Protein superfamilies and domain superfolds. *Nature*, **372**, 631-634.

Characteristics of the structure space



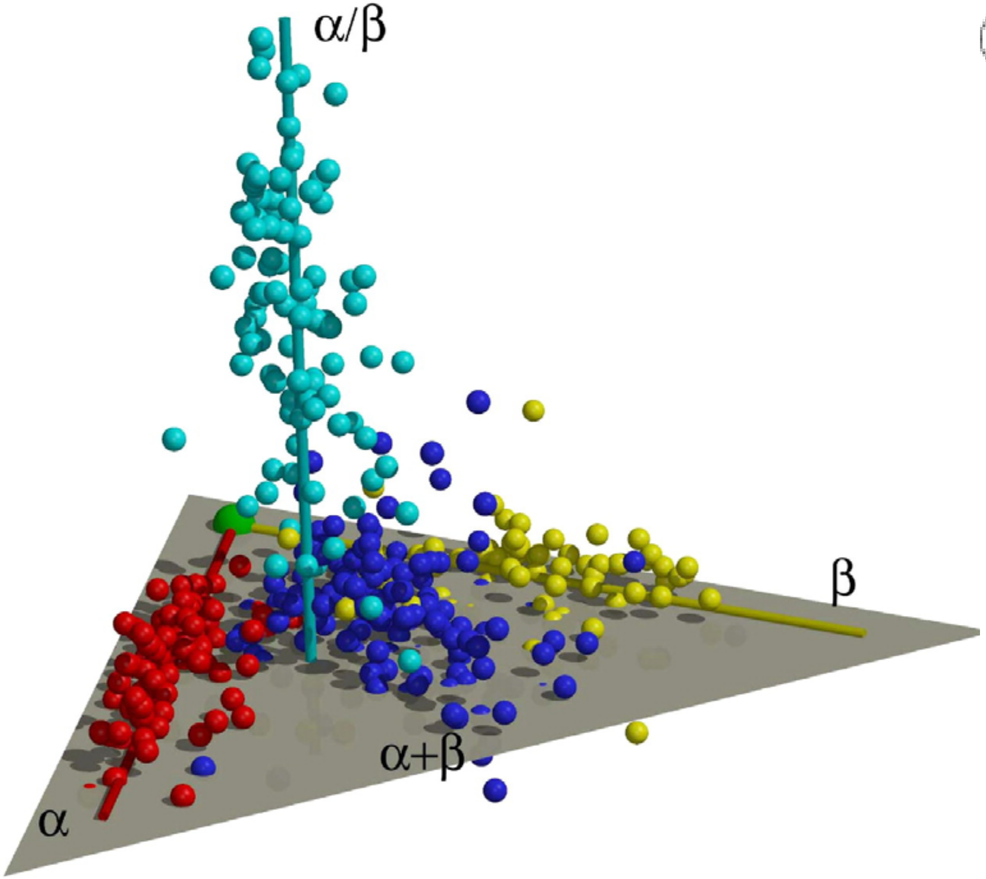
Zhang, Y., Hubner, I.A., Arakaki, A.K., Shakhnovich, E. and Skolnick, J. (2006) On the origin and highly likely completeness of single-domain protein structures. *Proc Natl Acad Sci U S A*, **103**, 2605-2610.

