Abstracting Knowledge from Protein Structures for Biology in the 21st Century

PDB40 Symposium

CSHL
October 2011

Janet Thornton
EMBL-EBI

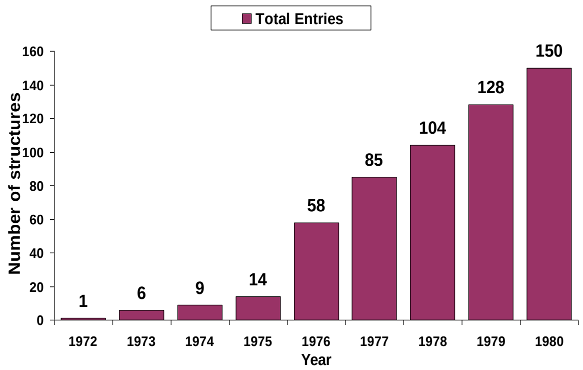
EMBL-EBI

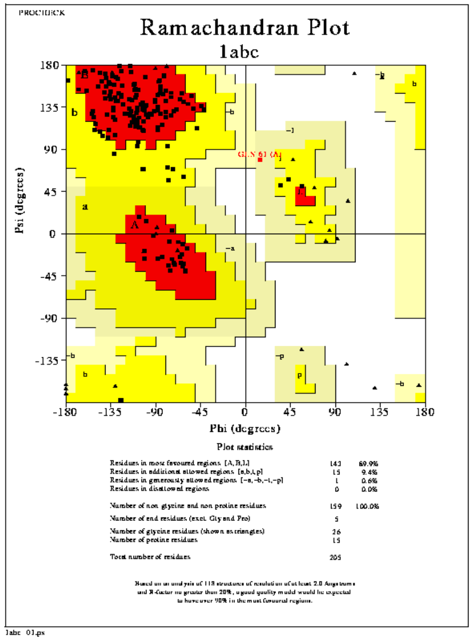
Overview

- Personal Recollections of PDB
- Abstracting knowledge from structures for biology in the past and today
- Thoughts about the Future of PDB
- Thanks

Personal Recollections of the PDB: 1974 - 1995

- 12” tapes about every 3 months from Brookhaven via Daresbury to Oxford Lab in ~1974
- Growth in number of entries (‘70s)
- Validation 1989 CCP4 ‘Errors in Protein Structures’ / PDBClean/ PROCHECK
- Visits to Brookhaven (Tom Koetzle, Frances Bernstein & Enrique Abola) as part of Scientific Advisory Board
- Challenges of data increase – move to RCSB: Helen, Phil & Gary





Personal Recollections of the PDB: 1995 onwards

PDBe
PROTEIN DATA BANK EUROPE

- Establishing PDBe – grant from Wellcome Trust (for 4 staff) to EMBL- EBI:
- 1995 – recruitment of Kim Henrick  & Geoff Barton
- Building relationships between PDBe & RCSB/PDBj/BMRB 1995 - 2005
- Kim & colleagues started to build the EMDB (2002)

EMDB
EM DATA BANK

- Establishment of wwPDB

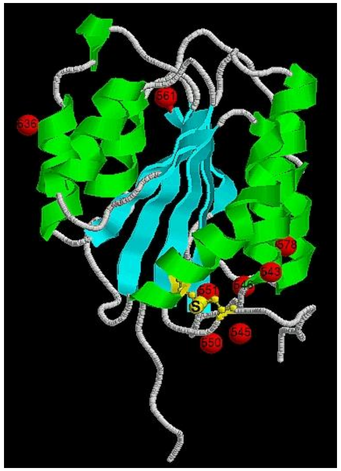
WORLDWIDE PDB
PROTEIN DATA BANK

- Recruiting Gerard (Kleywegt) – 2009
  'Bringing Structure to Biology'

Abstracting Knowledge from the PDB

- The knowledge contributed by an individual protein structure about how this particular protein performs its biological function remains the most important aspect of knowledge in the PDB e.g. **Von Willebrand Factor**

- BUT additional knowledge in many areas can also be abstracted by combining information over many structures. In practice most proteins interact with many other molecules, either as multimers or as parts of pathways



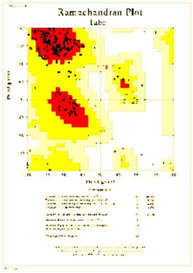
PDB code: 1auq
Emsley *et al* (1997)
J.B.C. 273 10396

Σ Information over all or subset of PDB entries to generate knowledge

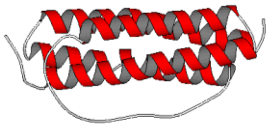
Abstracting Knowledge from PDB: Historical perspective

- Practical knowledge e.g. Which proteins are likely to crystallise
- Basics Principles of Protein Structure (physics/chemistry)
- The Universe of Proteins & evolutionary relationships
- Structure to Function

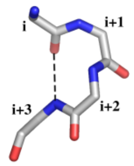
1970's Basic Principles of Protein Structure (Understanding Sequence to Structure)



- Properties of amino acids eg helix propensities
- Basic geometry of pp chain, e.g. phi,psi values
- Hydrophobic Core
- Secondary Structures



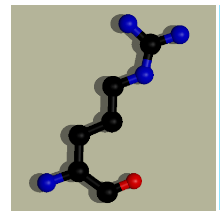
  - Helices - geometry; length, curvature; packing
  - Strands – twist; geometry; residue pairs
  - Turns – types; residue preferences

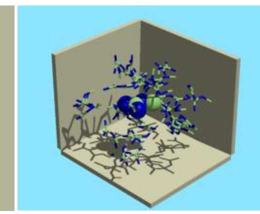


- Chirality
  - Twists of sheets, Right handed βαβ, Barrels
- Tools for 'describing' protein structures
  - Secondary Structure Assignment - DSSP
  - Hydrogen bonds - HBPlus
  - Accessibility - NACCESS

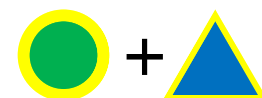
1980's The Universe of Protein Structures from the PDB

- Interactions:
  - Amino acid packing

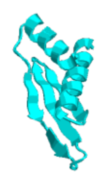




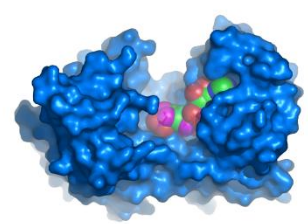
  - Tertiary packing – helix; sheet
- Domains & multi-domain architectures



- Folds
- Evolution – conserved structures
- New Tools

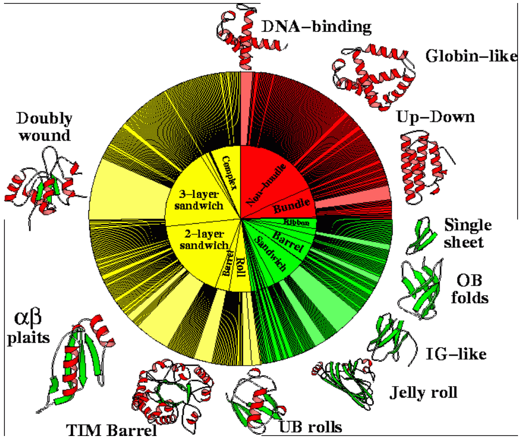


  - Visualisation
  - Homology Modelling
  - Simulations
  - Electrostatics

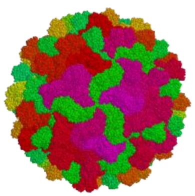


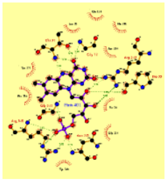
1990s Folds; Classification; Interactions

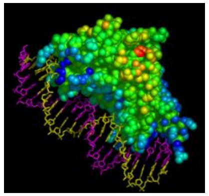
- Protein Structure Classifications CATH & SCOP

