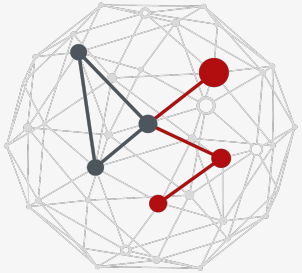
UNIVERSITÀ DEGLI STUDI DI PADOVA



DATA SCIENCE
UNIVERSITY OF PADOVA

SEQUENCE-STRUCTURE RELATIONSHIP

Master of Science in Data Science

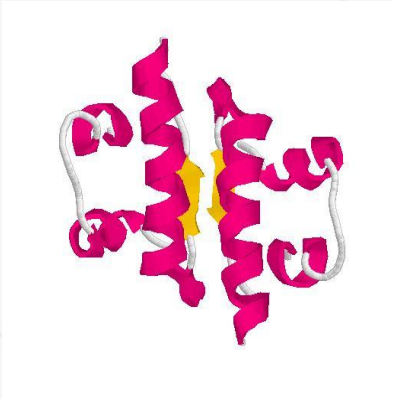
Damiano Piovesan

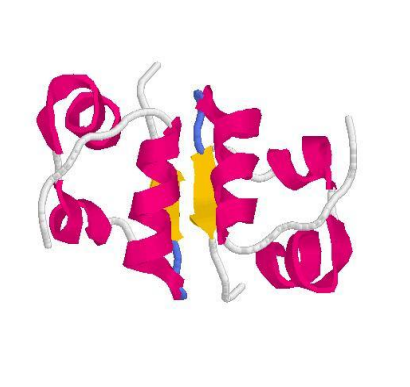
Same structure and high sequence similarity

- Human insulin (**1his**)
- Pig insulin (**3ins**)

- 91% sequence identity

```
sp|P01308|INS_HUMAN
sp|P01315|INS_PIG

MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED 60
MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAEN 60
**** ************.* ** ******************************:*****:

LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110
PQAGAVELGGGLGG--LQALALEGPPQKRGIVEQCCTSICSLYQLENYCN 108
 *.* ****** *.  **.*****. ************************
```



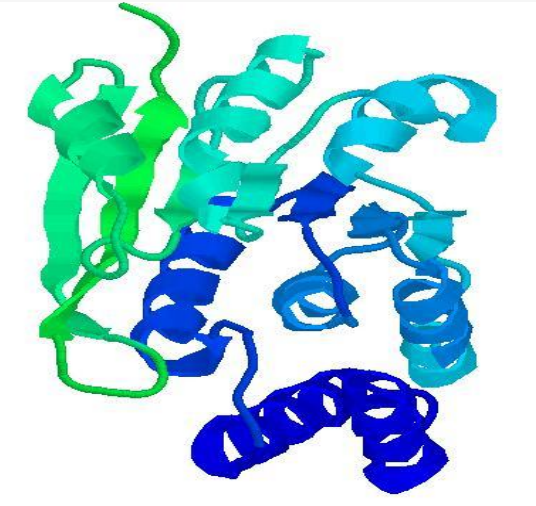


Same structure and low sequence similarity

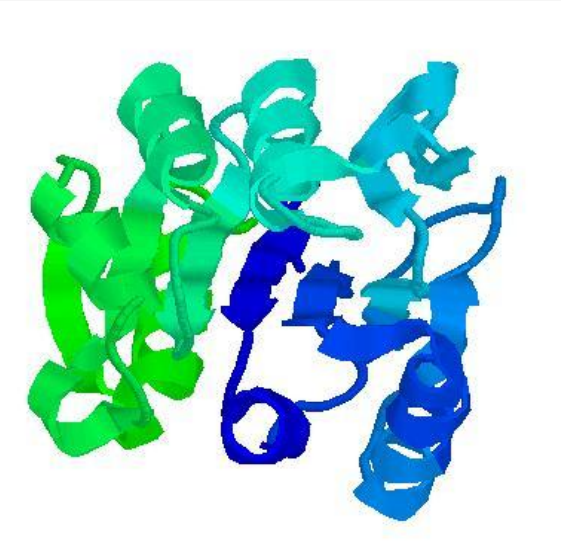
- **1vid** - Transferase (EC 2.1.1.6)
  - *Rattus norvegicus*
  - Inactivation of neurotransmitters
- **1chd** - Methylesterase (EC 3.1.1.61)
  - *Salmonella typhimurium*
  - Cell sensory response

```
1vid  TKEQRILRYVQQNAKPGDPQSVLEAIDTYCTQKEWAMNVGDAKGQIMDAVIREYSPSLVL
1chd  ..................................................llsseKLIA

1vid  ELGAYC.GYSAVRMARLLQ.PGARLLTMEMNP.DYAAITQQMLNFA.GLQD.........
1chd  IGAstggTEAIRHVLQPLPlSSPAVIITQHMPpGFTRSFAERLNKLcQISVkeaedgerv

1vid  ...KVTILN............................GASQDLIPQLKKKYDVDTLDMVF
1chd  lpgHAYIAPgdkhmelarsganyqikihdgppvnrhrPSVDVLFHSVAK..HAGRnAVGV

1vid  LDHWKDRYLPDTLLLEK.CGLLRKGTVLLADNVIVPGTPDFLAYVRGSSSFECTHYSSYL
1chd  ILTGMGN..dGAAGMLAmYQAG...aWTIAQNEA...................scvvfg

1vid  EYMKVVDGLEKAIYQGPSX..................
1chd  mpreainmggVSEVvdlsqvsqqmlakisagqairi
```

1vid



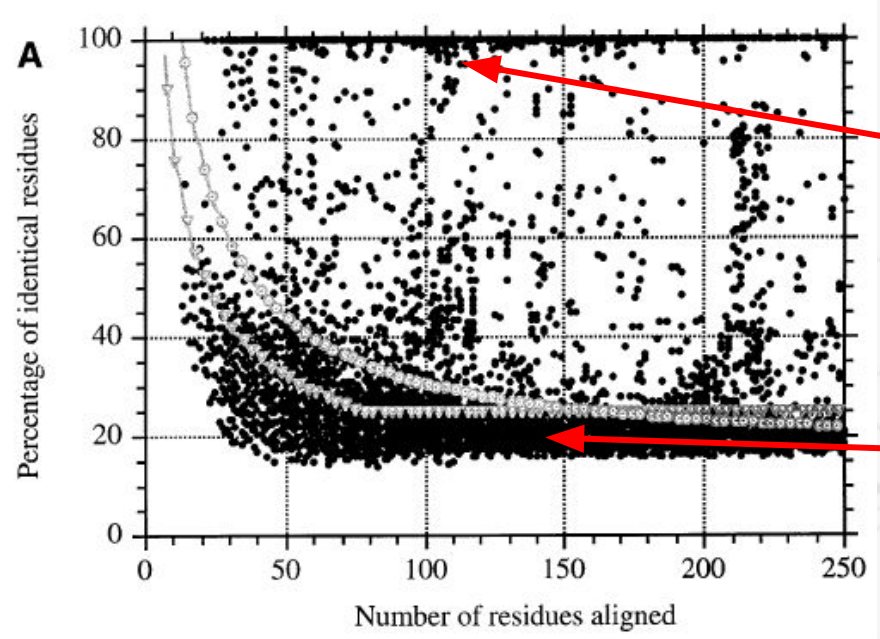
1chd

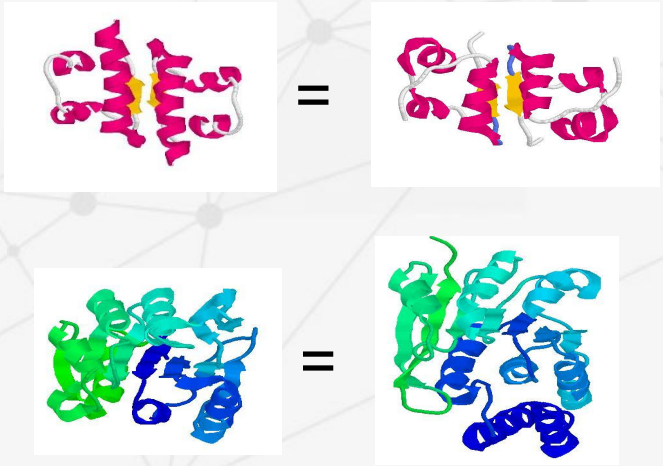


- Rossmann fold
- 10% sequence identity
- RMSD 3.0 Å for 104 out of 198 residues

Sequence similarity corresponds to structure similarity?

