1222·2022 800 ANNI

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO MATEMATICA

DATA SCIENCE
UNIVERSITY OF PADOVA

SEQUENCE-STRUCTURE RELATIONSHIP

Master of Science in Data Science

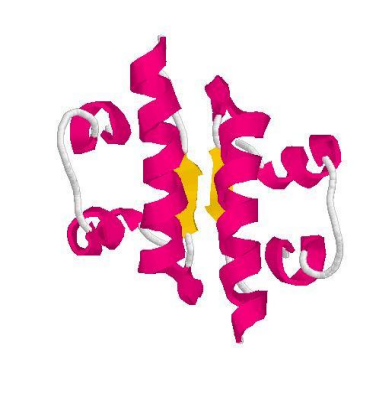
Damiano Piovesan

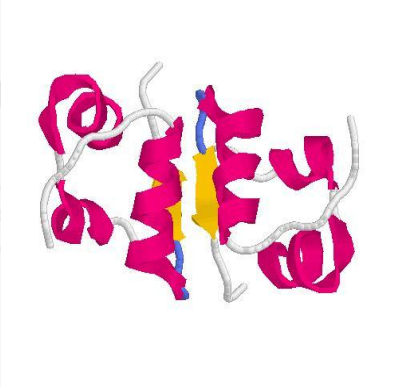
Same fold and high sequence similarity

- 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
 *.* ****** *.  **.*****. ************************
```



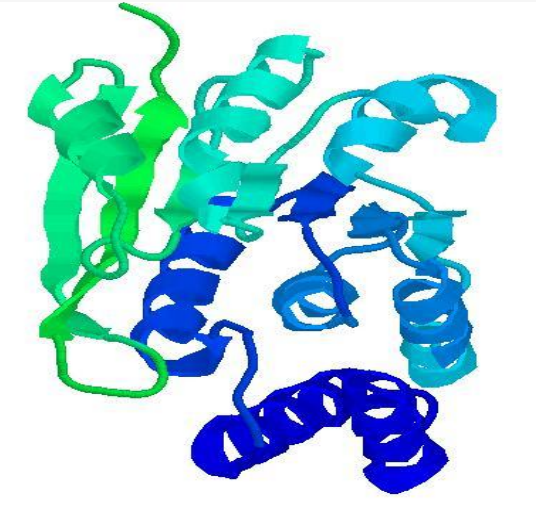


Low sequence similarity but same fold

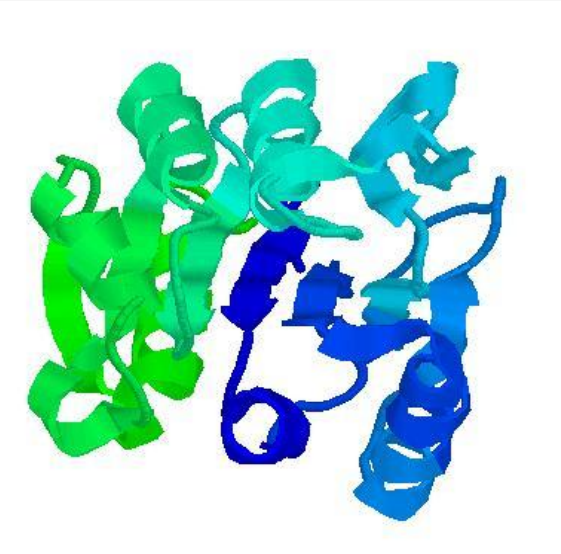
- **1vid** - Transferase (EC 2.1.1.6)
  - *Rattus norvegicus*
  - Inactivation of neurotransmitters
- **1chd** - Methylesterase (EC 3.1.1.61)
  - *Salmonella typhimurium*
  - Cell sensory response

```
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



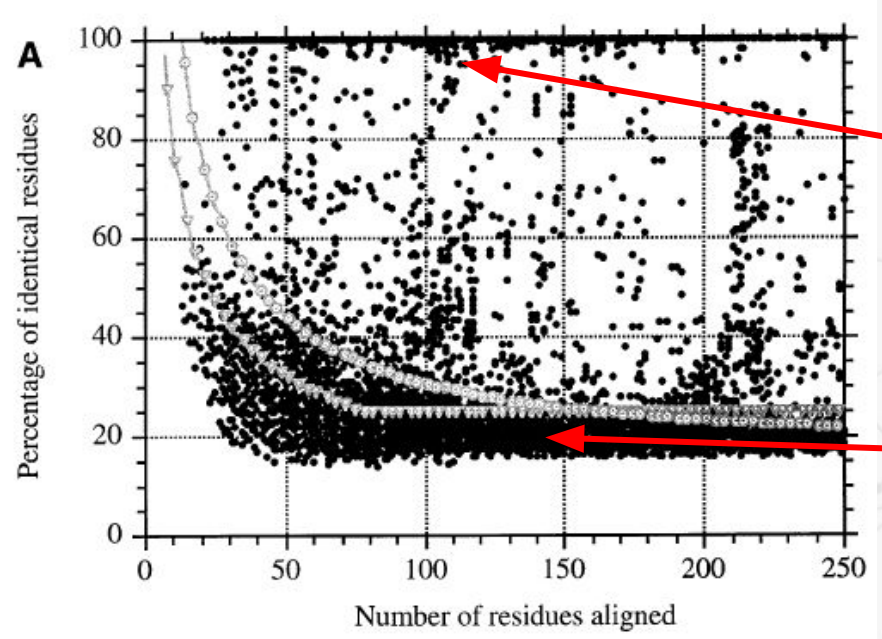
1chd

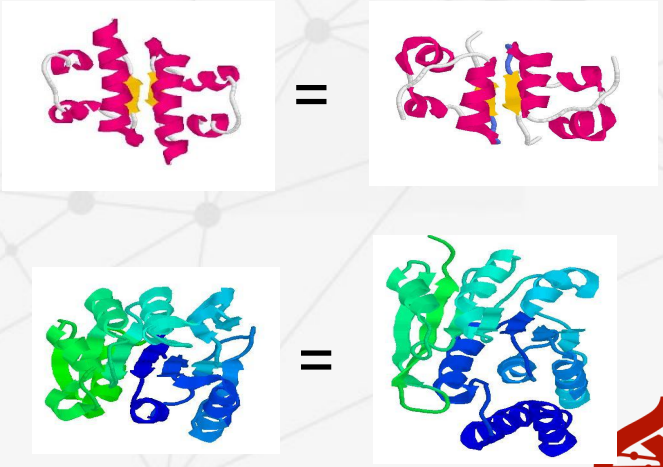


- Rossmann fold
- 10% sequence identity
- RMSD 3.0 Å for 104 out of 198 residues

Sequence similarity == Structure similarity ?