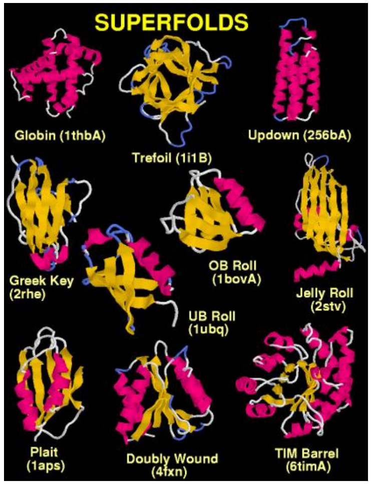


- Interactions
  - Protein-protein
  - Protein-Ligand
  - Protein-DNA





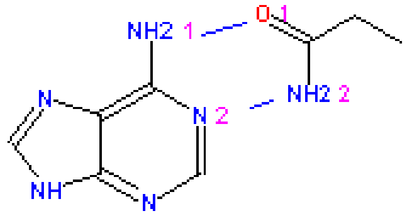




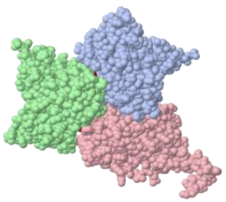
- New Tools:
  - Structure Comparison eg DALI
  - Patch Analysis for PPI
  - Docking
  - Fold Recognition - Threading

Many of Tools now provided by PDB as searches

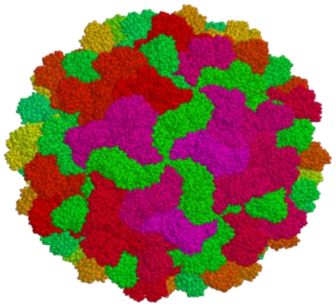
- PDBeMotif – to identify motifs



- PDBePISA – to assign multimeric status in crystal

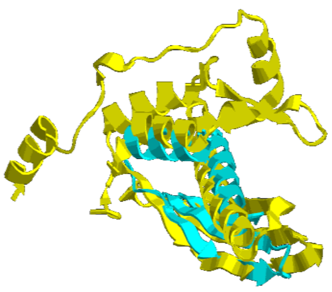


2TBV A trimer?



Biological unit 2TBV
180-mer!

- PDBeFold – to find all similar folds in PDB



Structural Genomics Projects ~2000

Taken from
www.isgo.org

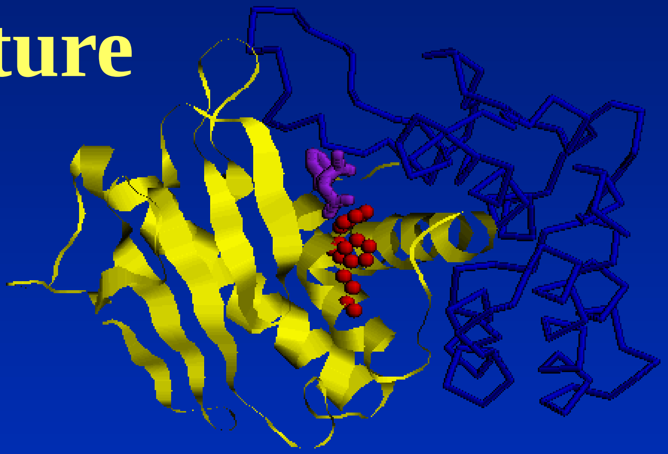
Ontario Centre for ***SG***	Canada
Montreal-Kingston Bacterial ***SG*** Initiative	
Montreal Network for Pharmaco-Proteomics and ***SG***	
CyberCell Project	
Structural Proteomics in Europe (SPINE)	Europe
SG of *Mycobacterium* pathogens	
SG of Eukaryotes	France
Yeast ***SG***	
SG of Orphan *E. coli* Genes	
Protein Structure Factory	Germany
RIKEN ***SG***/Proteomics Initiative	
National Project on Protein Structural and Functional A enters)	Japan
Biological Information Research Center (BIRC)	
The Korean Structural Proteomics Research Organization	Korea
National Centers for Competence in Research (NCCR)	Switzerland
North West ***SG*** Centre	
Oxford Protein Production Facility	UK
Cambridge Group	
New York ***SG*** Research Consortium	
Midwest Center for ***SG***	
Berkeley ***SG*** Center	
Northeast ***SG*** Consortium	
TB ***SG*** Consortium	USA
Southeast Collaboratory for ***SG***	
Joint Center for ***SG***	
SG of Pathogenic Protozoa Consortium	
Center for Eukaryotic ***SG***	
Structure 2 Function Project	