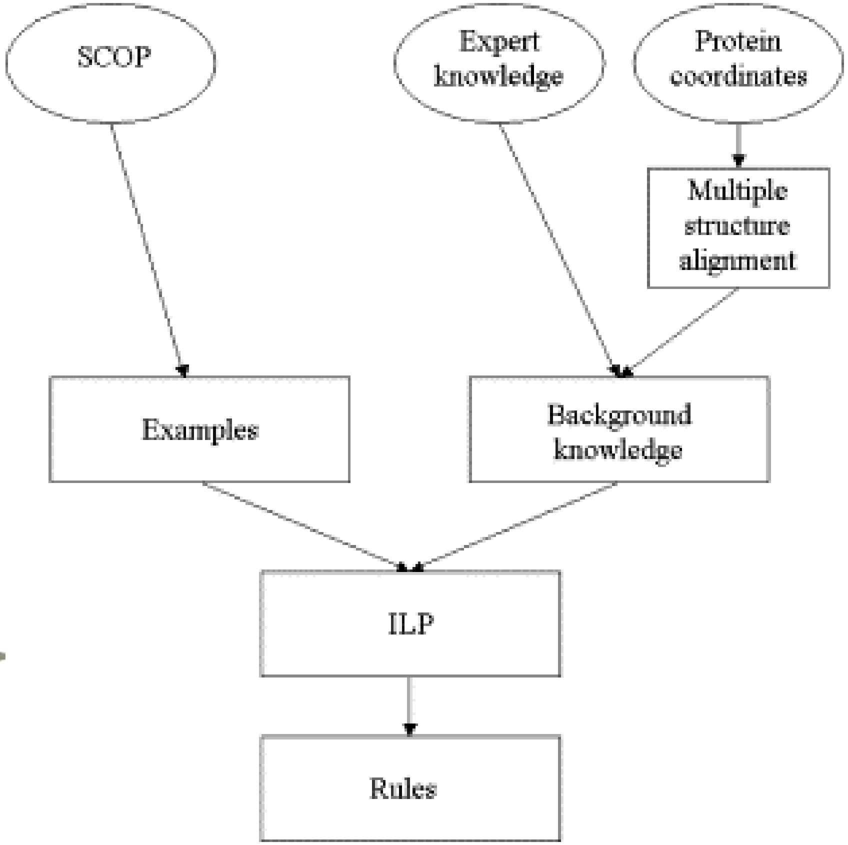
Characteristics of the structure space



http://www.rcsb.org

Characteristics of the structure space

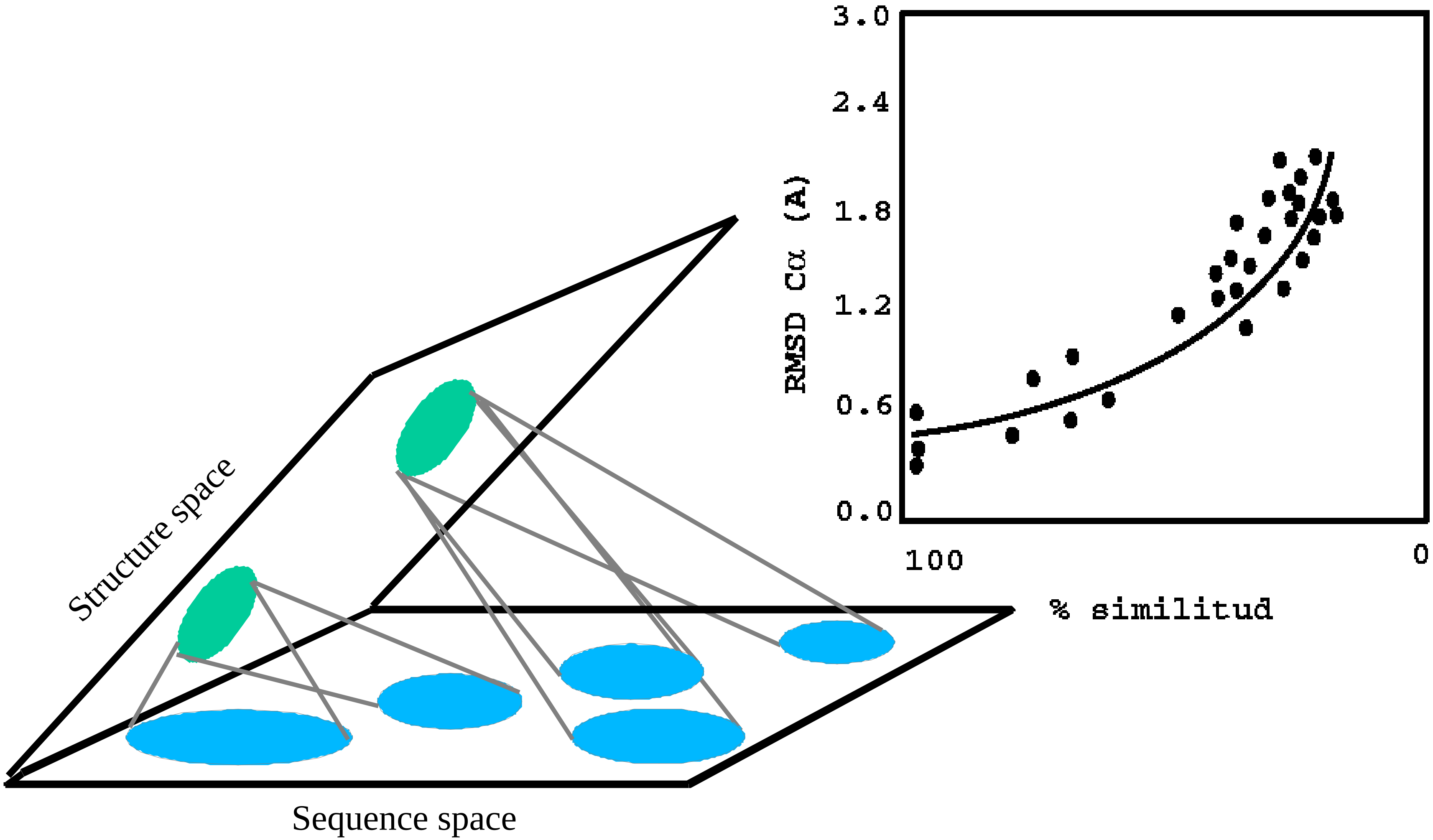