Pairs of proteins with **similar structure**

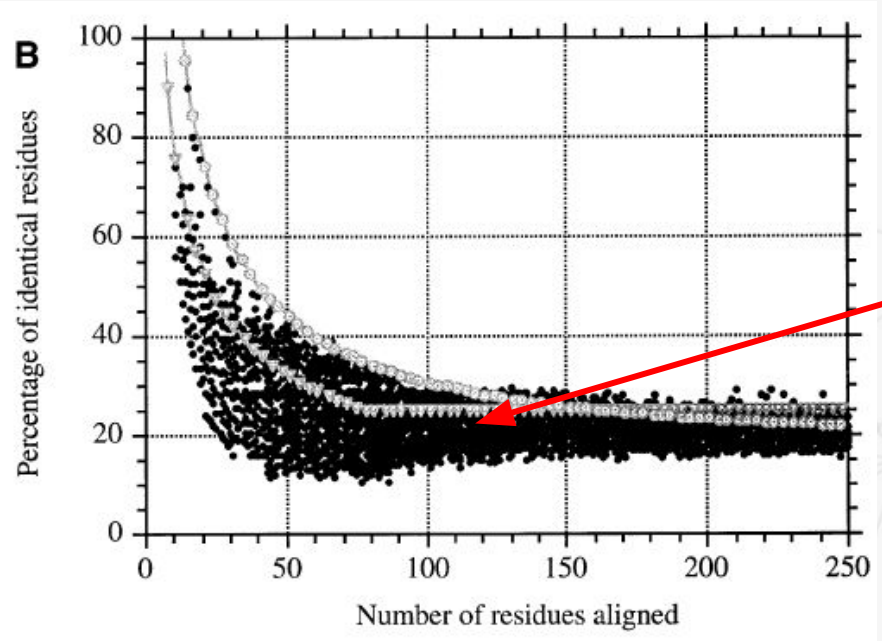




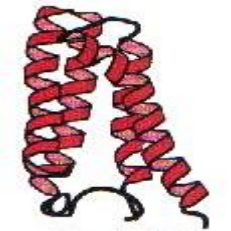
(Rost, 1999)

Sequence similarity == Structure similarity ?

Pairs of proteins with **different structure**



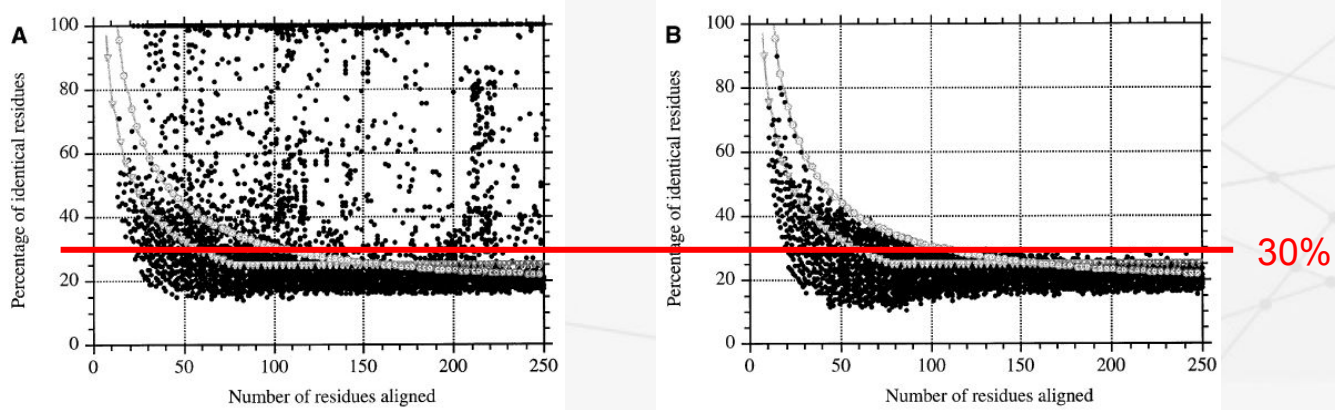
≠



(Rost, 1999)

Sequence similarity – Structure similarity?

Similar structure **Different structure**



- Proteins with at least ca. **30% identical residues**, likely 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
- Any pair of sequences have at least 15% sequence identity