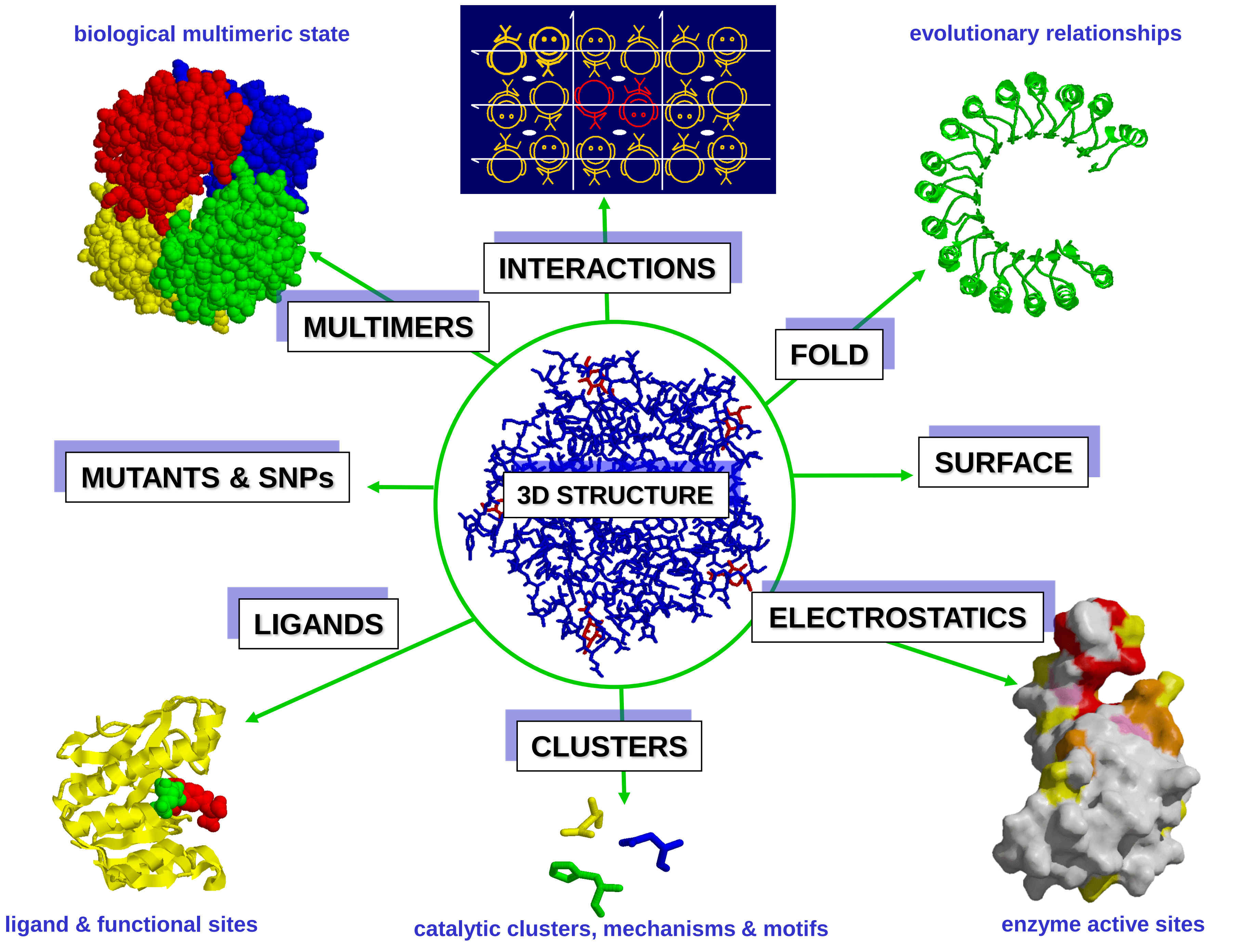
From Structure to Function

Protein Structure

↓

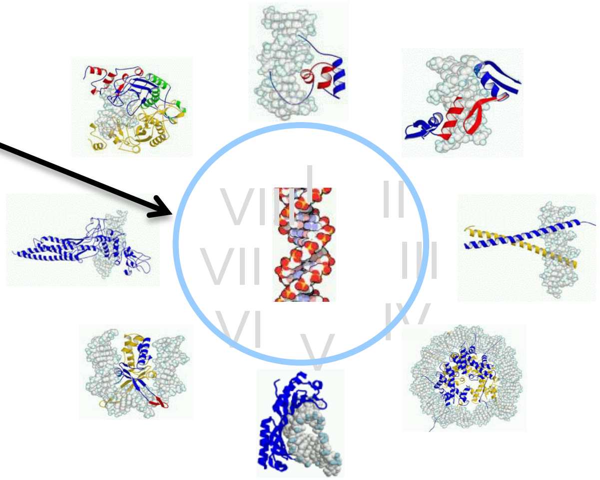
Molecular Function




biological multimeric state
evolutionary relationships
INTERACTIONS
MULTIMERS
FOLD
MUTANTS & SNPs
3D STRUCTURE
SURFACE
LIGANDS
ELECTROSTATICS
CLUSTERS
ligand & functional sites
catalytic clusters, mechanisms & motifs
enzyme active sites

Fold & Function

- No direct correlation between fold & function, though some tendencies
  - DNA binding proteins tend to be helical
  - Haem binding proteins tend to be helical
  - Enzymes tend to adopt αβ folds
  - Immune-related proteins tend to be β-sheet structures e.g. Ab
  - Membrane proteins are predominantly helical – apart from porins



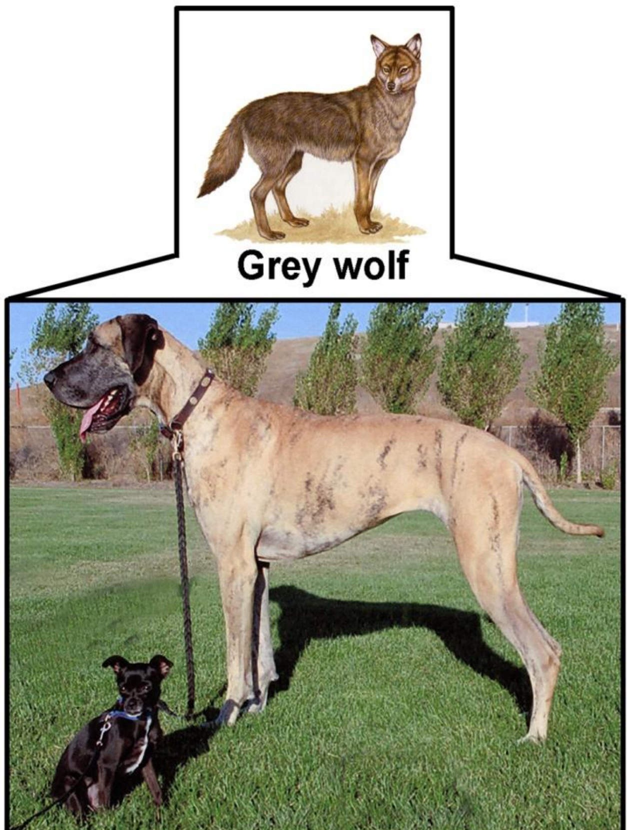
From Structure To Biochemical Function

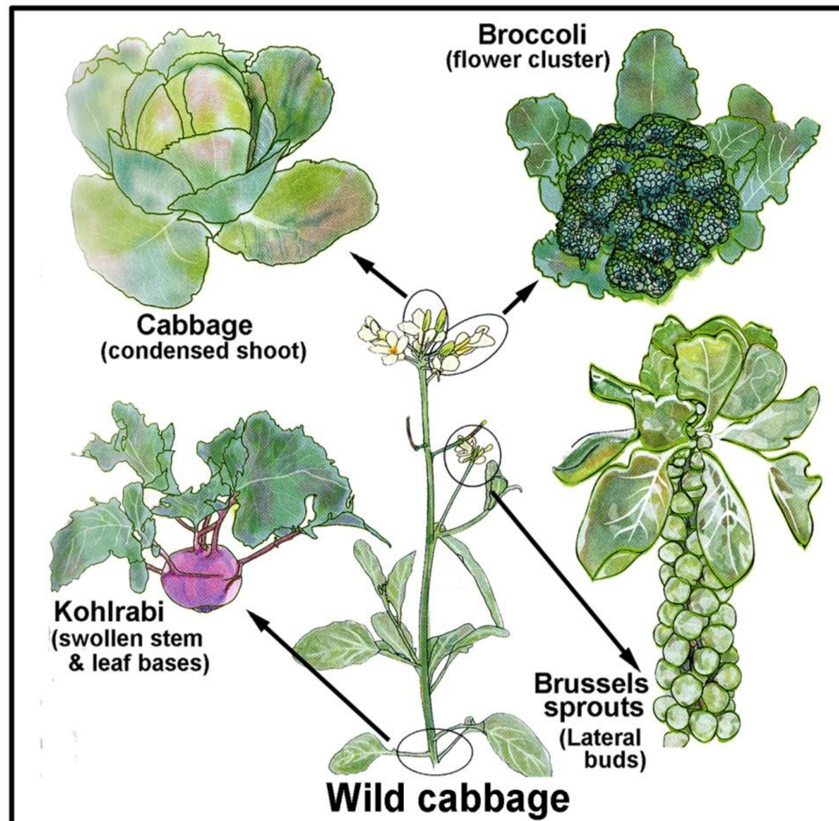
However identifying sequence or structural similarity (i.e. identifying an evolutionary relationship) is the most powerful route to function assignment

BUT members of the same protein superfamily often have a related but not identical function