Pairs of proteins with **similar structure**

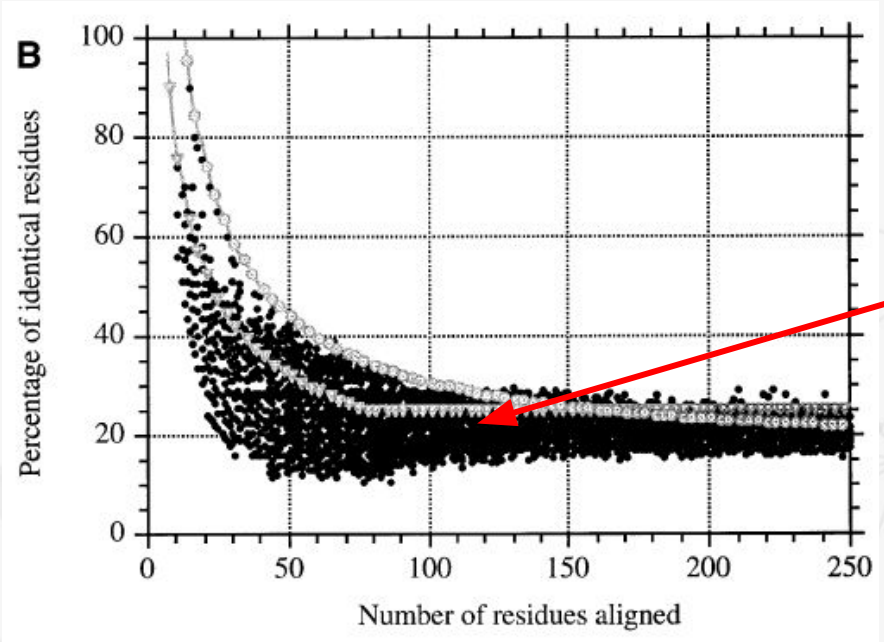




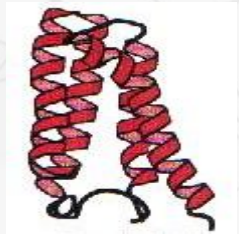
(Rost, 1999)

Sequence similarity corresponds to structure similarity?

Pairs of proteins with **different structure**



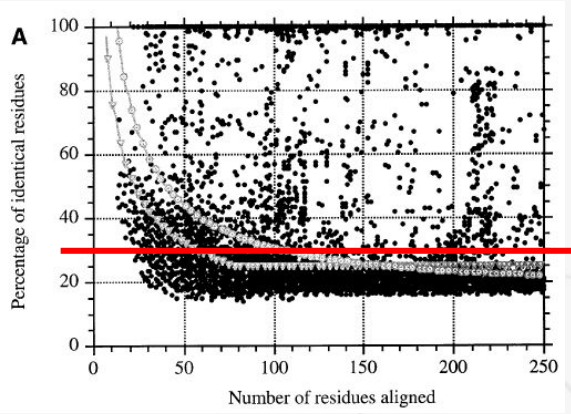
≠



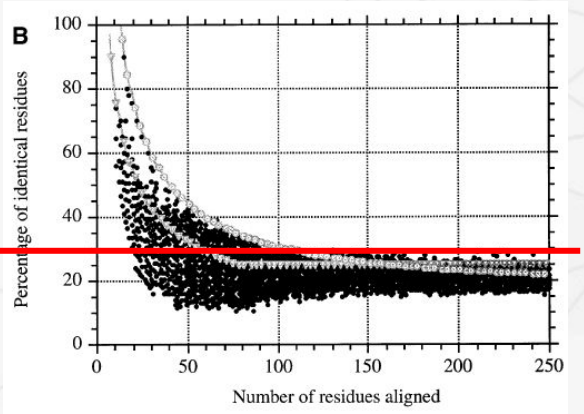
(Rost, 1999)

Sequence similarity corresponds to structure similarity?

Similar structure



Different structure



30%

- Any **random pair** of natural sequences have at least **15% sequence identity**
- Proteins with at least ca. **30% identical residues**, have the **same fold** (similar structure). For shorter alignments the threshold is higher
- In some cases proteins with less than **20% of sequence identity - "twilight zone" -** have the same fold

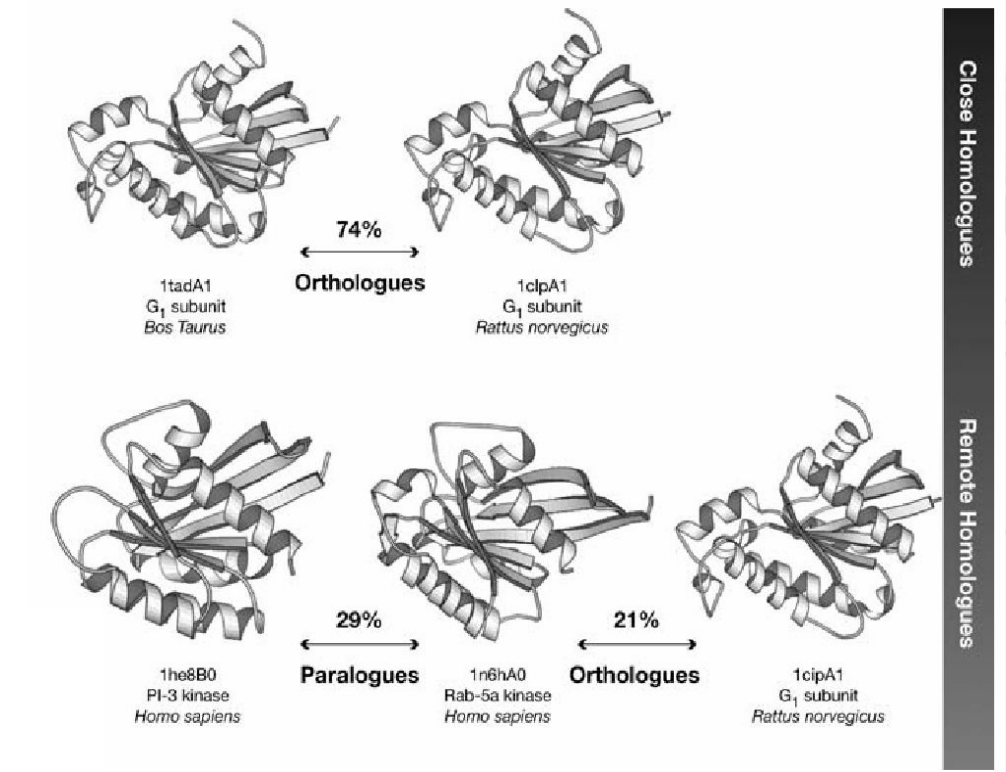
Homology

Orthologues

- Different species
- Same function
- Vertical descent

Paralogues

- Same species
- Similar (but different) function
- Horizontal evolution (duplication)