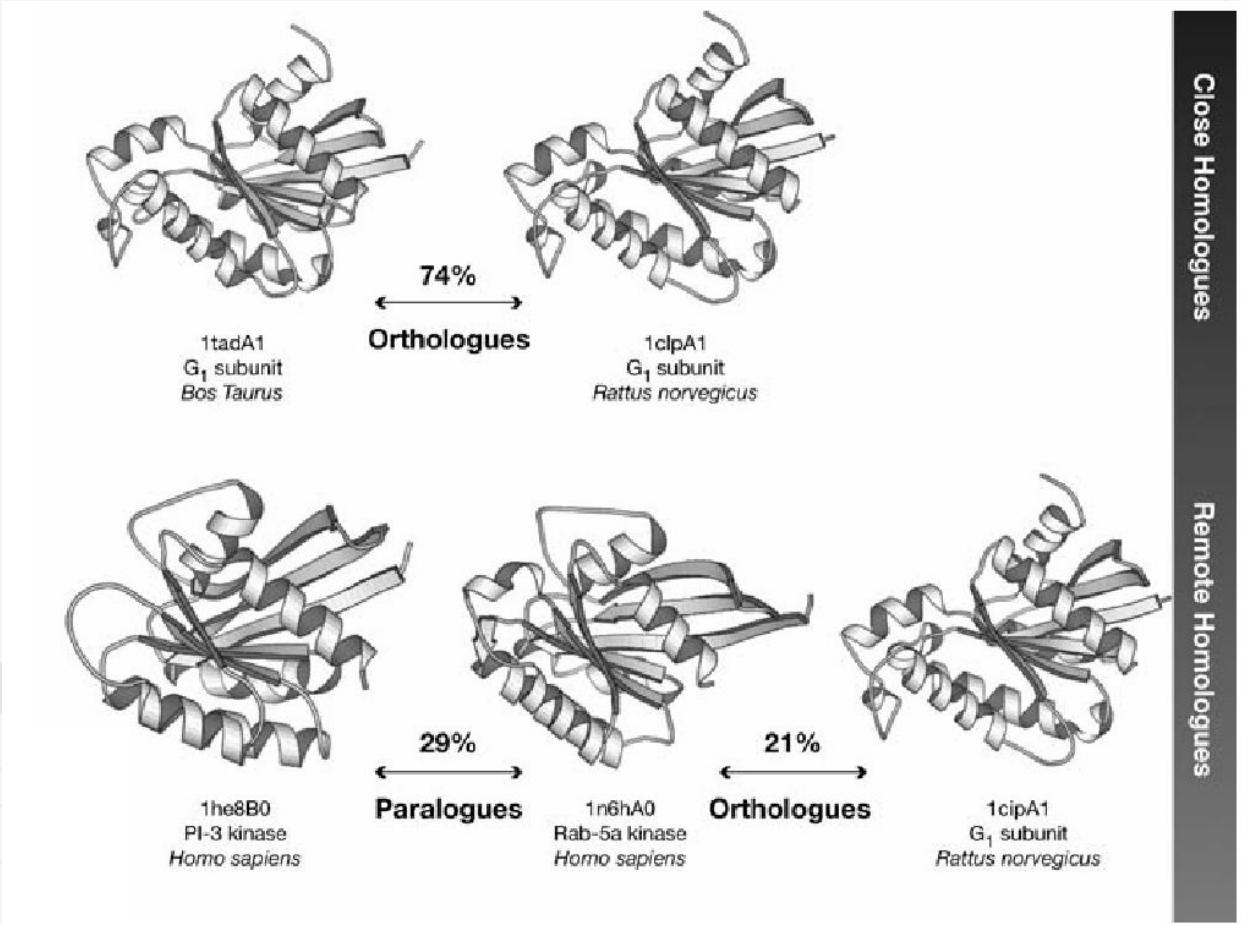
Homology

Orthologues

- Different species
- Same function
- Vertical descent

Paralogues

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



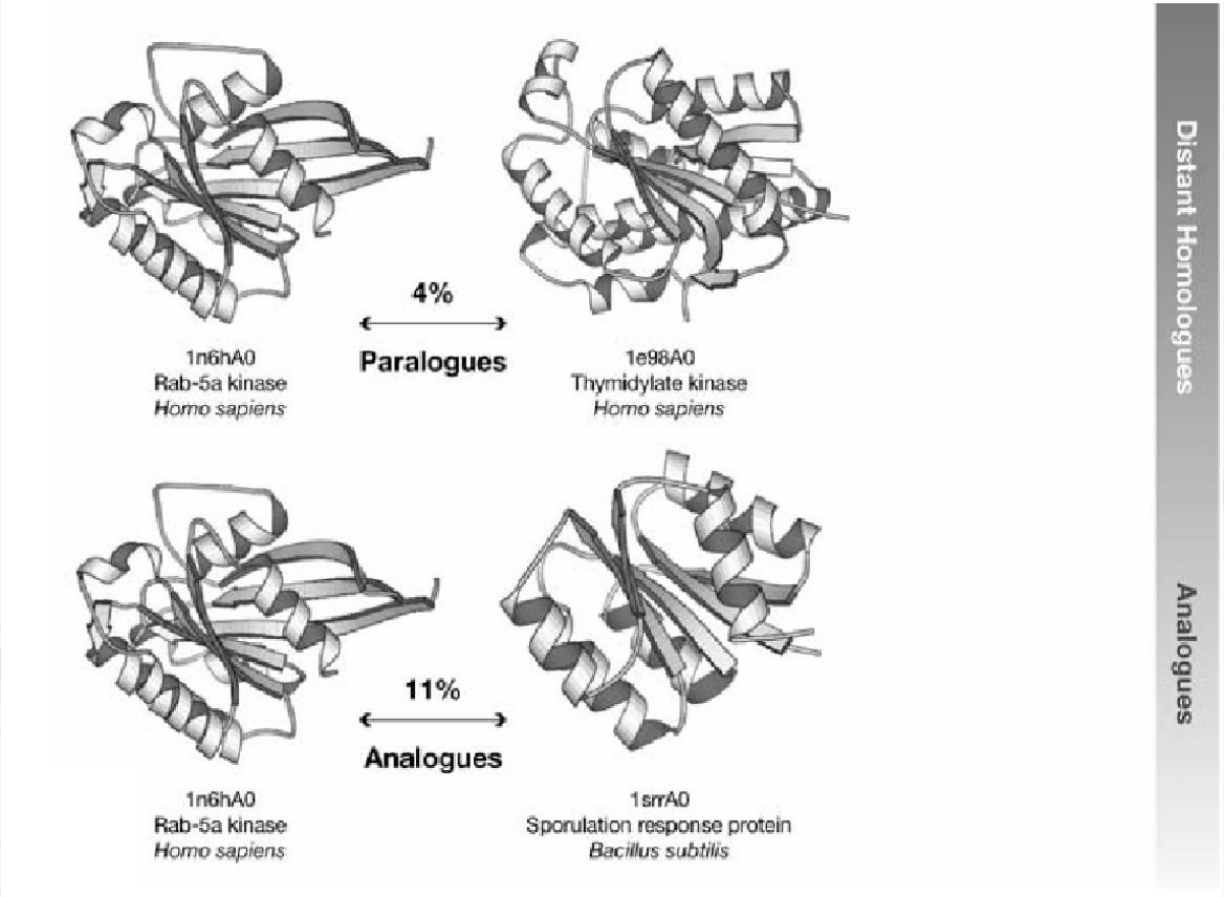
Homology

Remote homology

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

Analogues

- Same structure
- Different ancestor
- Same function ?



Università degli Studi di Padova

Dipartimento Matematica

DATA SCIENCE
UNIVERSITY OF PADOVA

STRUCTURAL EVOLUTION

Master of Science in Data Science

Damiano Piovesan

Structural evolution

- How protein structure has evolved?
  - Ancestral proteins?
  - Footprints of the evolutionary path?

- Structural evolution
  - Inference from structural classification
  - Hypothesis on the common elements
  - Theories on the origin of life