MICROEVOLUTION BY ARTIFICIAL SELECTION

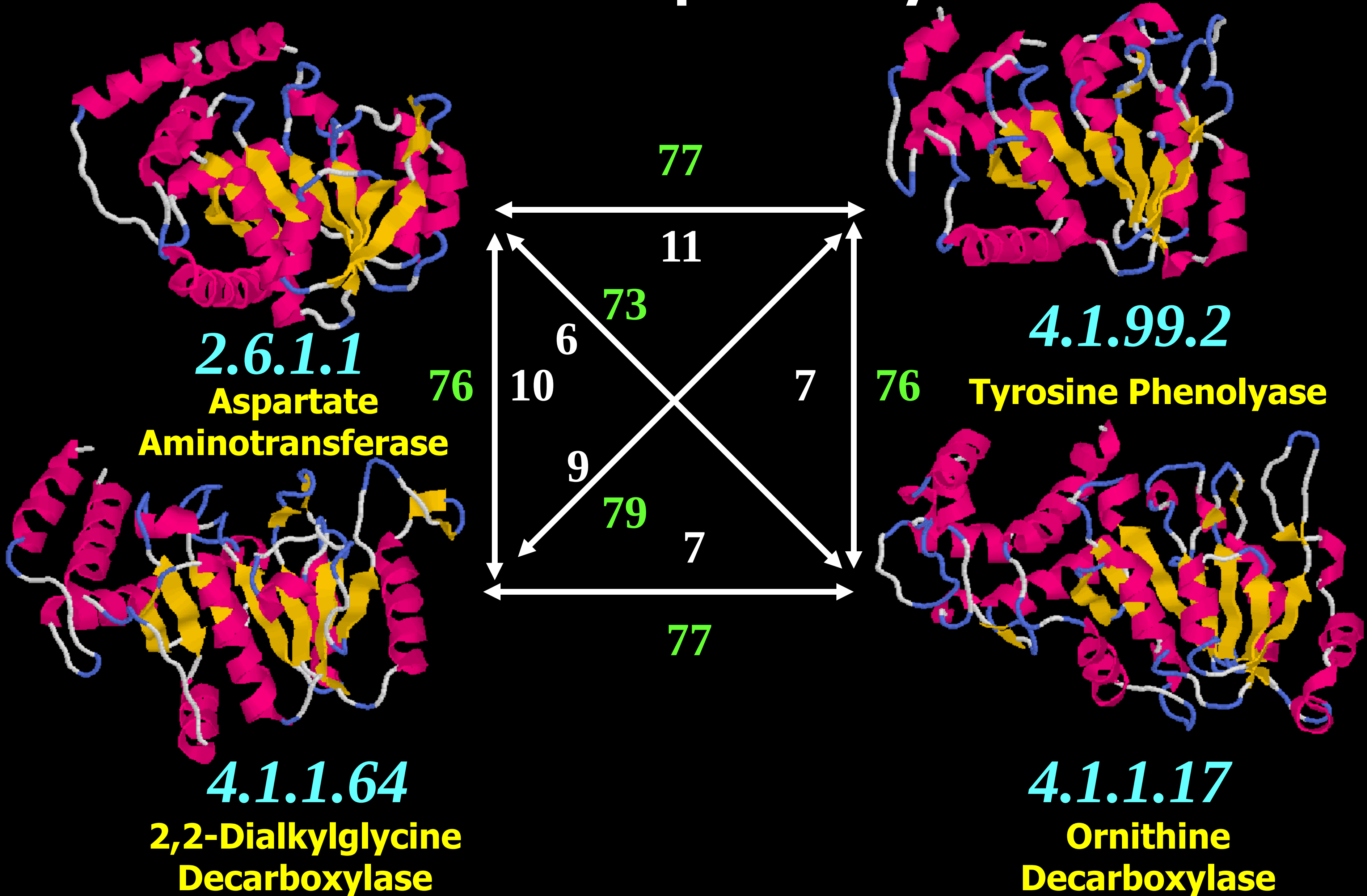
Rock pigeon



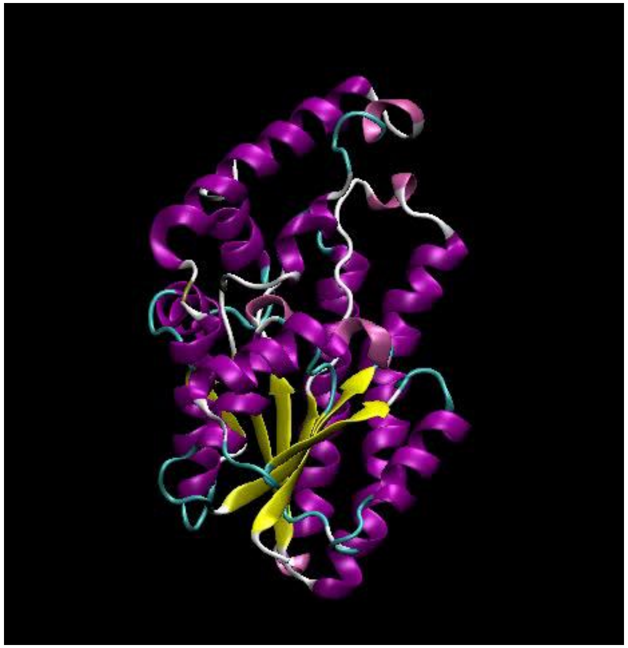


John Ellis

Aspartate Amino Transferase Superfamily

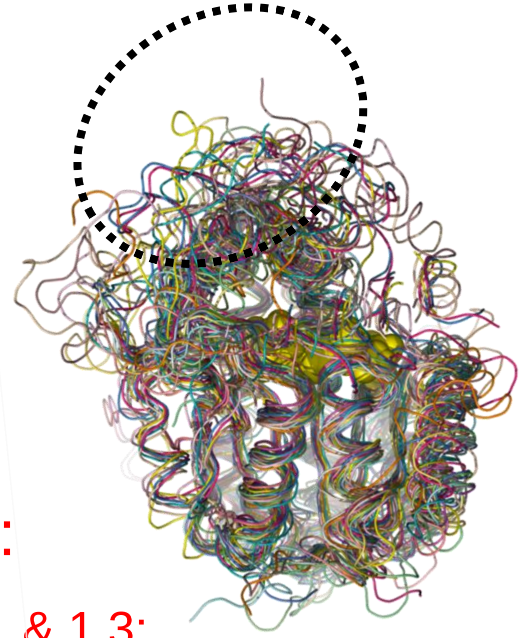


SDR Family
Short chain dehydrogenase/reductase family



>60 in humans

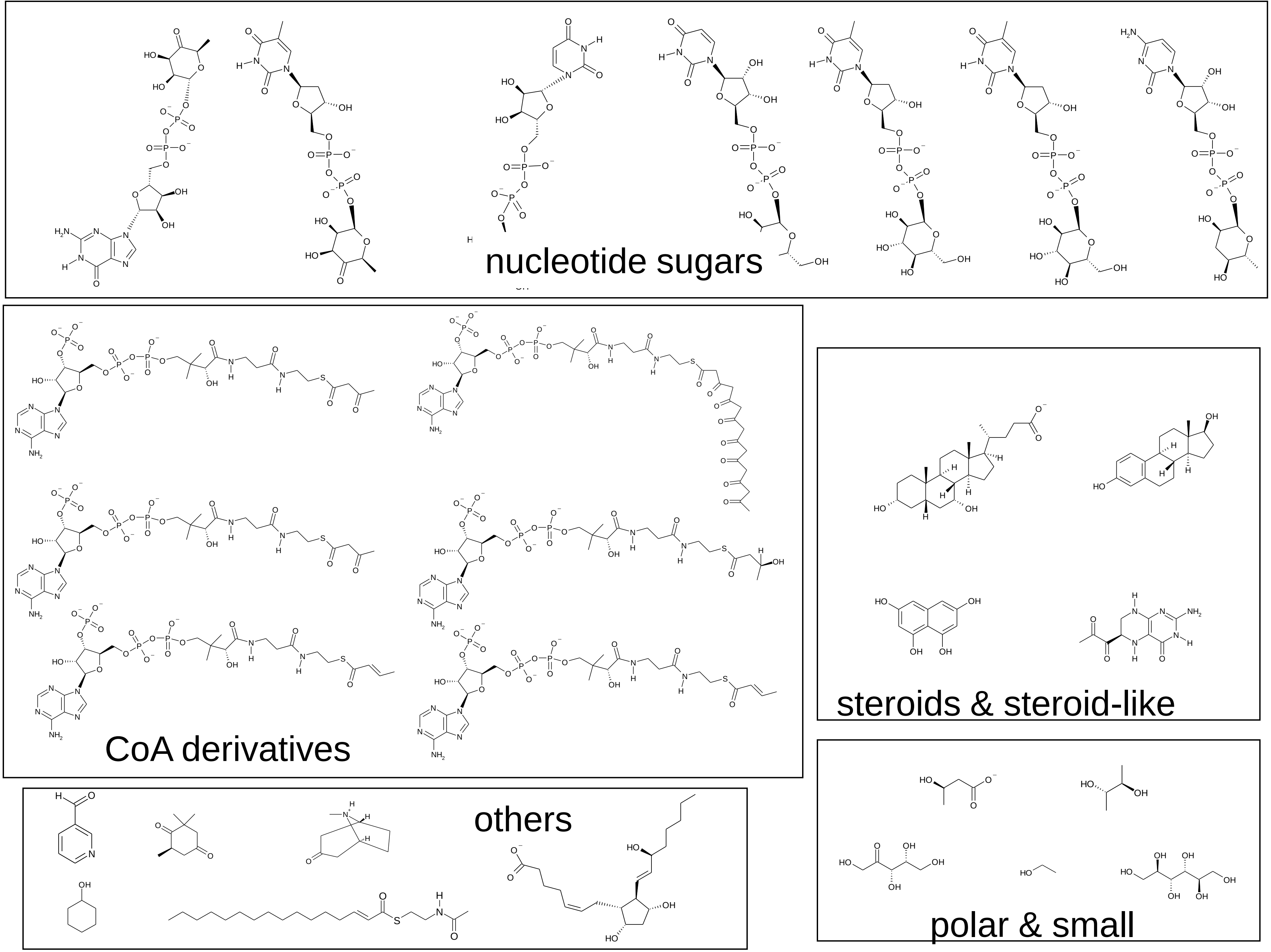
Catalytic Tetrad:
S,Y,K,N



Different Functions:

Oxidoreductases E.C. 1.1 & 1.3;
Lyases E.C. 4.3;
Isomerases E.C. 5.1

Many structures solved
Many different substrates

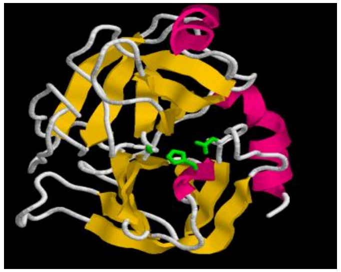

nucleotide sugars
CoA derivatives
steroids & steroid-like
others
polar & small