Hou, J., Sims, G.E., Zhang, C. & Kim, S.H. (2003) A global representation of the protein fold space. *Proc Natl Acad Sci USA*, **100**, 2386-2390.

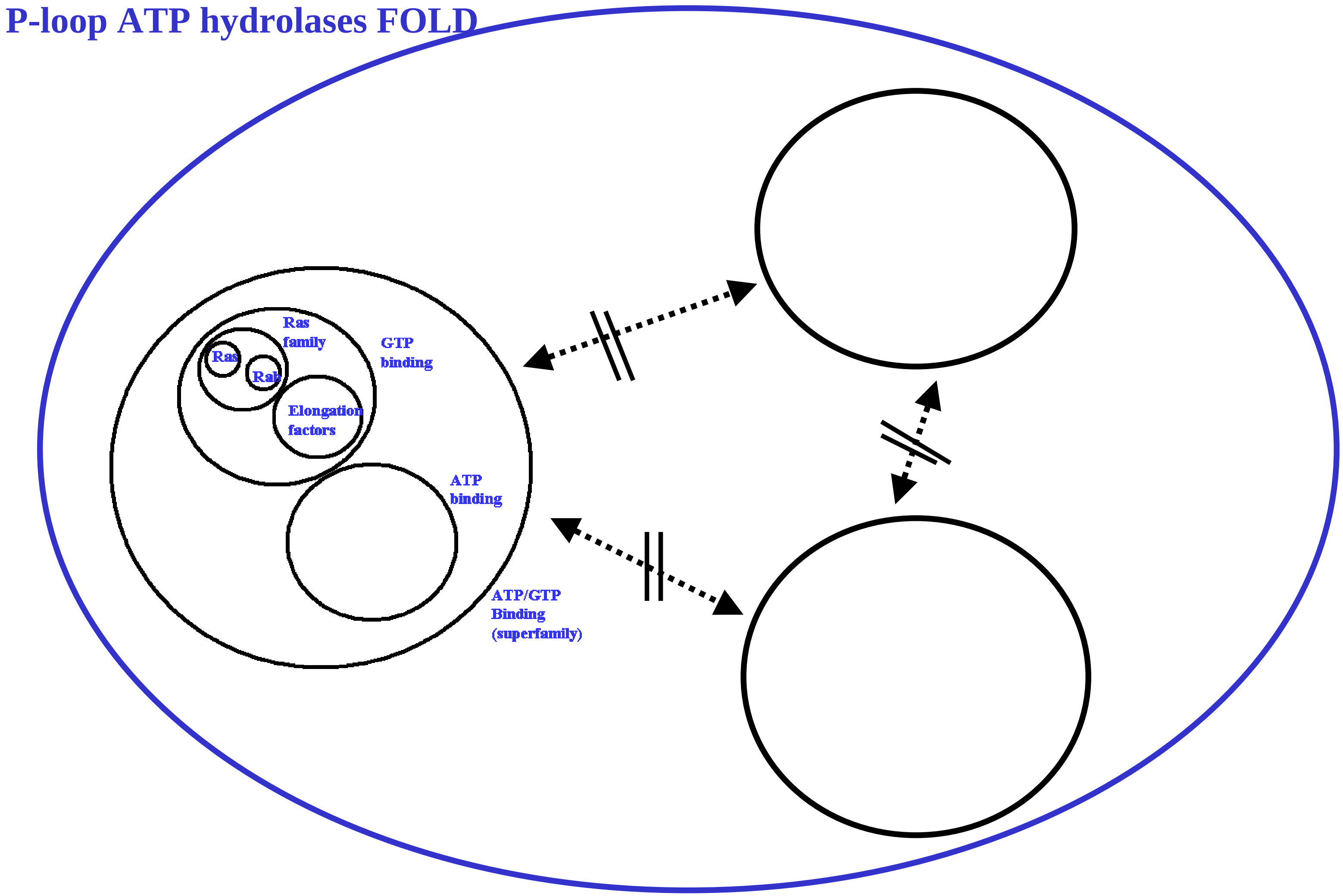
Cootes, A.P., Muggleton, S.H. & Sternberg, M.J. (2003) The automatic discovery of structural principles describing protein fold space. *J Mol Biol*, **330**, 839-850.