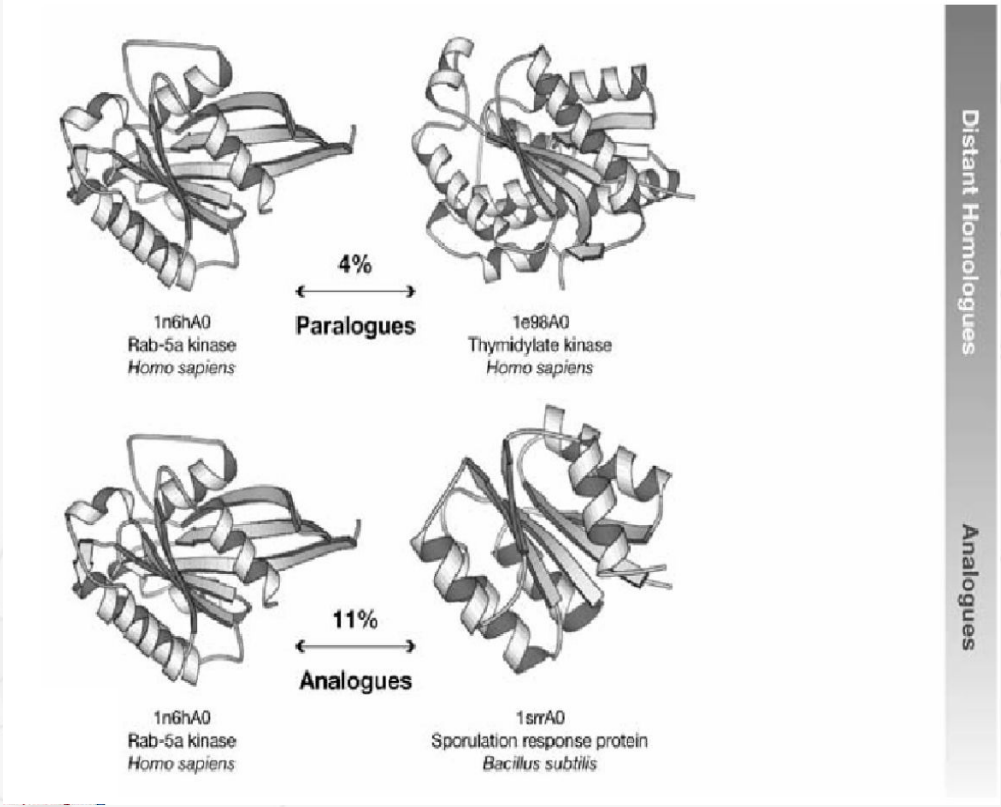
Homology

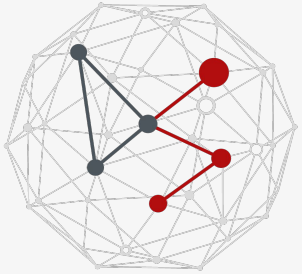
Remote homology

- Same structure
- Same ancestor
- Same function
- Low sequence identity

Analogues

- Same structure
- Different ancestor
- Same function ?





DATA SCIENCE
UNIVERSITY OF PADOVA

STRUCTURAL EVOLUTION

Master of Science in Data Science

Damiano Piovesan

Structural evolution

- How has protein structure evolved?
- Do current structures derive from a limited number of ancestral proteins?
- Can structure be seen as a footprint of the evolutionary path?

Topics to explore

- Inference from structural classification
- Hypotheses on common elements
- Theories on the origin of life

Structural complexity

DIPARTIMENTO MATEMATICA

Complexity

- Millions of species → each with thousand of coding genes (with different sequences)
- Point mutations, insertion, deletions

Mechanisms

- Random drift + natural selection
- Parental inheritance, acquisition, duplication

Observations

- Proteins display **substantial similarity** in sequence and 3D structure.
- Structures diverge much more slowly than sequences, providing evidence of **common ancestry**.

Why some folds are so common?

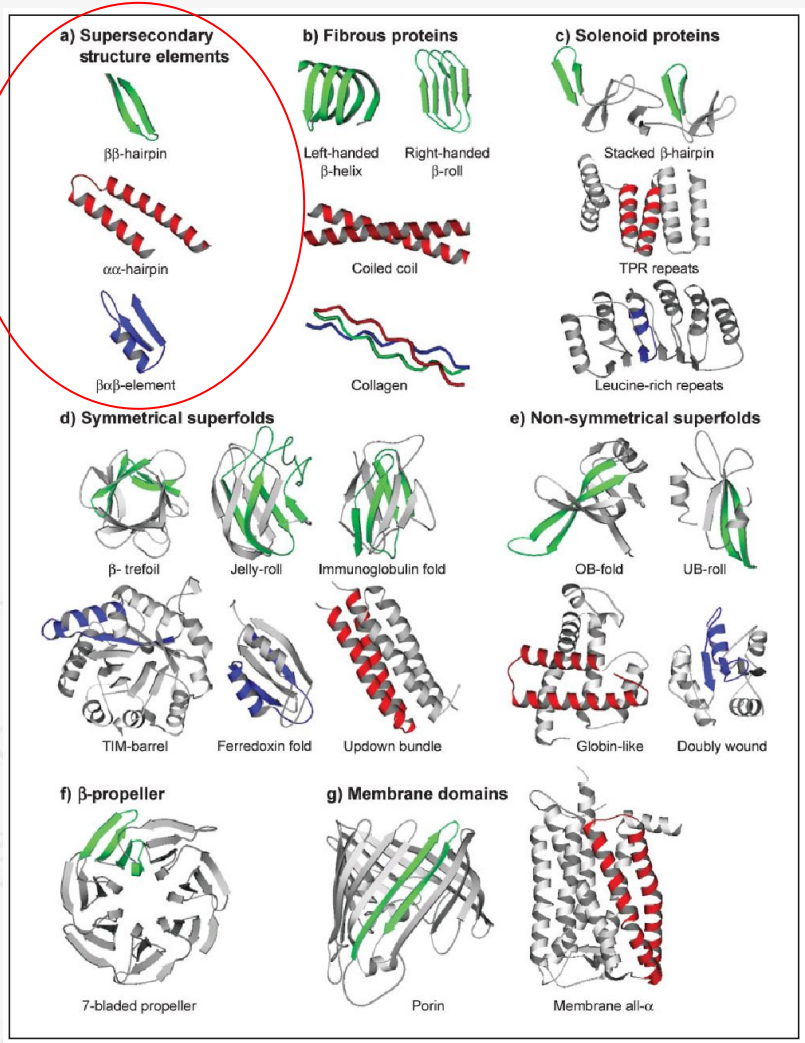
DIPARTIMENTO MATEMATICA

- Stability and folding efficiency
- Better scaffold for active sites (fold competition)
- Limited number of supersecondary structures

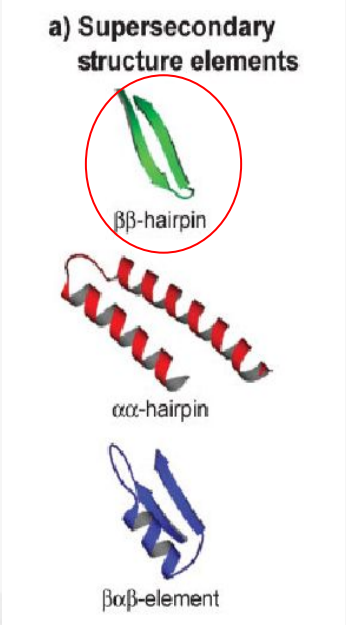
“More than the sum of their parts: on the evolution of proteins from peptides“