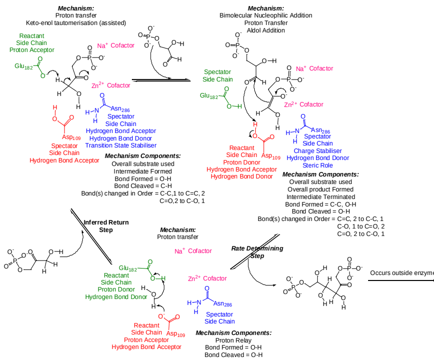
EBI Nick Furnham,
Gemma Holliday

UCL Christine Orengo
Ian Sillitoe, Alison Cuff

Understanding Enzyme Families and Evolution

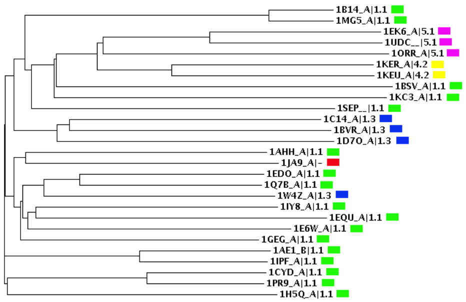
Understanding Enzyme Families & Evolution





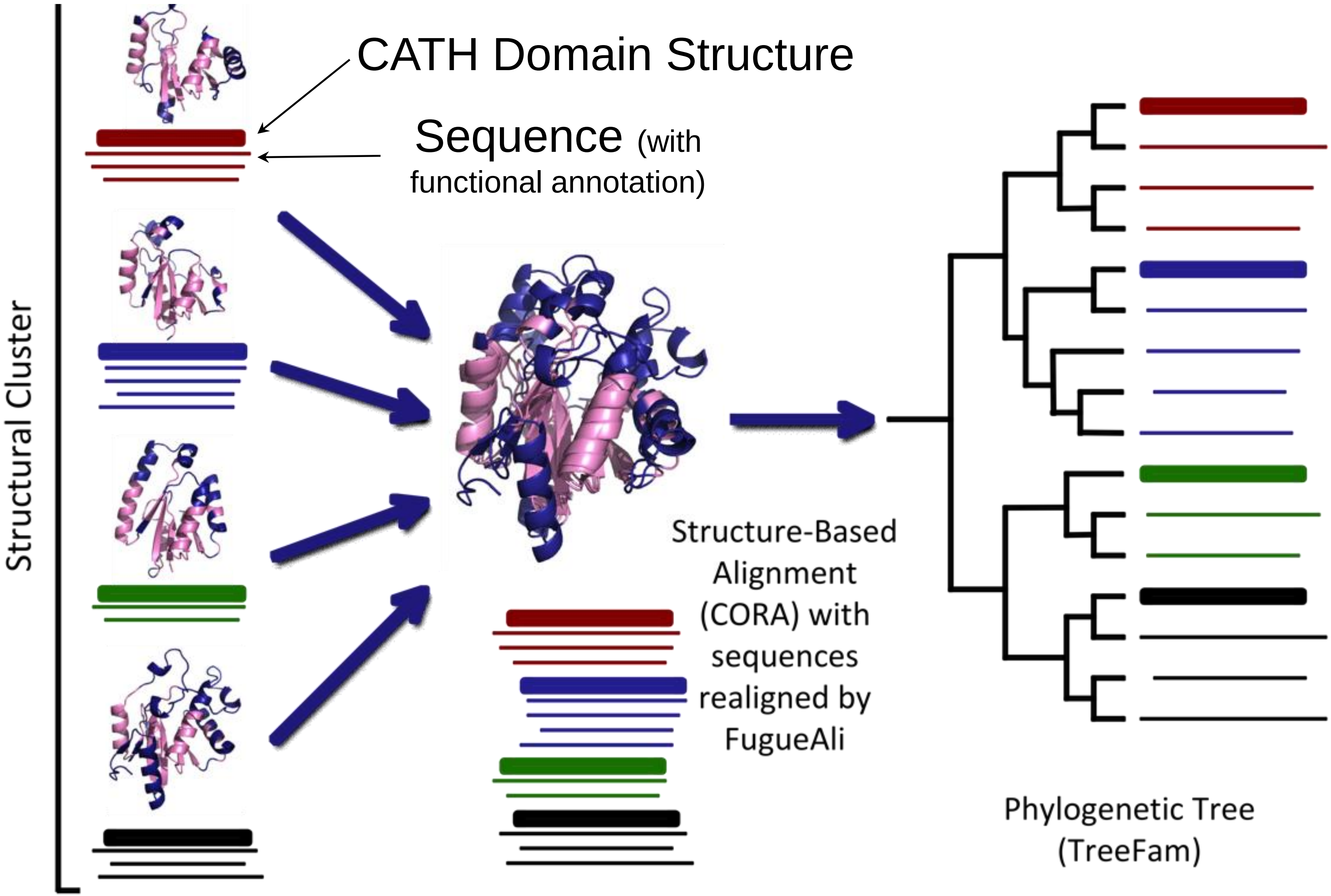
- Data
  - Protein Sequences
  - Protein Structures with ligands!
  - Substrate Knowledge (promiscuity)
    - in vitro
    - In vivo
  - Reaction mechanisms

- Computational tools for:
  - Sequence comparison
  - Structure comparison
  - Small molecule comparison
  - Reaction comparison



- Then we need to **integrate and visualise** all these data!!