Characteristics of the structure space
Relationships with sequence space

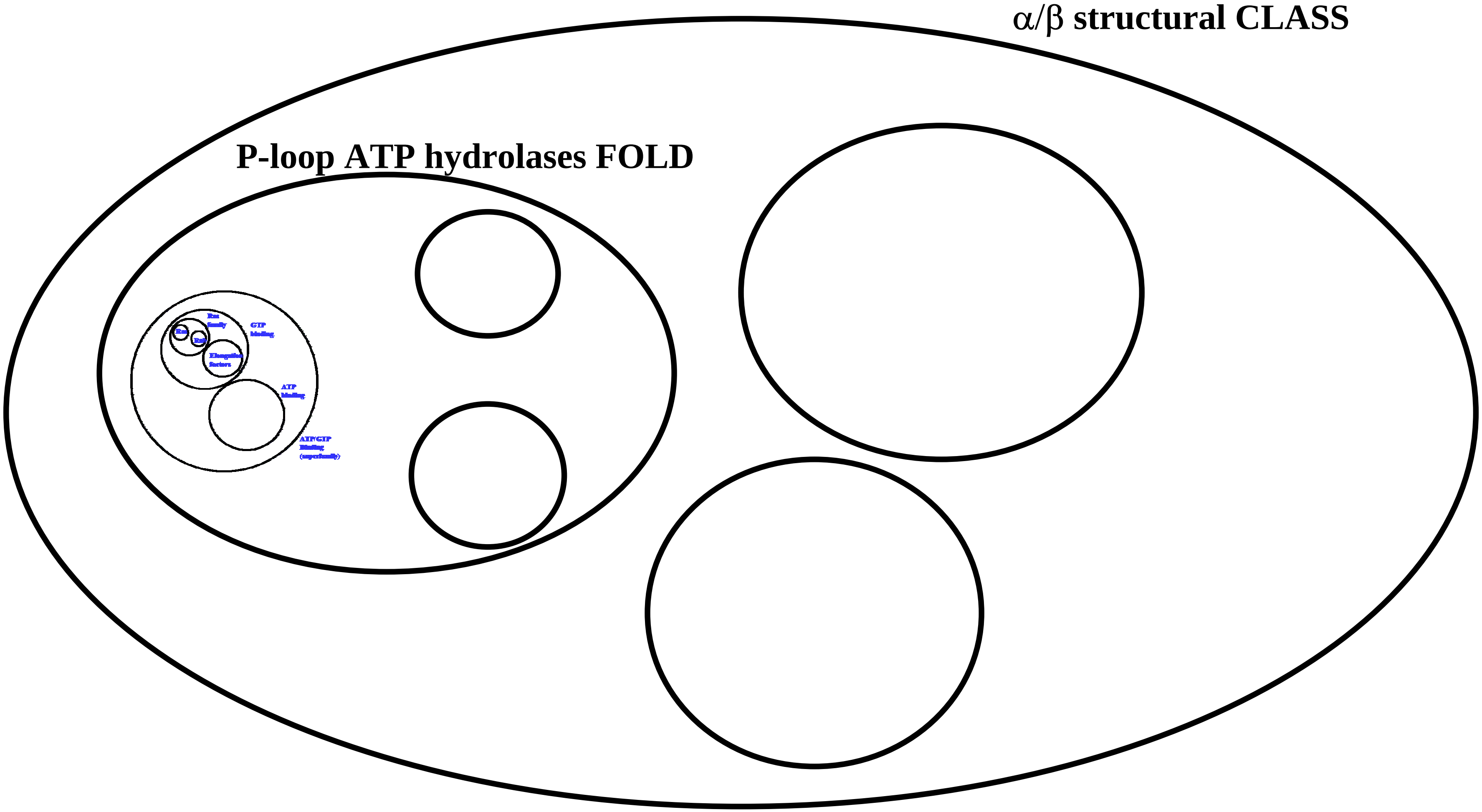


Chothia, C. & Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823-826.

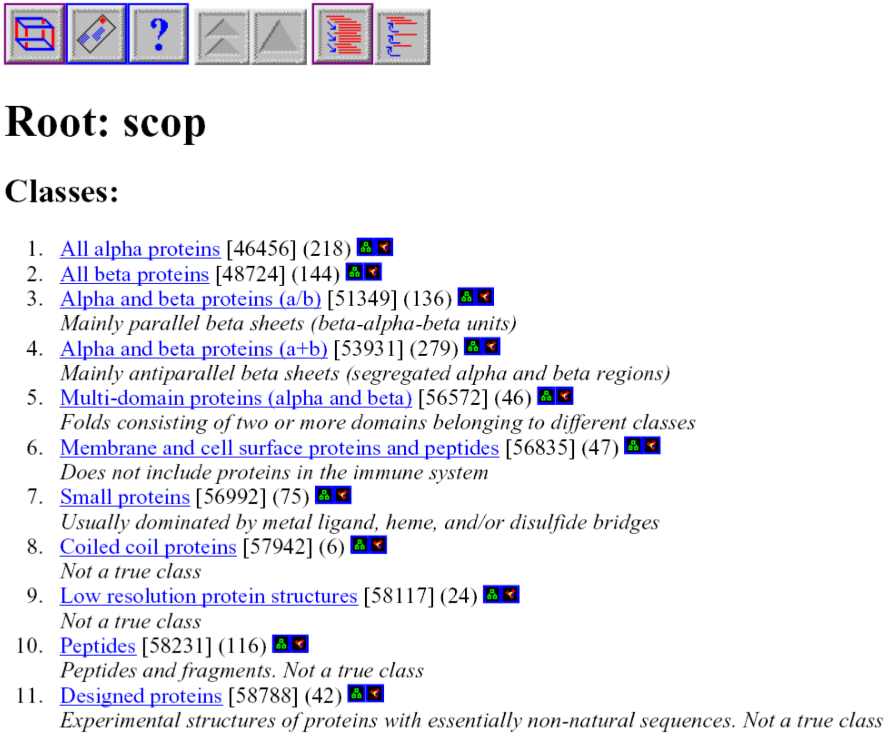
Characteristics of the structure space
Relationships with sequence space