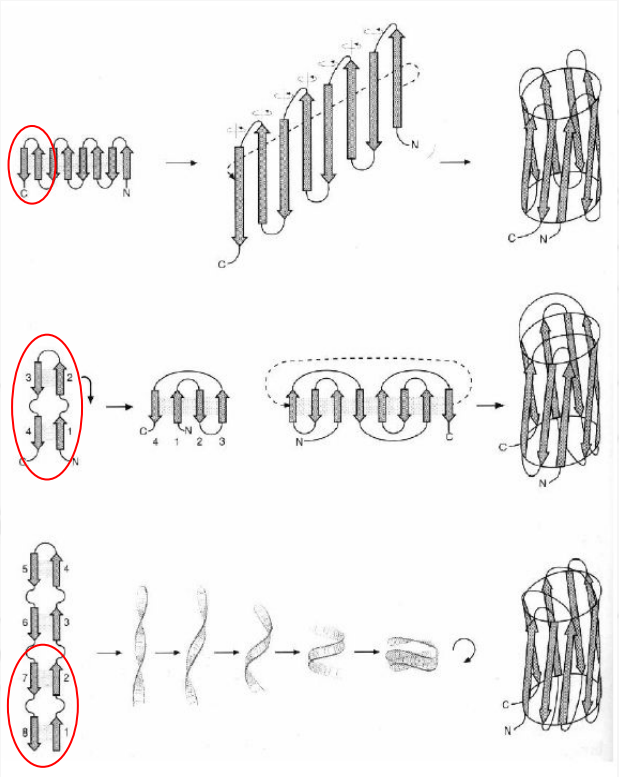
- There is a **basic complement** of autonomously folding units (**domains**)
- The complement was established at the time of the “**last universal common ancestor” (LUCA)**

(J. Söding & A. Lupas, BioEssays, 2003)



“More than the sum of their parts: on the evolution of proteins from peptides“



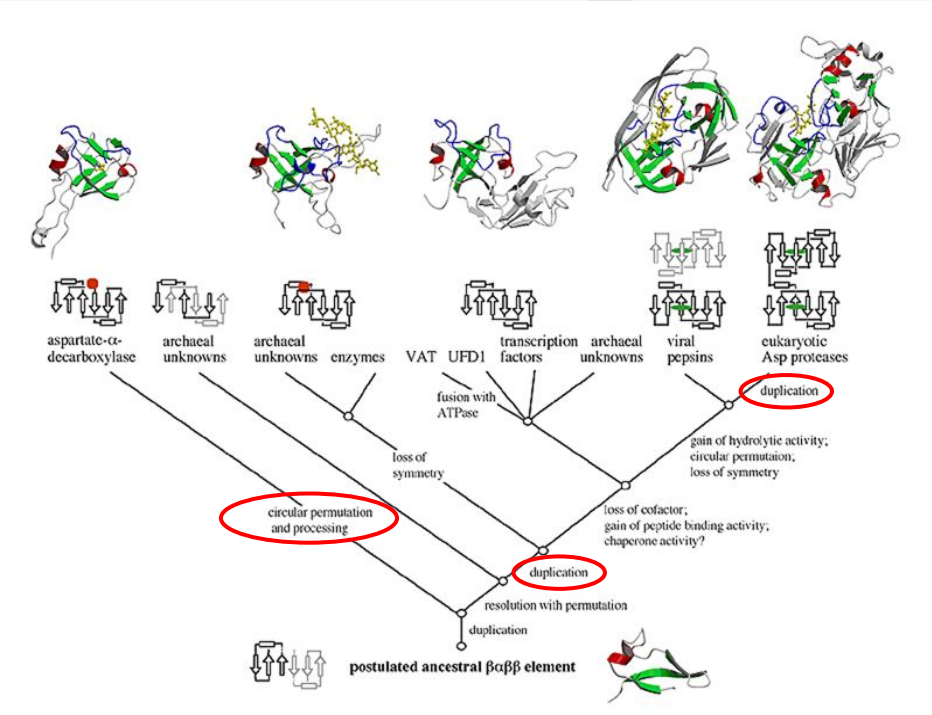


(J. Söding & A. Lupas, BioEssays, 2003)

Model of structural evolution

Common “operators”

- Oligomerization (repetition)
- Fusion
- Circular permutation
- Decoration

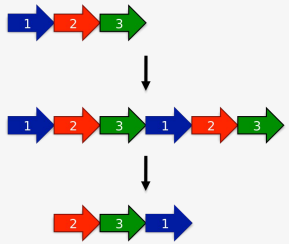


Circular permutations

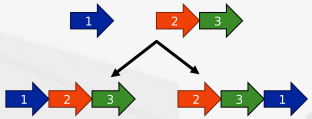
DIPARTIMENTO MATEMATICA

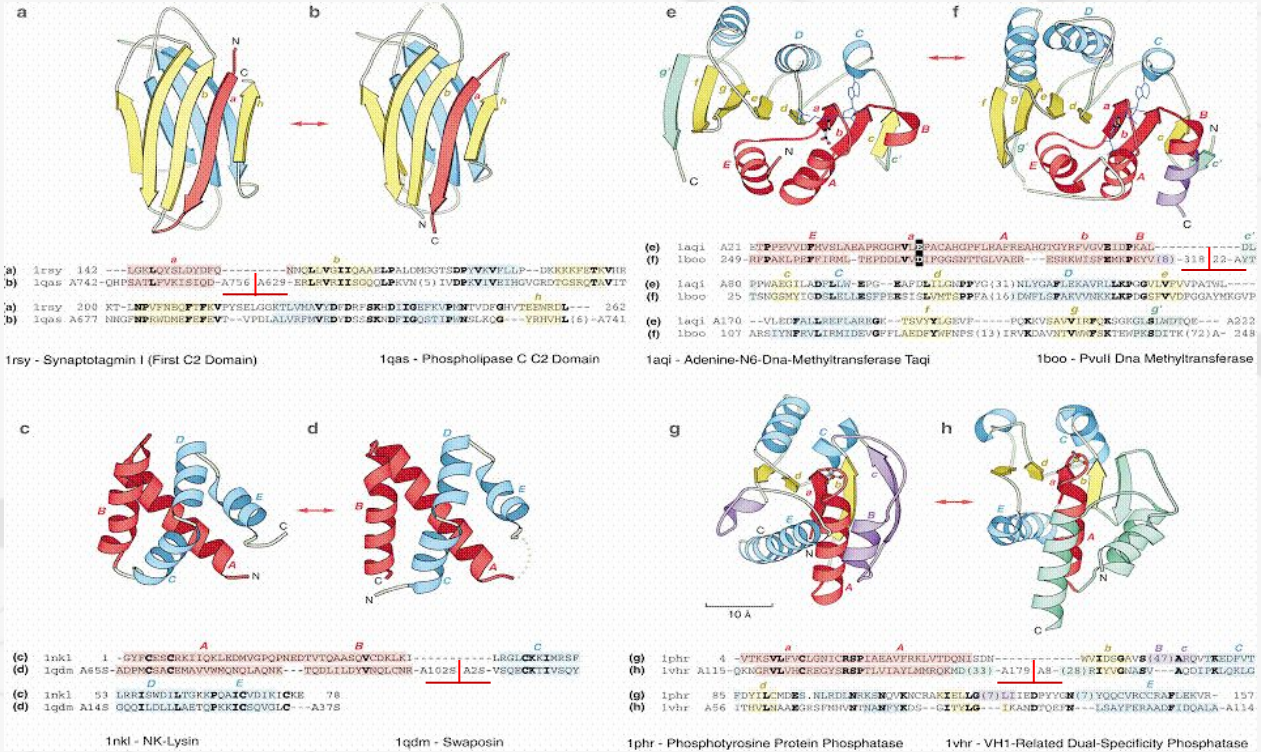
Close N- and C-termini can be found in different parts of the protein