Structurally Similar Groups



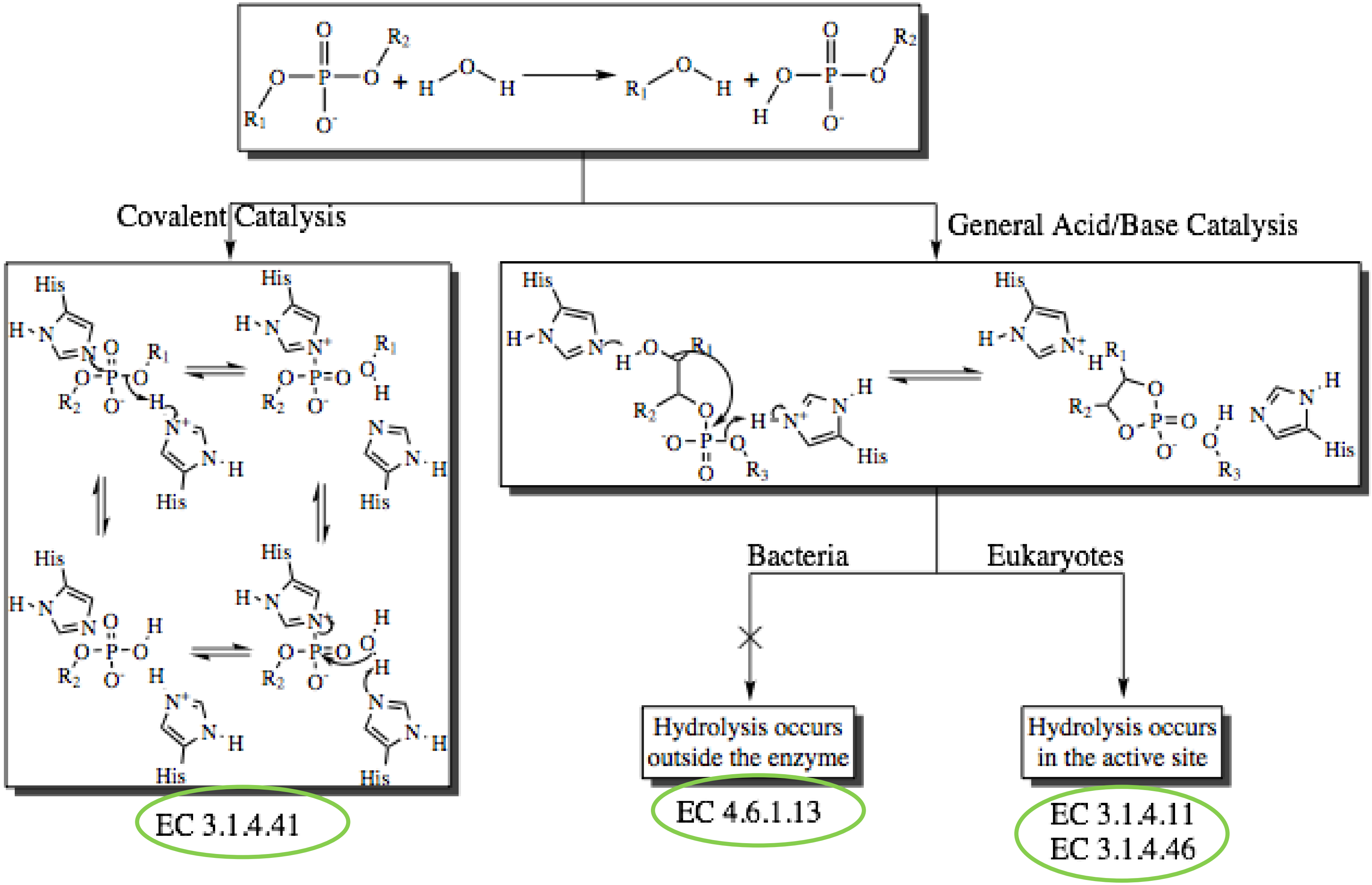
The pipeline

MACiE

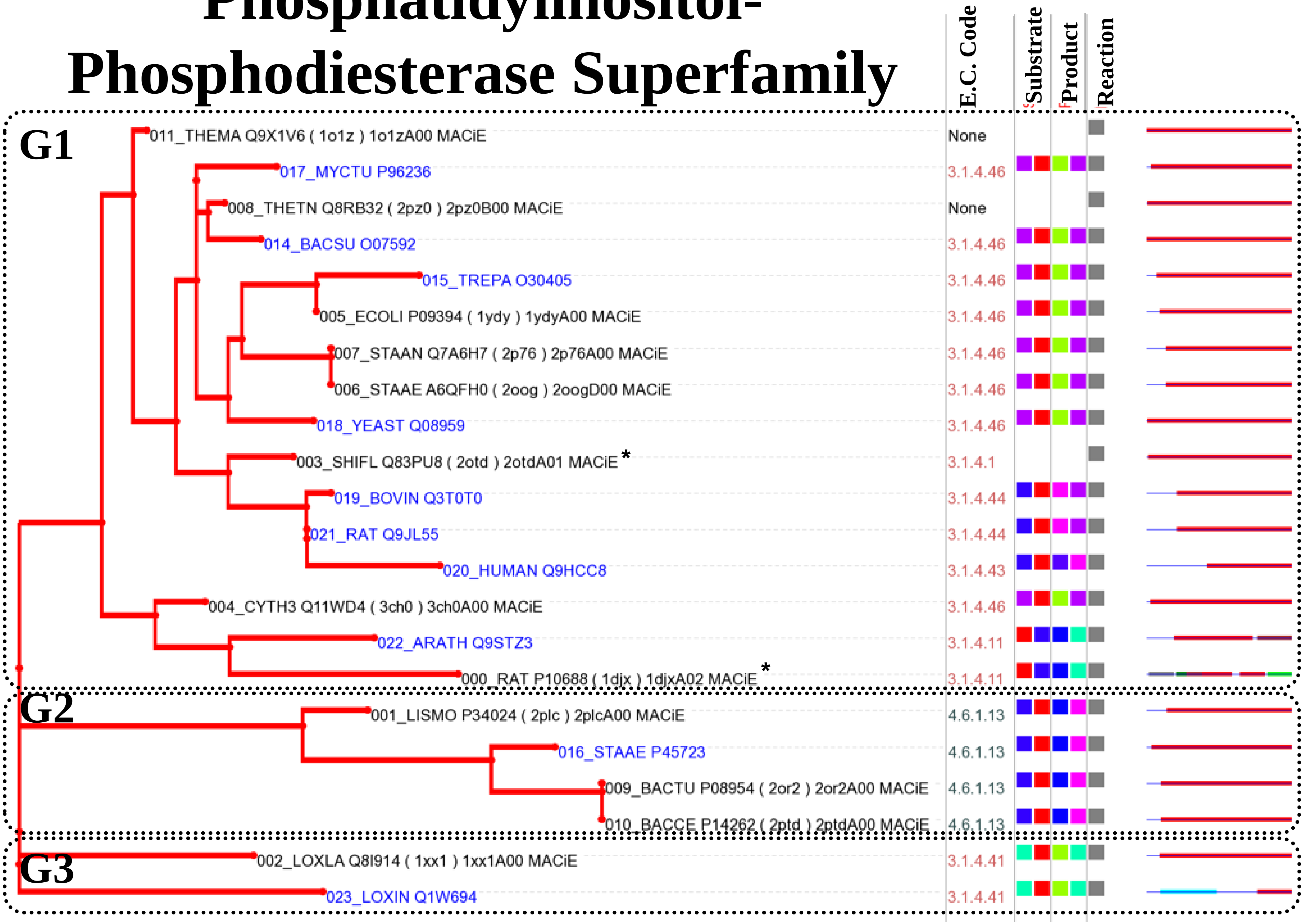
Structure and sequence alignments for
enzyme families -> Phylogenetic trees

Annotate with functional information
and small molecule data (eg substrates, mechanism)