Structural complexity

DIPARTIMENTO MATEMATICA

Complexity

- Millions of species → each with thousand of coding genes

Mechanisms

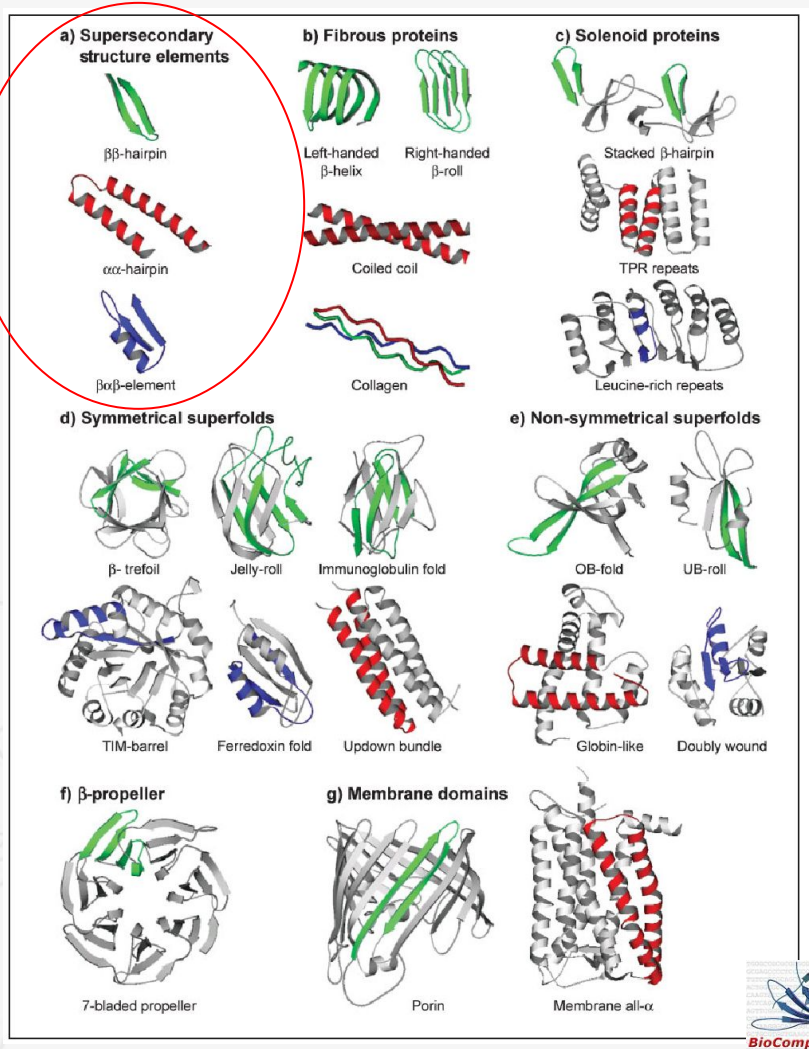
- Point mutations, insertion, deletions
- Random drift + natural selection
- Parental inheritance, acquisition, duplication

Observations

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

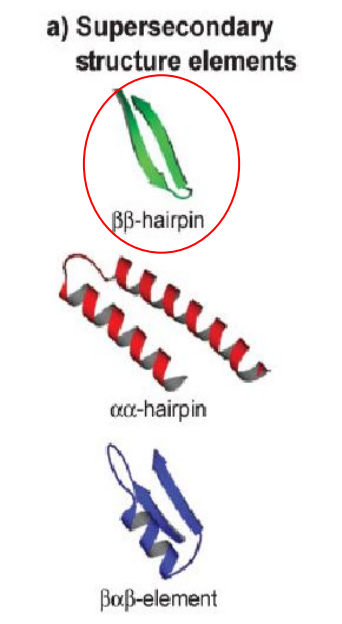
“More than the sum of their parts: on the evolution of proteins from peptides“

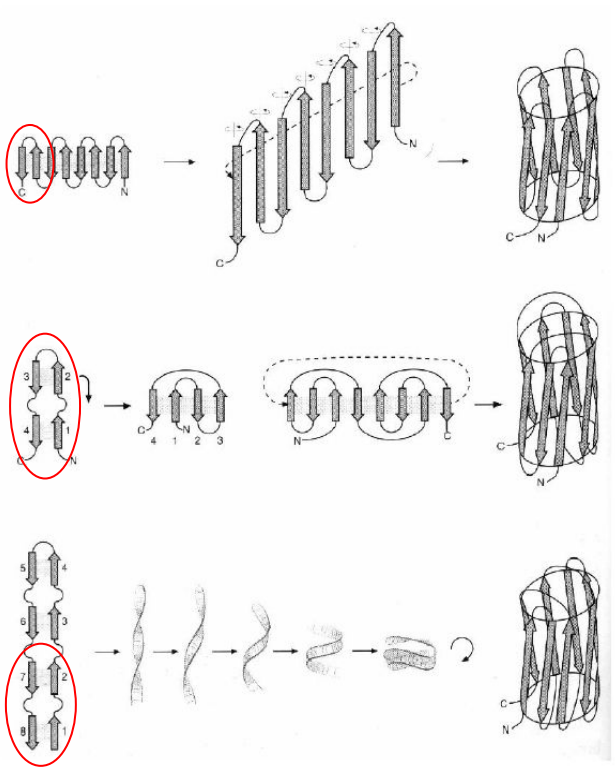
- There is a **basic complement** of autonomously folding units (**domains**)
- The complement was established at the time of the “**last common ancestor**”



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

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



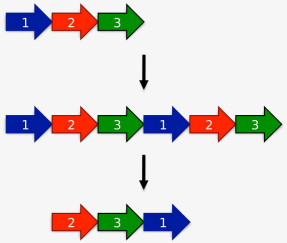


(J. Söding & A. Lupas, BioEssays, 2003)

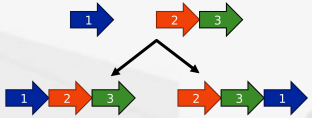
Examples of circular permutations

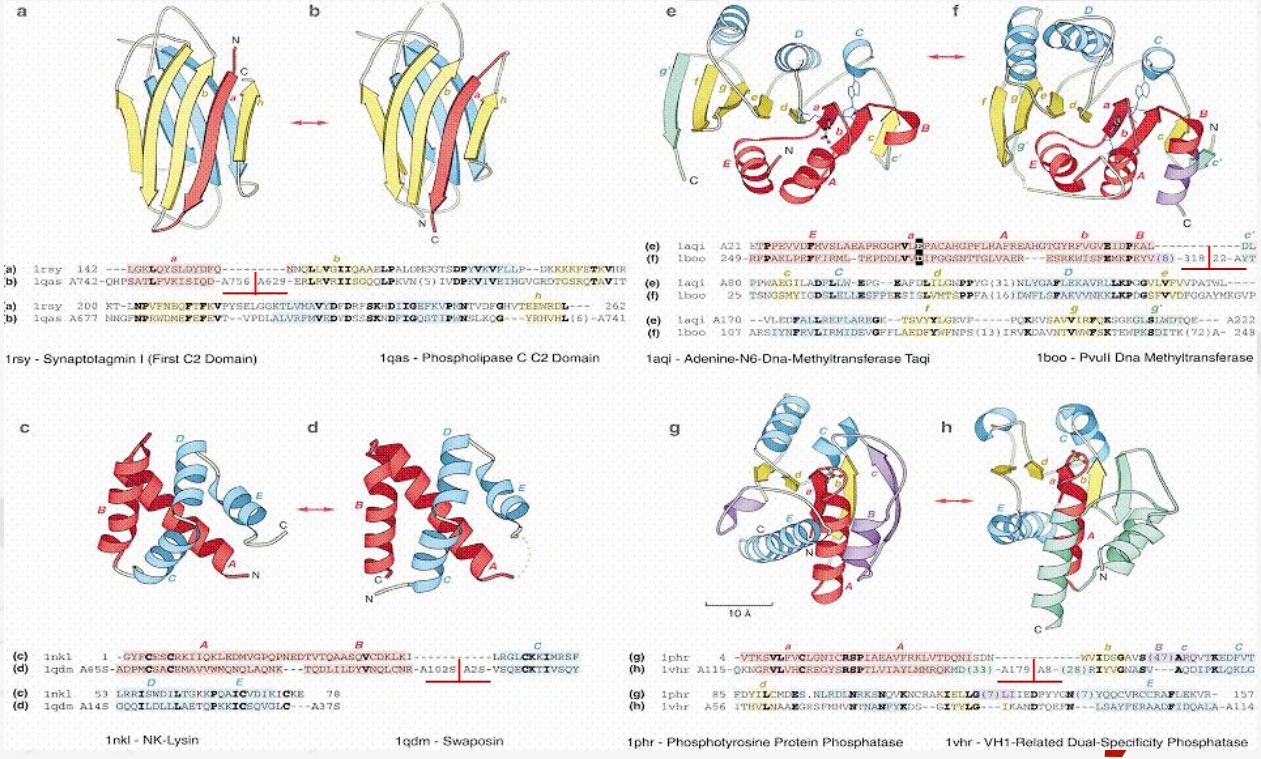
Close N- and C-termini can be found in different parts of the protein