Characteristics of the structure space



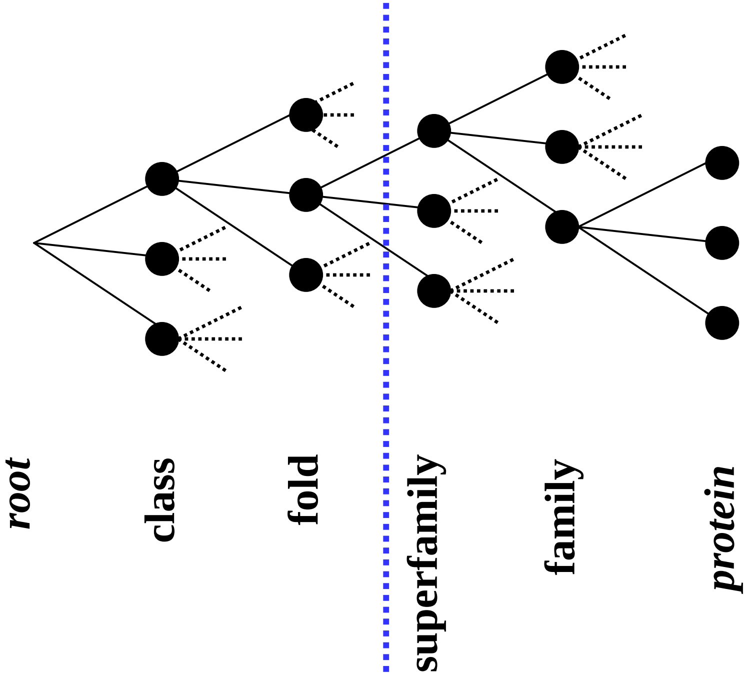
Structural Classifications
SCOP



Root: scop

Classes:

1. All alpha proteins [46456] (218)
2. All beta proteins [48724] (144)
3. Alpha and beta proteins (a/b) [51349] (136)
   Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) [53931] (279)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) [56572] (46)
   Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides [56835] (47)
   Does not include proteins in the immune system
7. Small proteins [56992] (75)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins [57942] (6)
   Not a true class
9. Low resolution protein structures [58117] (24)
   Not a true class
10. Peptides [58231] (116)
    Peptides and fragments. Not a true class
11. Designed proteins [58788] (42)
    Experimental structures of proteins with essentially non-natural sequences. Not a true class

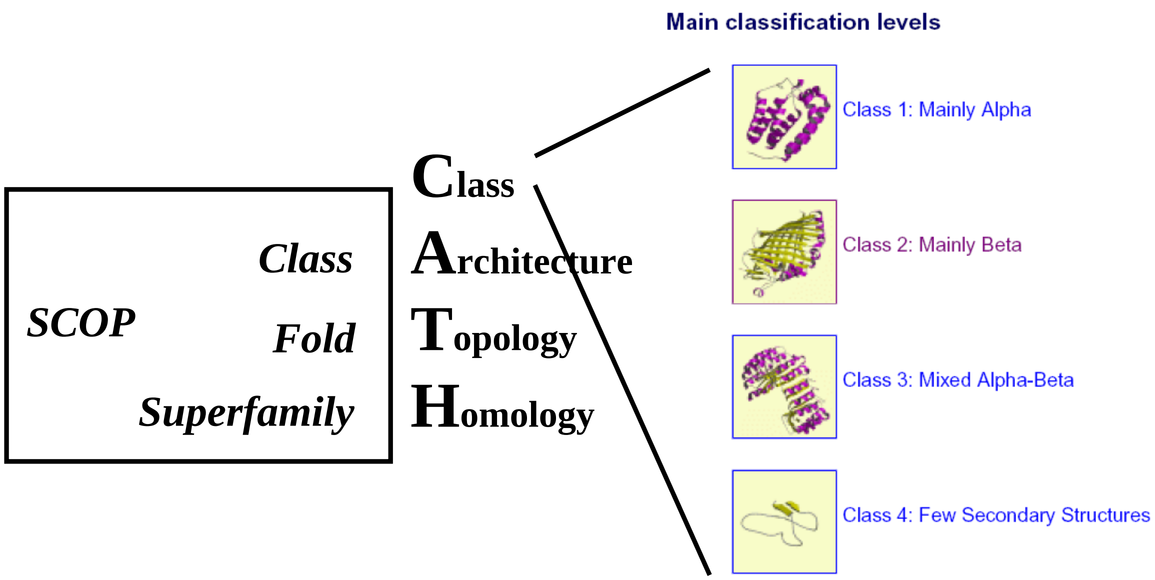


Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226-229.