- Duplication



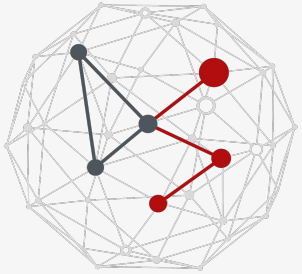
- Fission and fusion





Università degli Studi di Padova

DIPARTIMENTO MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

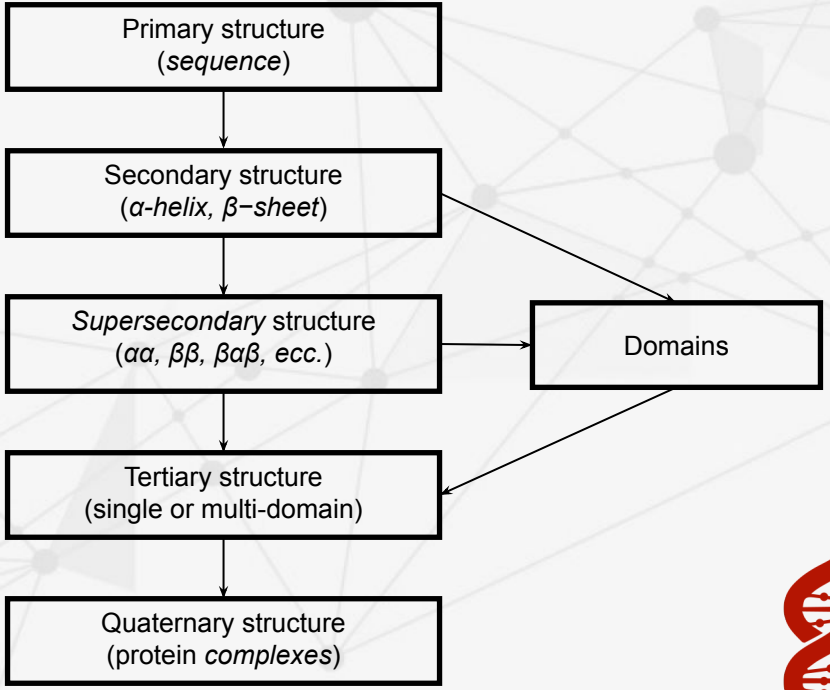
STRUCTURAL CLASSIFICATION

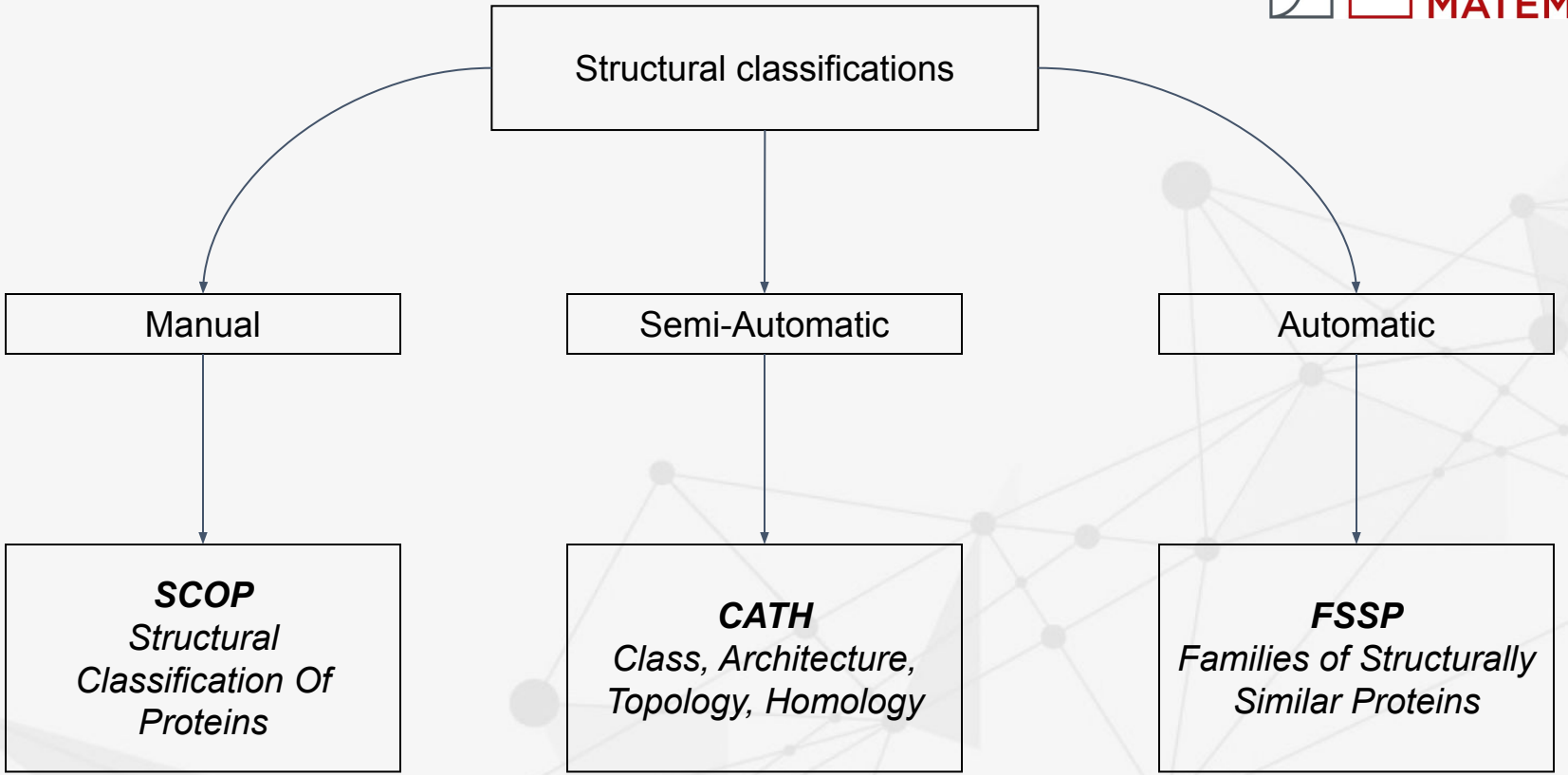
Master of Science in Data Science

Damiano Piovesan

Domain based classification

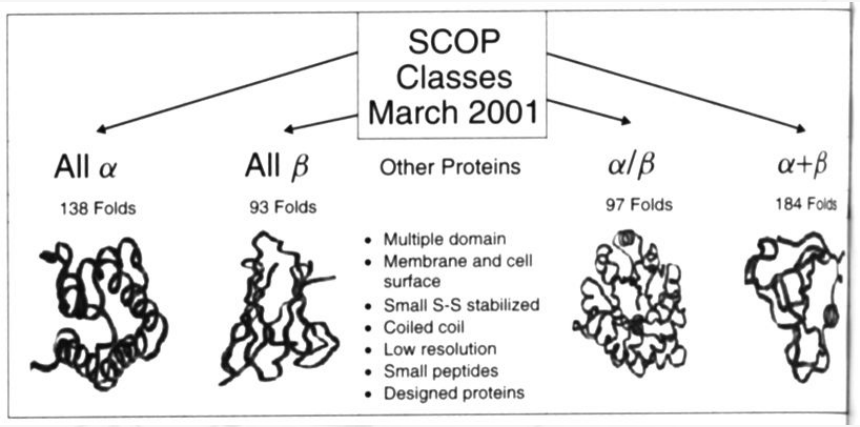
- A **domain** is a region of the protein that has its own **hydrophobic core** and has relatively little interaction with the rest of the protein making it **structurally independent**
- Recognition of **distant evolutionary events** make it possible to describe the basic **complement** of domains in the **last common ancestor**