- Duplication



- Fission and fusion

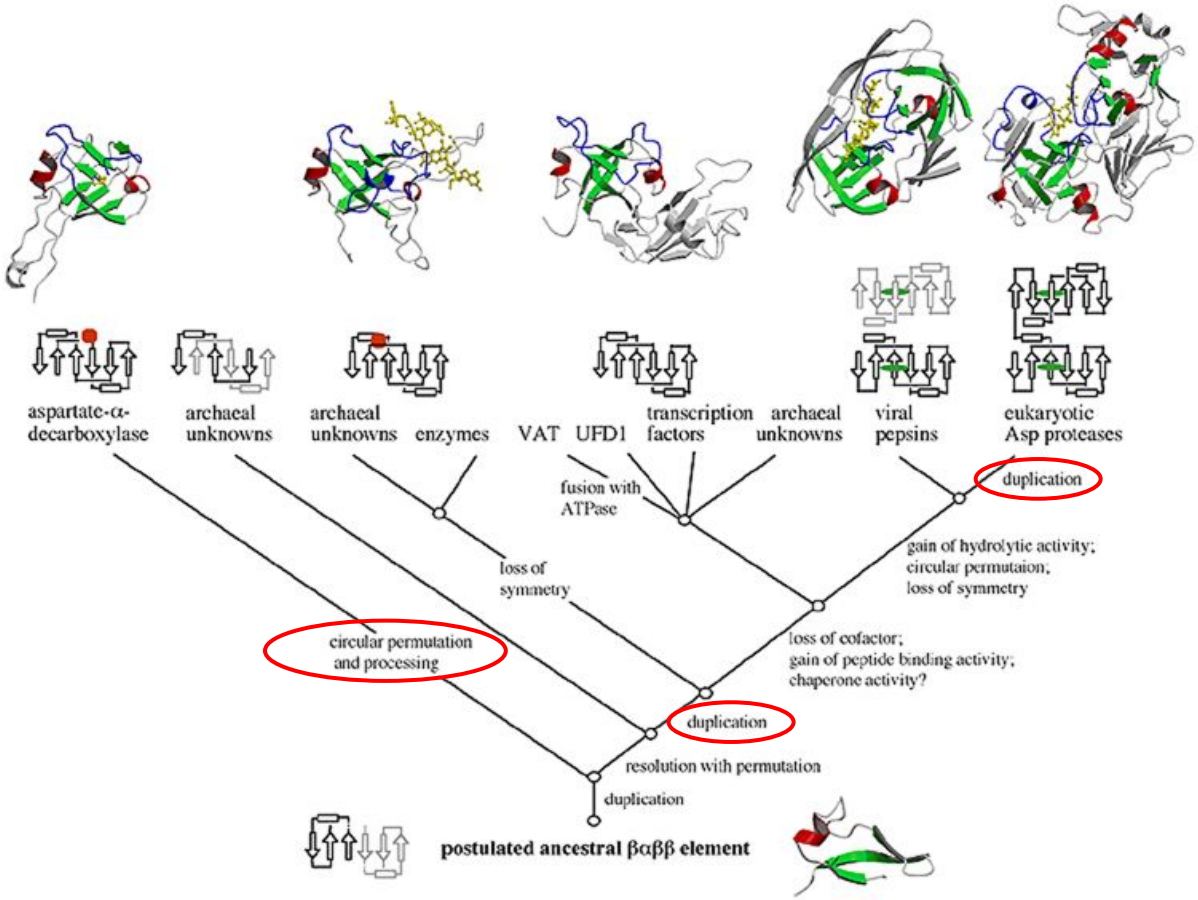




Model of structural evolution

Common "operators"

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



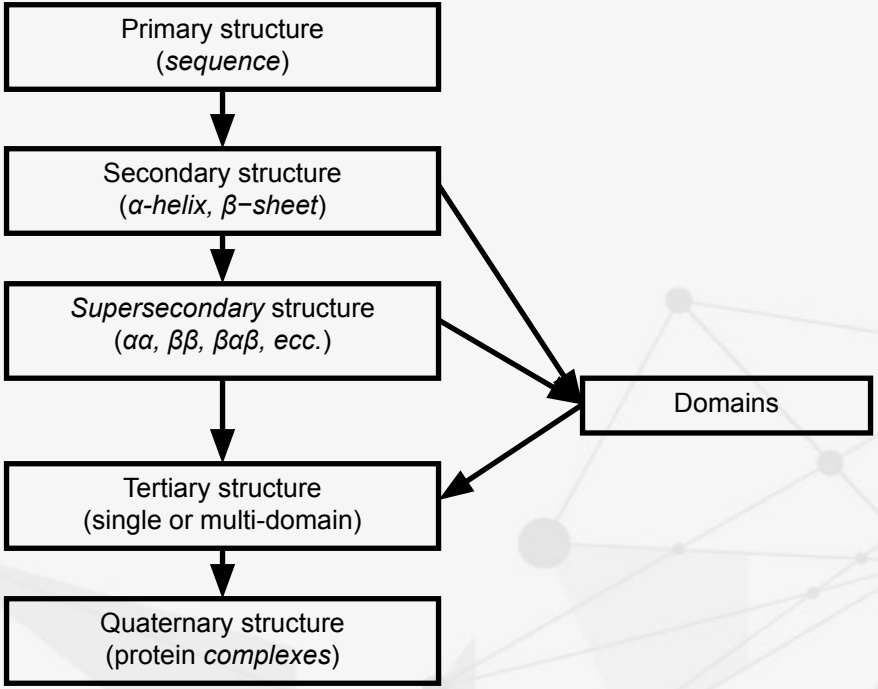
1222·2022 800 ANNI
Università degli Studi di Padova
Dipartimento Matematica
Data Science
University of Padova