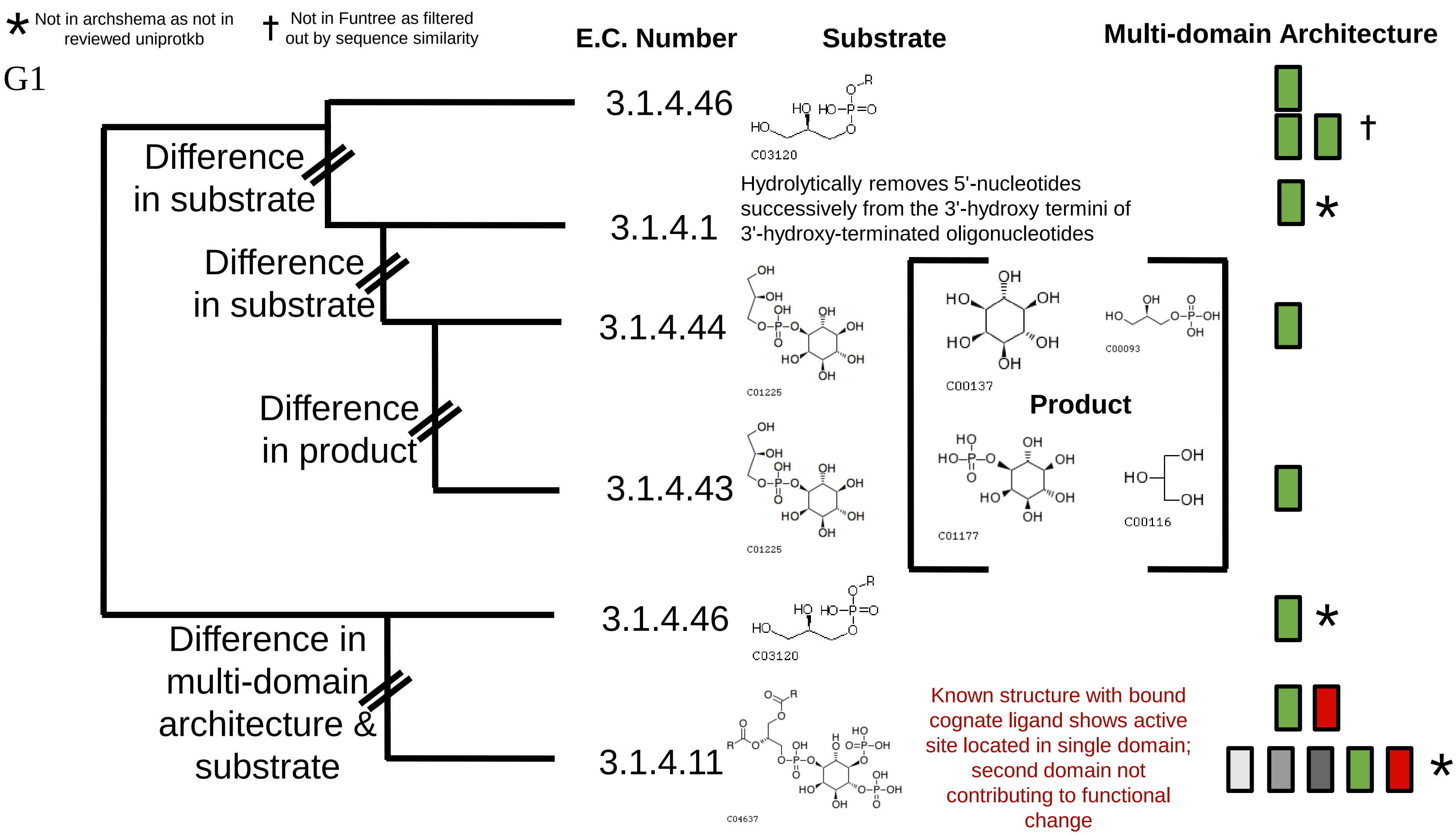
Phosphatidylinositol-Phosphodiesterase (PIP) Superfamily