Structural classifications
Manual
Semi-Automatic
Automatic
SCOP
Structural Classification Of Proteins
CATH
Class, Architecture, Topology, Homology
FSSP
Families of Structurally Similar Proteins

SCOP - Structural Classification of Proteins

- Class
  - α, β, α/β, α+β, ...
- Fold
  - Structural similarity - *Convergent evolution*
- Superfamily
  - Homology - *Divergent evolution*
- Family
  - Homology and function



Created by Alexey Murzin

Mainly manually curated

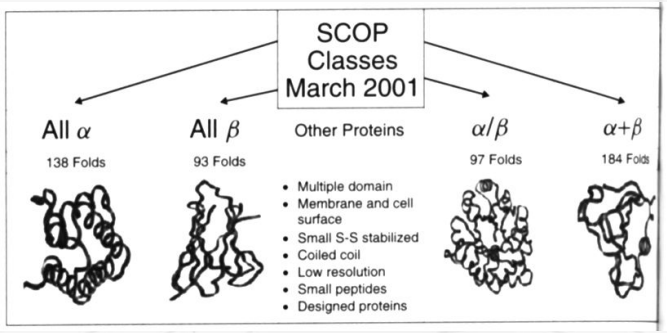
Represent the gold standard for fold classification

SCOP - Class

Initially a protein structure is classified based on its domains

Domain types

- mainly α
- mainly β
- α / β → the β-sheet and α-helices are mixed (typically β-strands connected by α-helices)
- α + β → domains that have the α and β units largely separated in sequence
- Multidomain
- Membrane and cell surface
- Small proteins



SCOP - Families and Superfamilies

DIPARTIMENTO MATEMATICA

- **Superfamilies**
  - Share a **common fold**, perform **similar functions**, usually **low sequence identity**
  - A strong functional relationship (eg the conserved interaction with substrate or cofactor molecules) can compensate for a different fold (provided it includes the active site)
- **Families**
  - **Sequence identity >30%** or functions and structures are very similar
  - Common evolutionary origin
- **“Strange” families**
  - Sequence similarity below family definition but above the superfamily level
  - Similar **domain organization**, common fold in the **catalytic domain** → likely to be closely related

SCOP website

DIPARTIMENTO MATEMATICA

Structural Classification of Proteins

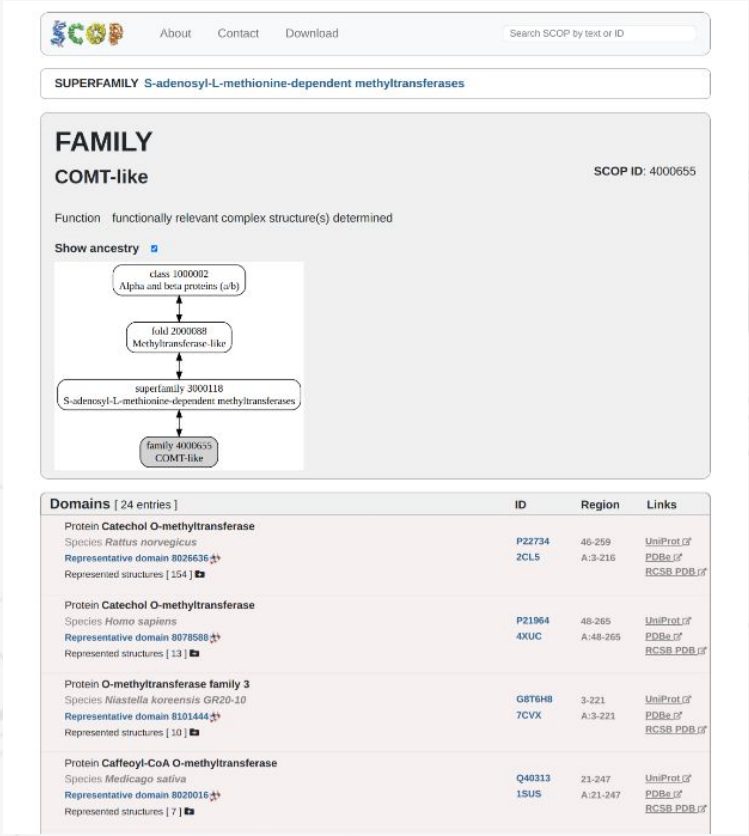
Protein: Catechol O-methyltransferase, COMT from Rat (Rattus norvegicus) [TaxId: 10116]

Lineage:

1. Root: scop
2. Class: Alpha and beta proteins (a/b) [51349]
   Mainly parallel beta sheets (beta-alpha-beta units)
3. Fold: S-adenosyl-L-methionine-dependent methyltransferases [53334]
   core: 3 layers, a/b/a; mixed beta-sheet of 7 strands, order 3214576; strand 7 is antiparallel to the rest
4. Superfamily: S-adenosyl-L-methionine-dependent methyltransferases [53335]
   Superfamily
5. Family: COMT-like [53336]
6. Protein: Catechol O-methyltransferase, COMT [53337]
7. Species: Rat (Rattus norvegicus) [TaxId: 10116] [53338]

PDB Entry Domains:

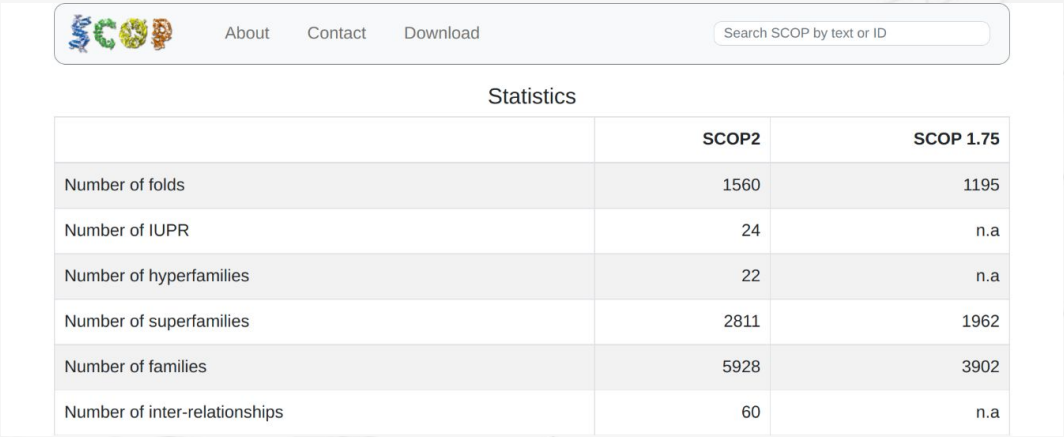
1. 2cl5
   *automatically matched to d1h1da_*
   complexed with bie, bu3, mes, mg, sam
   1. region a:3-216 [130570]
2. 2cl5
   *automatically matched to d1h1da_*
   complexed with bie, bu3, mes, mg, sam
   1. region b:3-215 [130571]
3. 1h1d
   complexed with bia, mg, sam
   1. chain a [83452]
4. 1vid
   complexed with dnc, mg, sam
   1. chain a [34178]
5. 2zlb
   *automatically matched to d1h1da_*
   complexed with so4
   1. region a:3-214 [154628]
6. 1jr4
   complexed with cl4, mg
   1. chain a [71820]



https://scop.mrc-lmb.cam.ac.uk/

Folds in SCOP

- The most difficult stage of classification
- Same major secondary structures, same arrangement, same topological connections
- Peripheral elements of secondary structure and turn regions may differ in size and conformation
- Useful to infer evolutionary relationship for distant homologs