STRUCTURAL CLASSIFICATION

Master of Science in Data Science

Damiano Piovesan

Recognition of distant evolutionary events make it possible to describe the basic complement of domains in the last common ancestor



Domain based classification

Families

Superfamilies

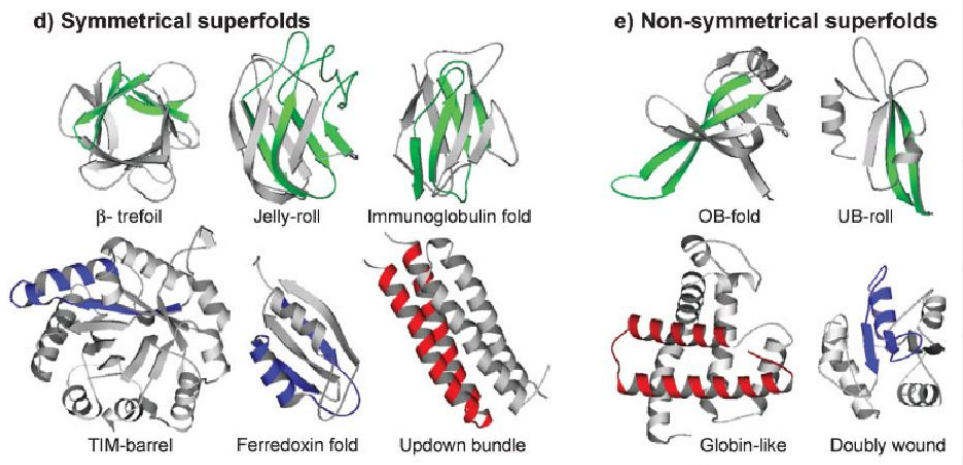
- Homologous families - Divergent evolution

Folds

- Analogous superfamilies - Convergent evolution

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”



Why some folds are so common?

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

Why some folds are so common?

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

Structural classifications

Manual

Semi-Automatic

Automatic

SCOP
Structural Classification Of Proteins

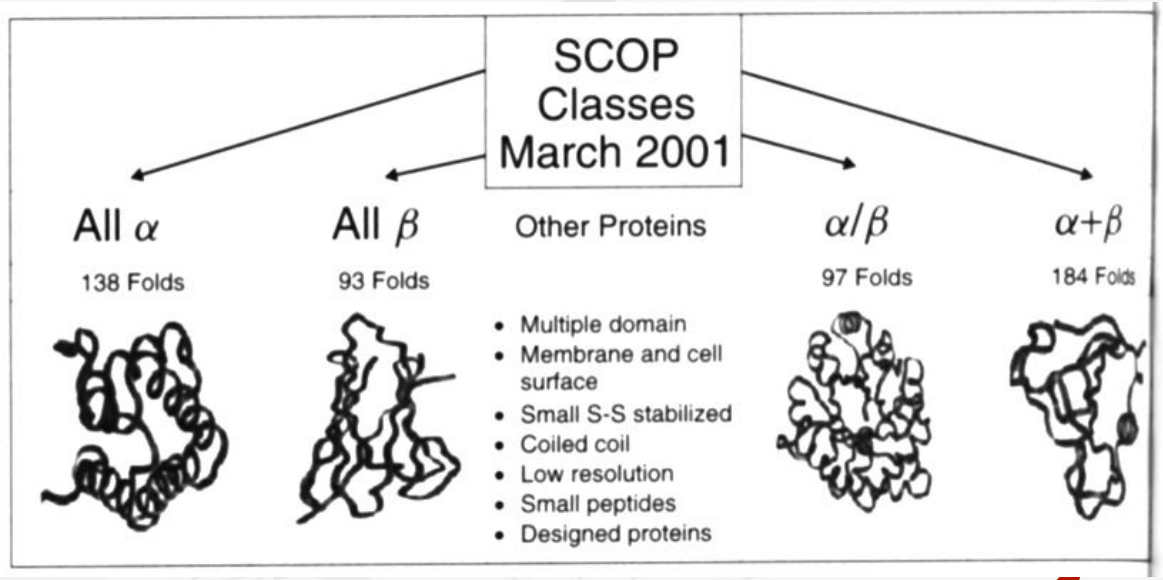
CATH
Class, Architecture, Topology, Homology

FSSP
Families of Structurally Similar Proteins

SCOP - Structural Classification of Proteins

- Alexey Murzin
- Mainly manually curated
- Gold standard

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



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]

SCOP About Contact Download Search SCOP by text or ID

SUPERFAMILY S-adenosyl-L-methionine-dependent methyltransferases

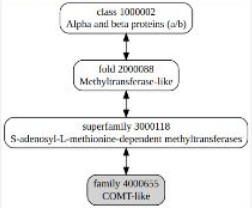
FAMILY

COMT-like

SCOP ID: 4000655

Function functionally relevant complex structure(s) determined

Show ancestry



Domains [24 entries]	ID	Region	Links
Protein Catechol O-methyltransferase Species *Rattus norvegicus* Representative domain 8026636 Represented structures [154]	P22734 2CL5	46-259 A:3-216	UniProt PDBe RCSB PDB
Protein Catechol O-methyltransferase Species *Homo sapiens* Representative domain 8078588 Represented structures [13]	P21964 4XUC	48-265 A:48-265	UniProt PDBe RCSB PDB
Protein O-methyltransferase family 3 Species *Niastella koreensis GR20-10* Representative domain 8101444 Represented structures [10]	G8T6H8 7CVX	3-221 A:3-221	UniProt PDBe RCSB PDB
Protein Caffeoyl-CoA O-methyltransferase Species *Medicago sativa* Representative domain 8020016 Represented structures [7]	Q40313 1SUS	21-247 A:21-247	UniProt PDBe RCSB PDB

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

SCOP - Structural Classification of Proteins

DIPARTIMENTO MATEMATICA

- Initially a protein structure is classified into **domains**
- 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**
- 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 - Folds

- 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

SCOP About Contact Download Search SCOP by text or ID

Statistics

	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

SCOP

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)