http://scop.mrc-lmb.cam.ac.uk/scop/index.html

Structural Classifications

CATH



SCOP

Class

Fold

Superfamily

Class

Architecture

Topology

Homology

Main classification levels

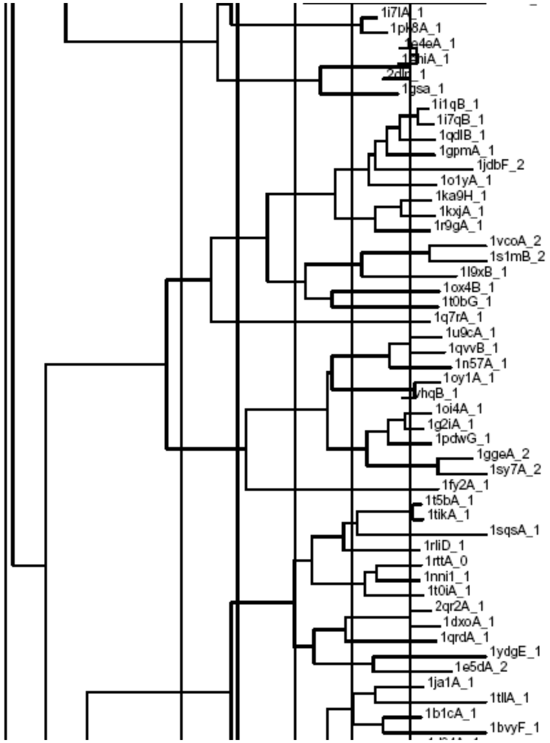
Class 1: Mainly Alpha

Class 2: Mainly Beta

Class 3: Mixed Alpha-Beta

Class 4: Few Secondary Structures

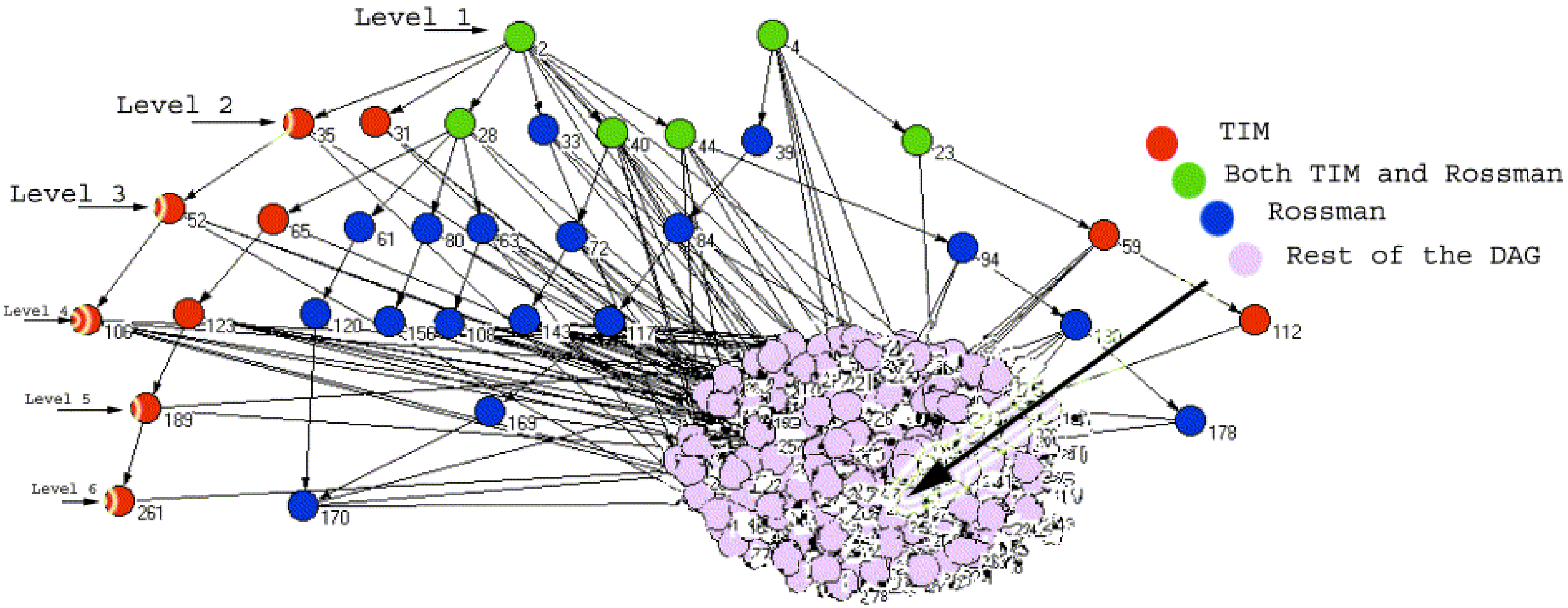
Dali/FSSP



Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247-251.