	SCOP2	SCOP 1.75
Number of folds	1560	1195
Number of IUPR	24	n.a
Number of hyperfamilies	22	n.a
Number of superfamilies	2811	1962
Number of families	5928	3902
Number of inter-relationships	60	n.a

Folds

- Estimated total **folds** 10,000
- A quarter of domains is inside **superfolds**
- 80% of all domains is inside 400 **mesofolds**
- The rest are called **unifolds**

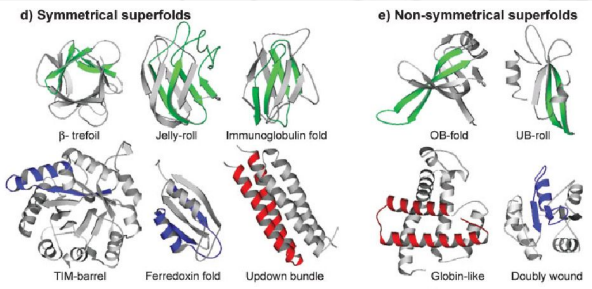
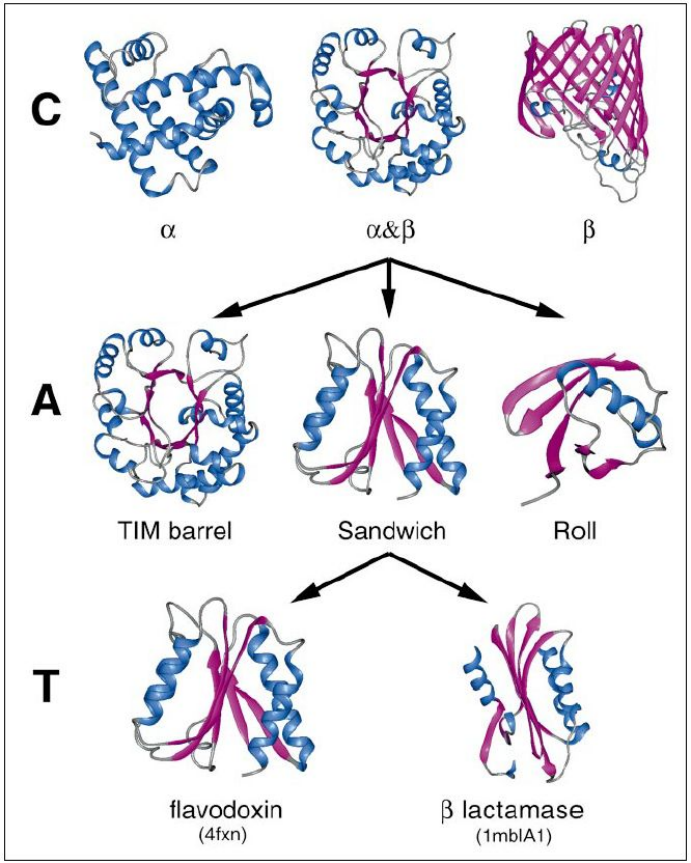


Table 1. Superfolds and the fraction of their residues contained in the supersecondary structure elements αα, ββ, βαβ[21]

Fold	Internal symmetry: Sequence*	Internal symmetry: Structure	Number of superfamilies (%)†	% Supersecondary structure content
β-trefoil	+	+	2 (0.1)	83
Jelly roll	–	+	17 (1.2)	47
Immunoglobin-like	–	+	55 (4.0)	67
TIM-barrel	+	+	28 (2.0)	82
Ferredoxin-like	+	+	65 (4.7)	38
Updown bundle	+	+	17 (1.2)	90
OB fold	–	–	16 (1.1)	77
UB-roll	–	–	16 (1.1)	55
Globin-like	–	–	4 (0.3)	88
Doubly wound	–	–	122 (8.8)	68
All superfolds			342 (24.7)	65
All folds			1386 (100)	62