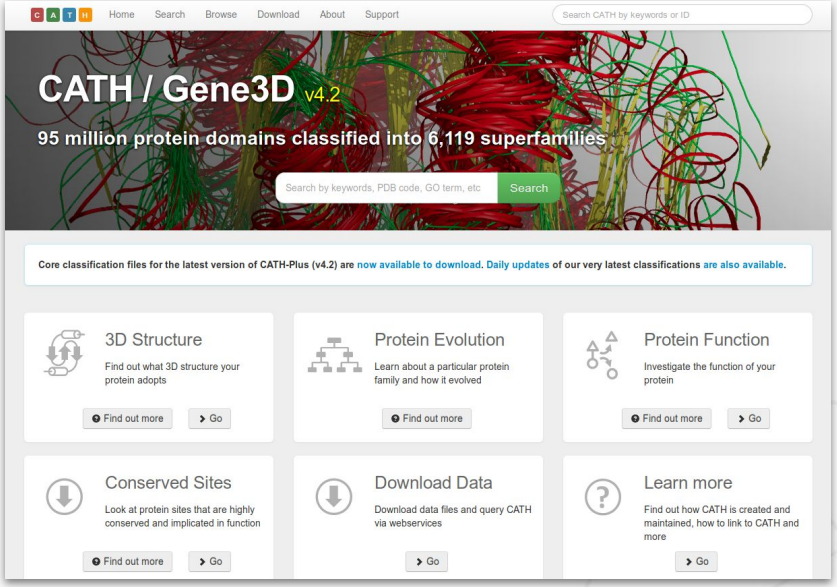
Families

- **Sequence identity +30%** or **functions** and **structures** are **very similar**
- Common evolutionary origin

“Strange” families

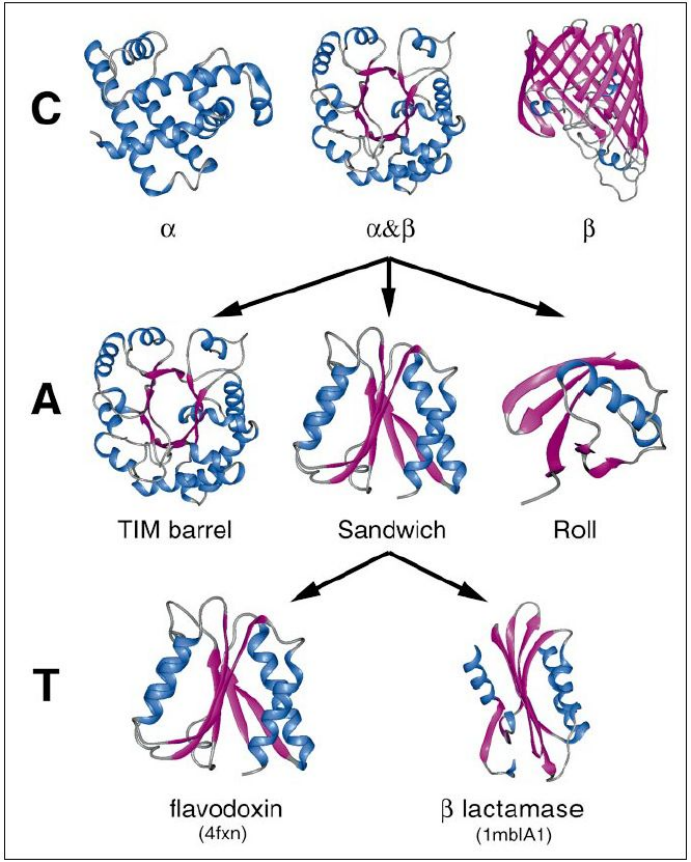
- Sequence similarity below family definition but above the superfamily level
- Similar domain organization, **common fold in the catalytic domain** → likely to be closely related