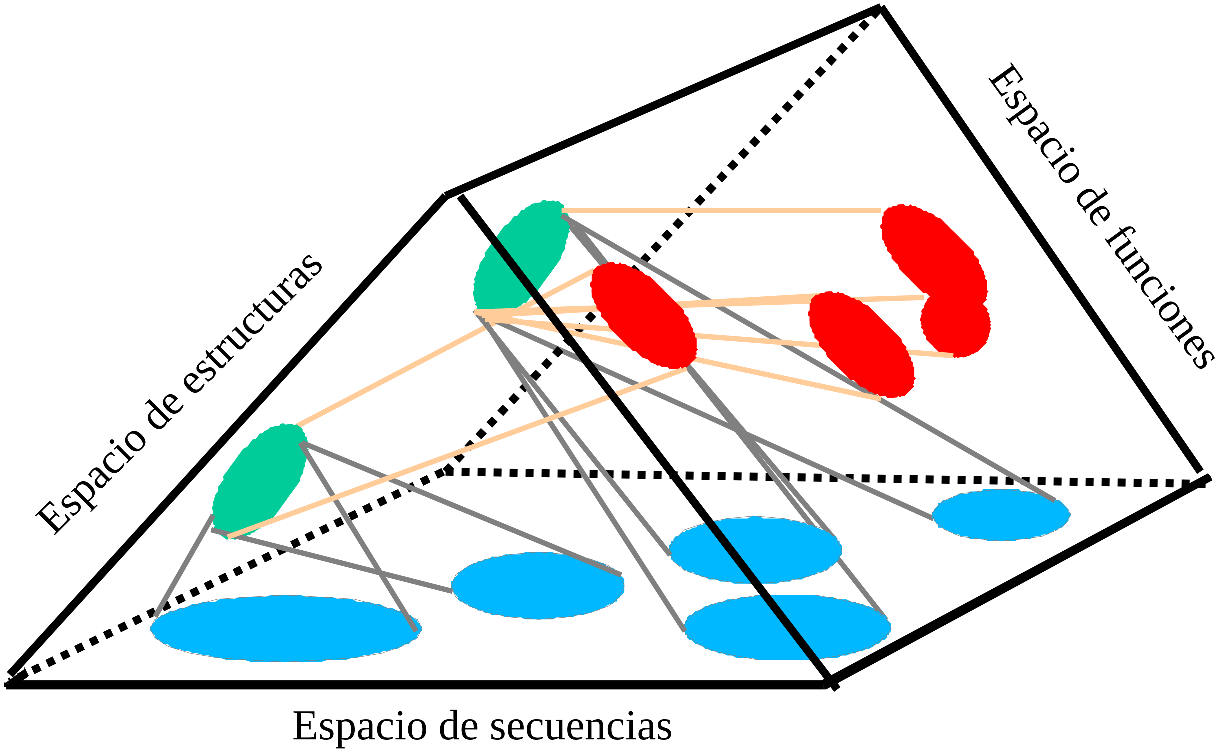
Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M. and Holm, L. (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.*, **29**, 55-57.

Relationship between structure and function spaces



Shakhnovich, B.E., Dokholyan, N.V., DeLisi, C. and Shakhnovich, E.I. (2003) Functional Fingerprints of Folds: Evidence for Correlated Structure-Function Evolution. *J Mol Biol*, **326**, 1-9.

Relationships between protein sequence structure and function spaces



The protein “Universe”

10^{400} Possible sequences

10^{10} Sequences in the biosphere

10^{5} Families

10^{3}-10^{4} Folds

10^{4} *Functions* (GO)

Vendruscolo, M. and Dobson, C.M. (2005) A glimpse at the organization of the protein universe. *Proc Natl Acad Sci U S A*, **102**, 5641-5642.