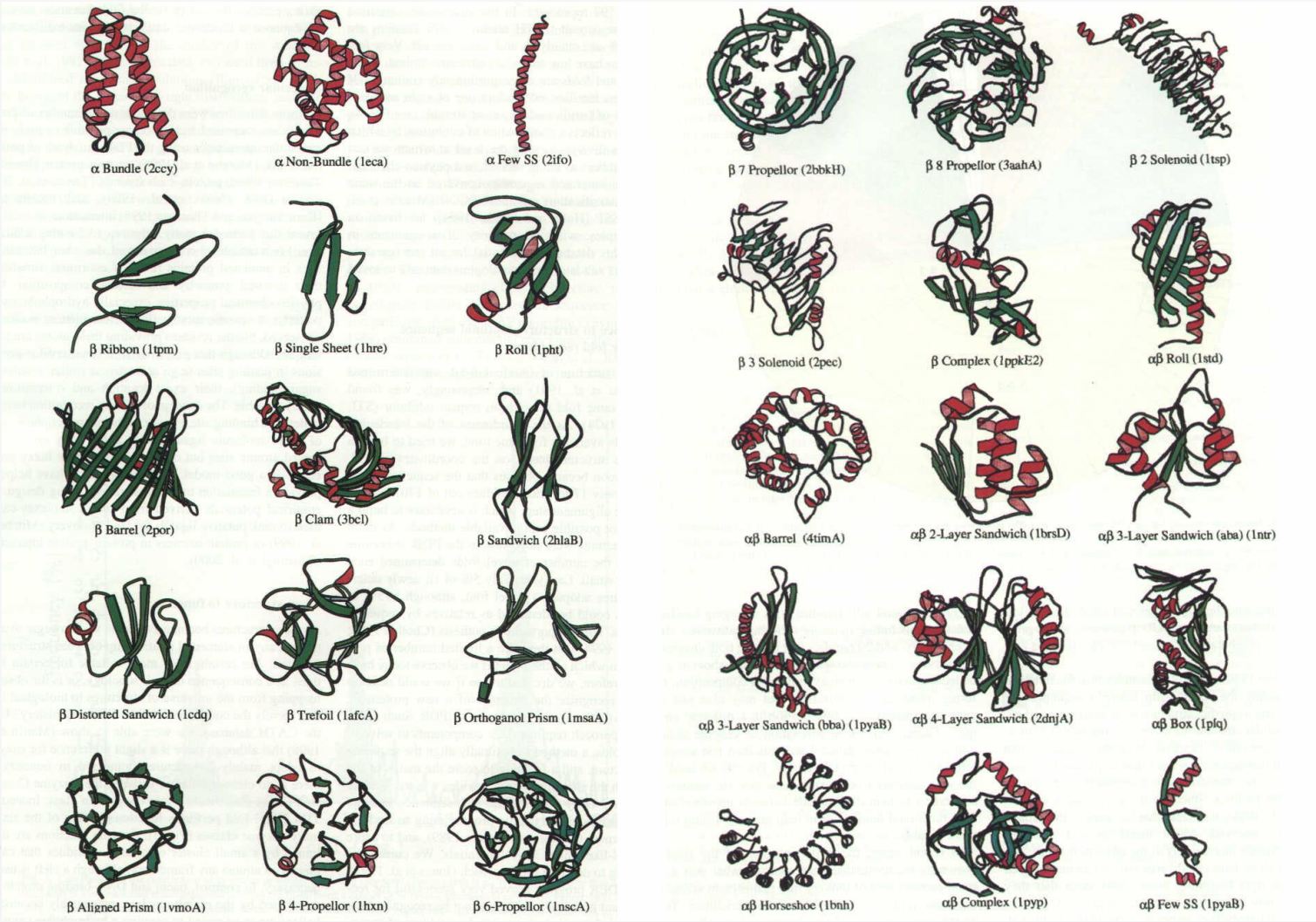
CATH



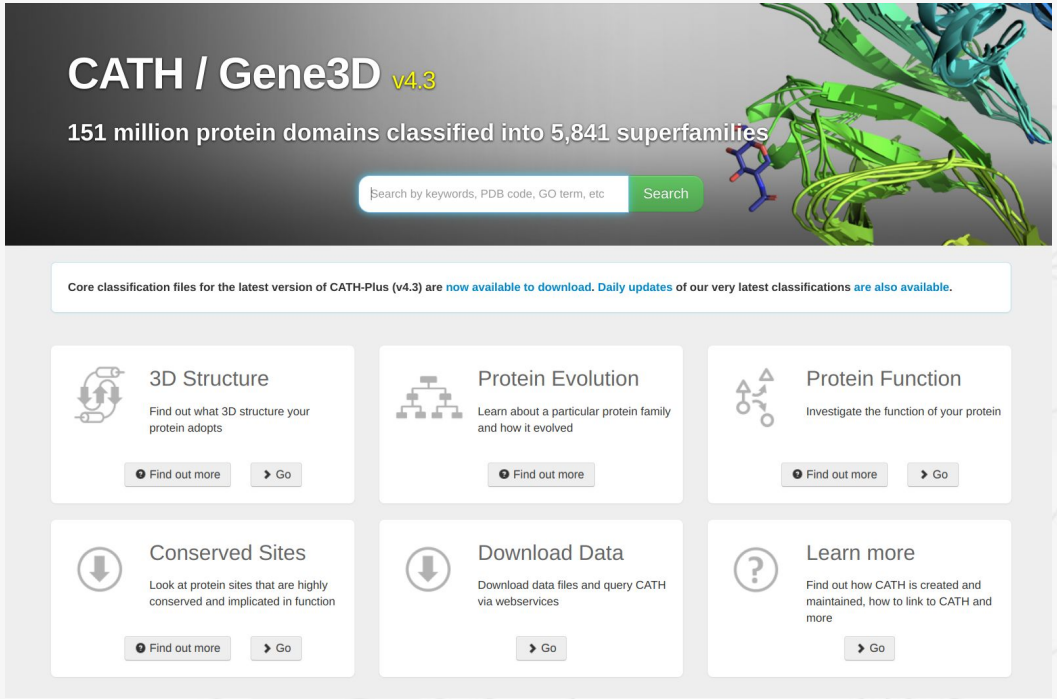
Semi-Automatic, only Architectures are manually assigned

- **Class** → secondary structure **content**
  - mainly-alpha, mainly-beta, mixed alpha/beta, “few secondary structures”
- **Architecture** → general **arrangement of the secondary structures** irrespective of connectivity between them
  - Eg. alpha/beta sandwich
- **Topology (fold)** → **connectivity** of secondary structures in the chain
- **Homologous Superfamily** → domains (believed to be) related by a **common ancestor**

CATH architectures

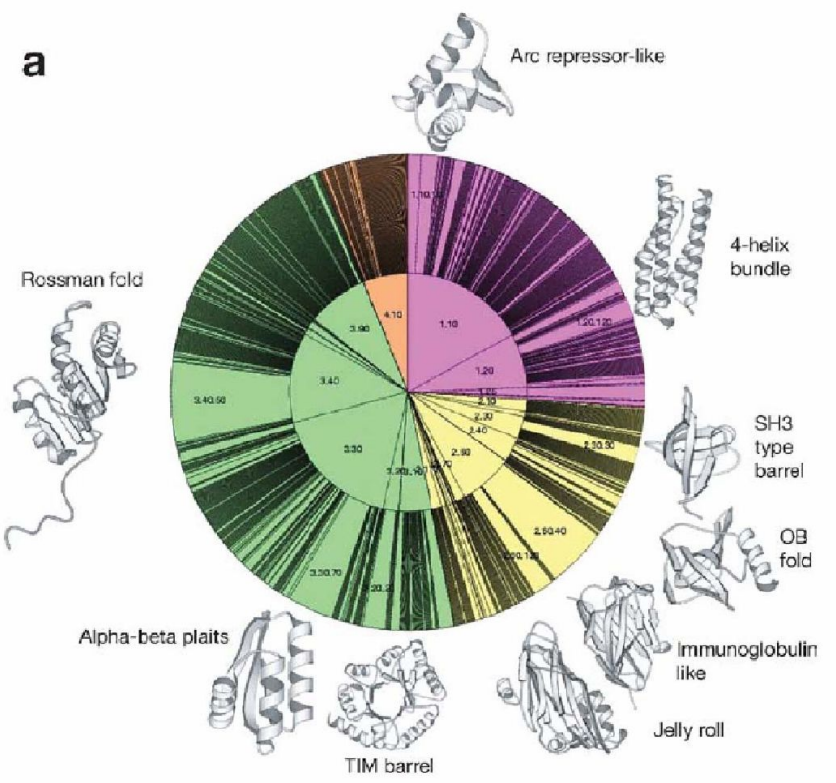


CATH website



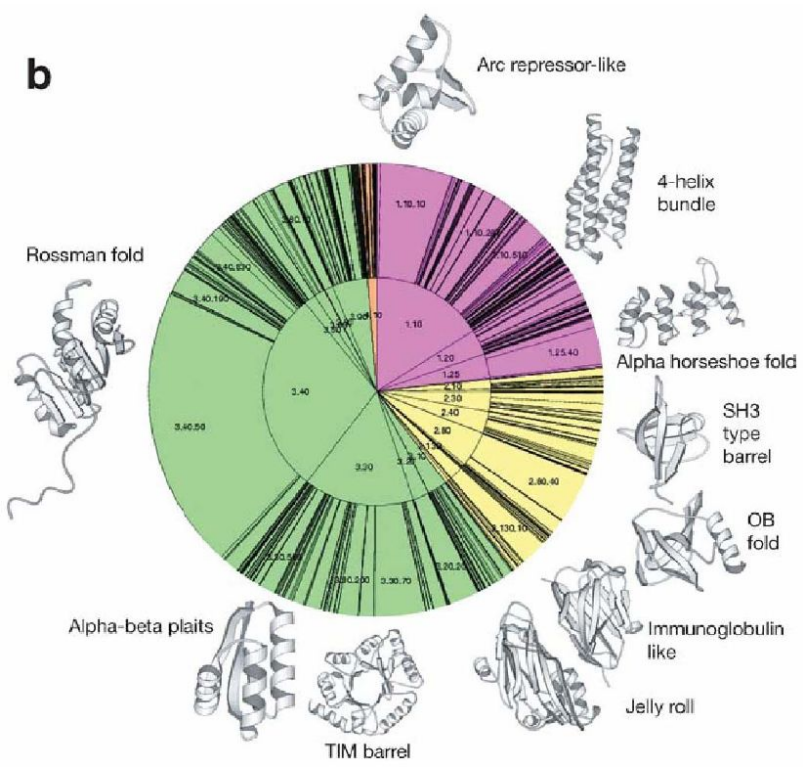
http://www.cathdb.info/

CATH hierarchy in the PDB



- Wheels
  - Inner → architectures
  - Outer → folds
- Classes
  - Pink → mainly *α*
  - Yellow → mainly *β*
  - Green → *α*-*β*
- Slice size proportional to the number of folds and superfamilies

CATH hierarchy in 150 genomes - Gene3D



Gene3D

HMM models of CATH families