CATH



- Semi-Automatic
- Only Architectures are manually assigned

CATH

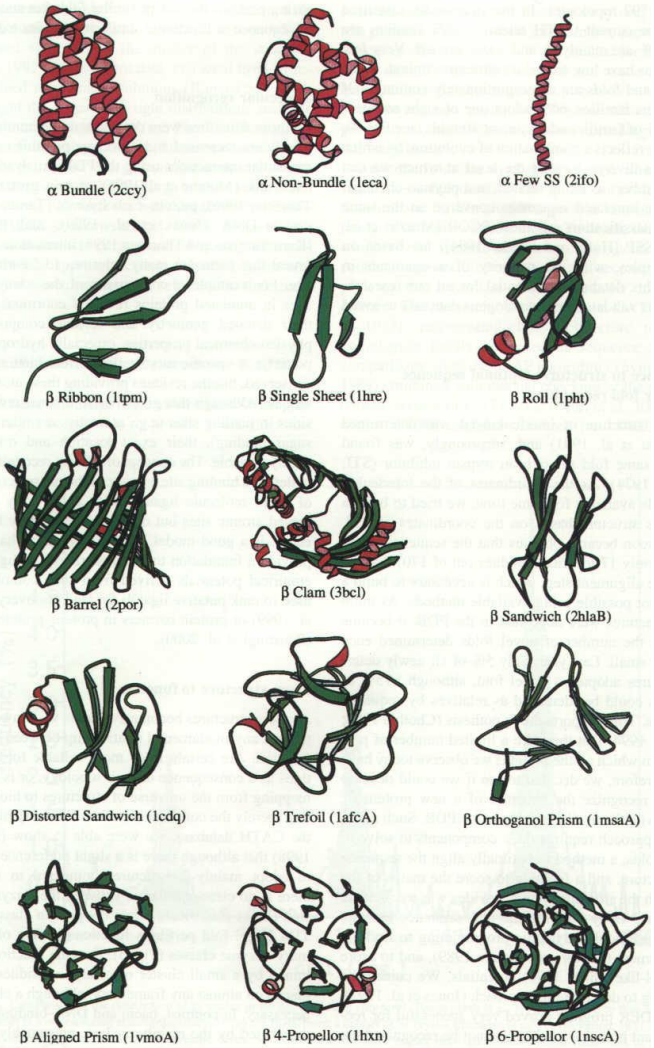


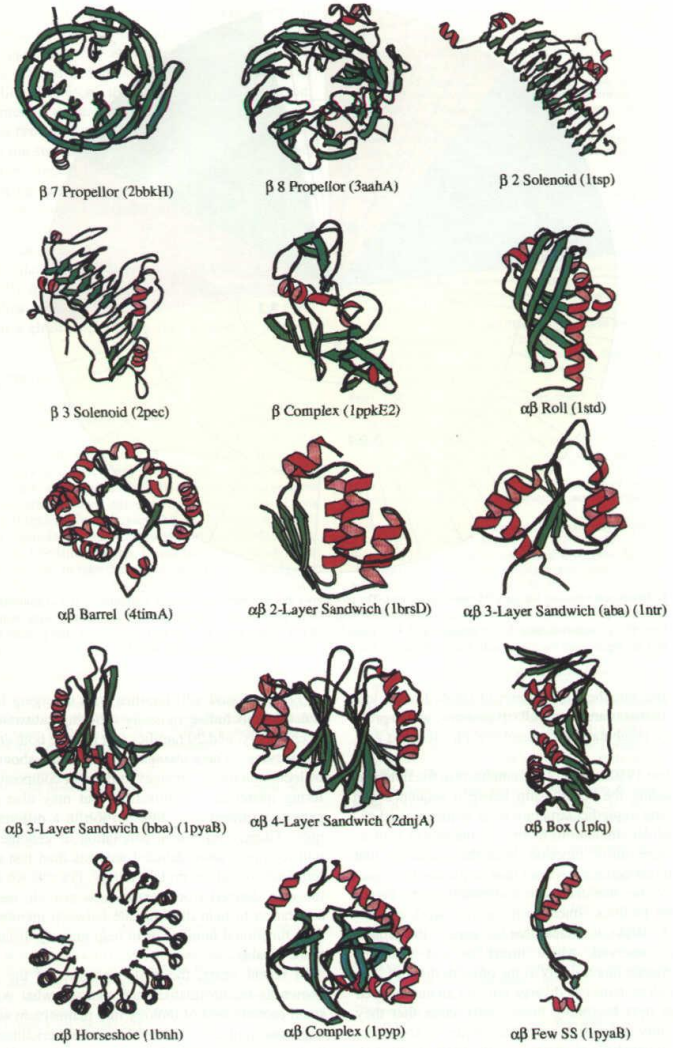
Semi-Automatic, only Architectures are manually assigned

- **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**

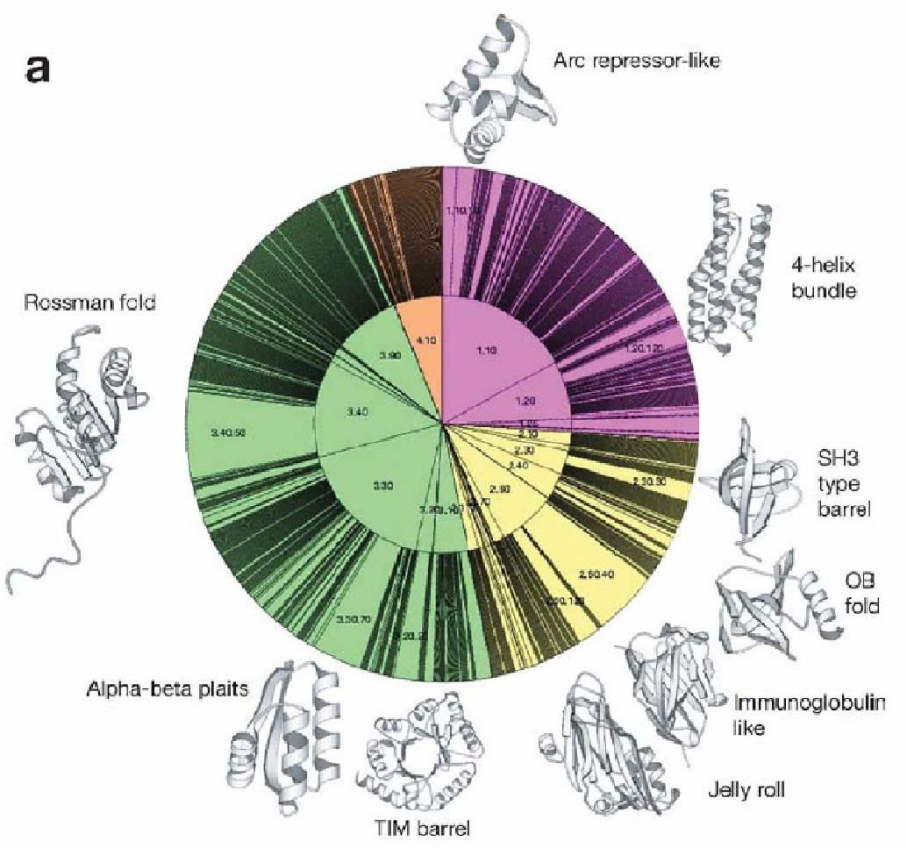
CATH

architectures



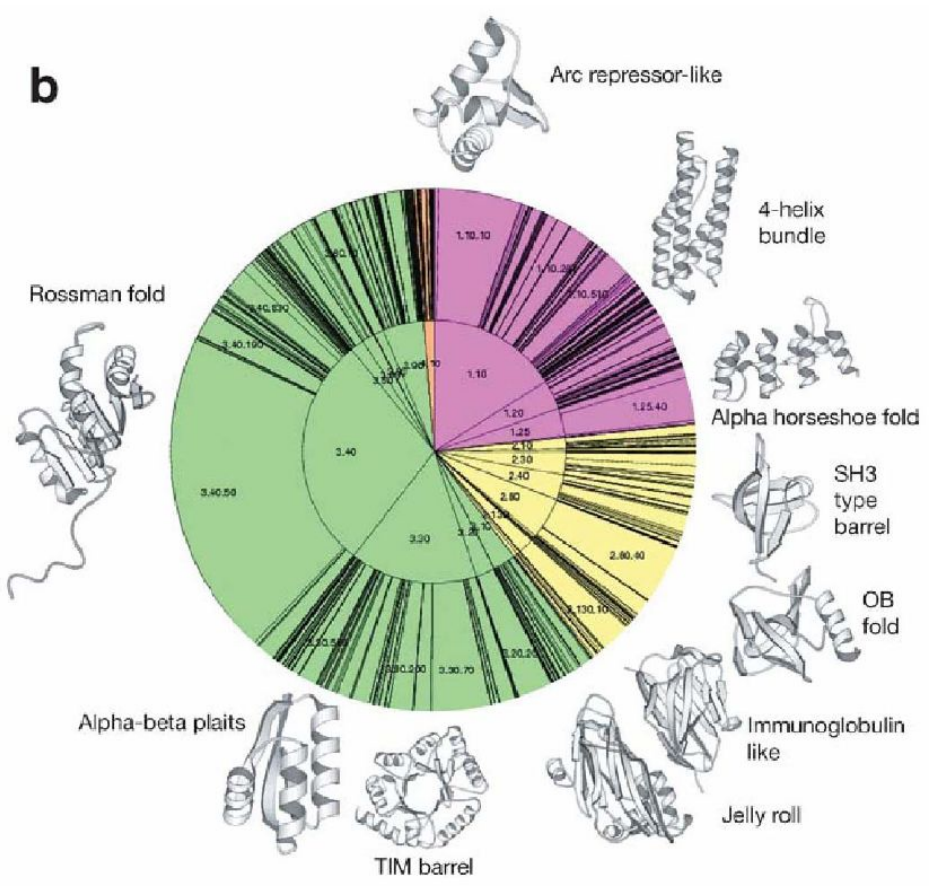


CATH hierarchy in the PDB



- 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