Phosphatidylinositol-Phosphodiesterase Superfamily

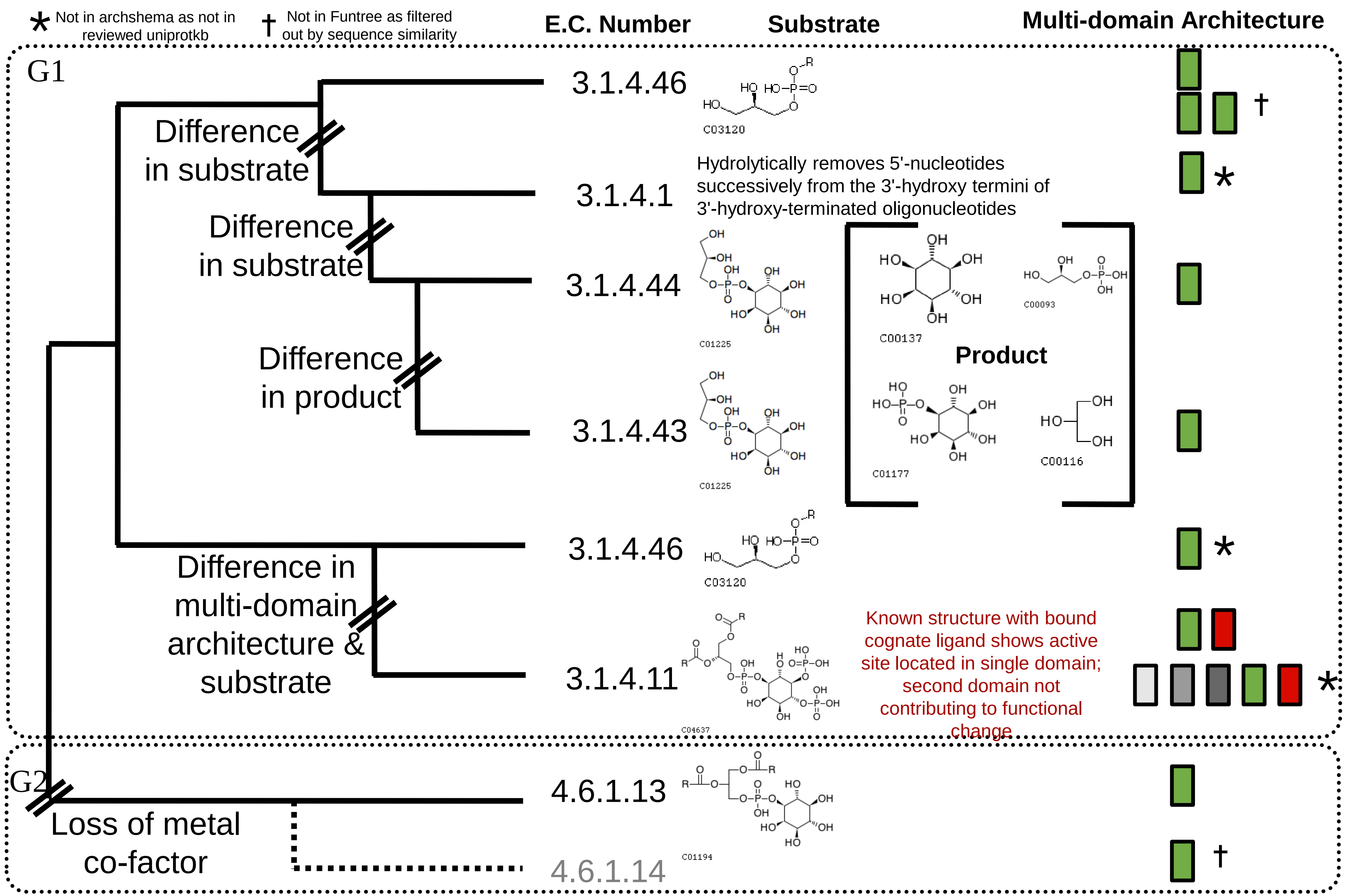


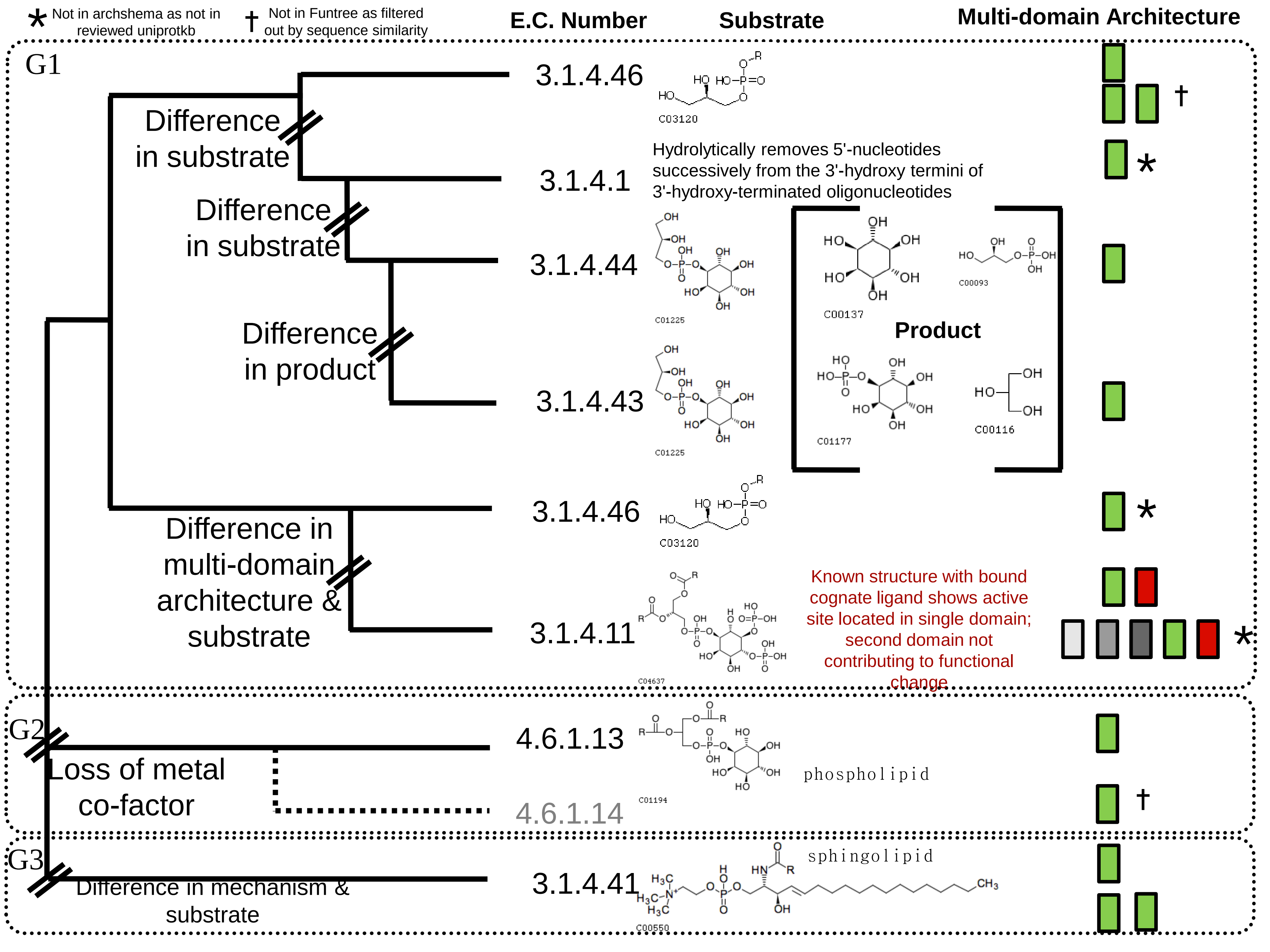
Phosphatidylinositol-Phosphodiesterase Superfamily



Phosphatidylinositol-Phosphodiesterase Superfamily

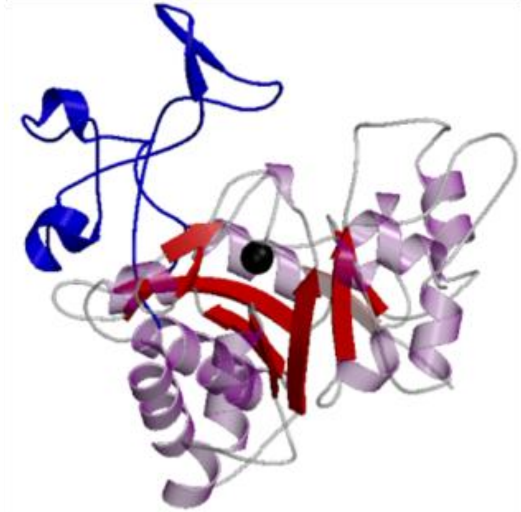
EMBL-EBI



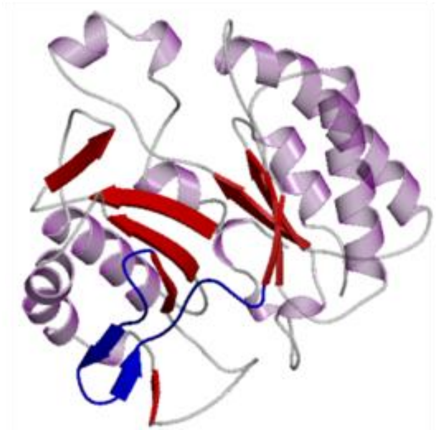


* Not in archshema as not in reviewed uniprotkb
† Not in Funtree as filtered out by sequence similarity
E.C. Number
Substrate
Multi-domain Architecture
G1
Difference in substrate
Difference in substrate
Difference in product
Difference in multi-domain architecture & substrate
3.1.4.46
C03120
3.1.4.1
Hydrolytically removes 5'-nucleotides successively from the 3'-hydroxy termini of 3'-hydroxy-terminated oligonucleotides
3.1.4.44
C01225
3.1.4.43
C01225
Product
C00137
C00093
C01177
C00116
3.1.4.46
C03120
3.1.4.11
C04637
Known structure with bound cognate ligand shows active site located in single domain; second domain not contributing to functional change
G2
Loss of metal co-factor
4.6.1.13
4.6.1.14
C01194
phospholipid
G3
Difference in mechanism & substrate
3.1.4.41
sphingolipid
C00550

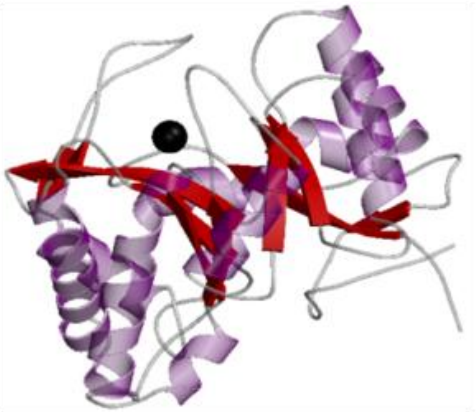
Phosphatidylinositol-Phosphodiesterase Superfamily



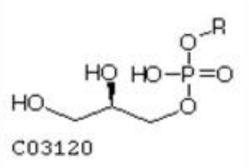
Eukaryotes (clade 1)



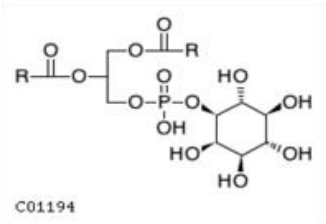
Bacteria (clade 2)



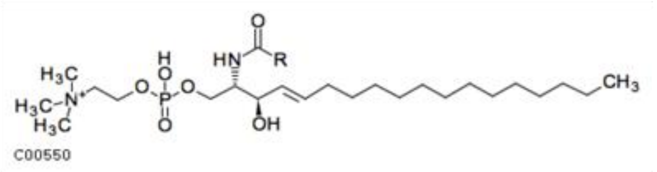
Spider venom (clade 3)



Glycerophosphodiester

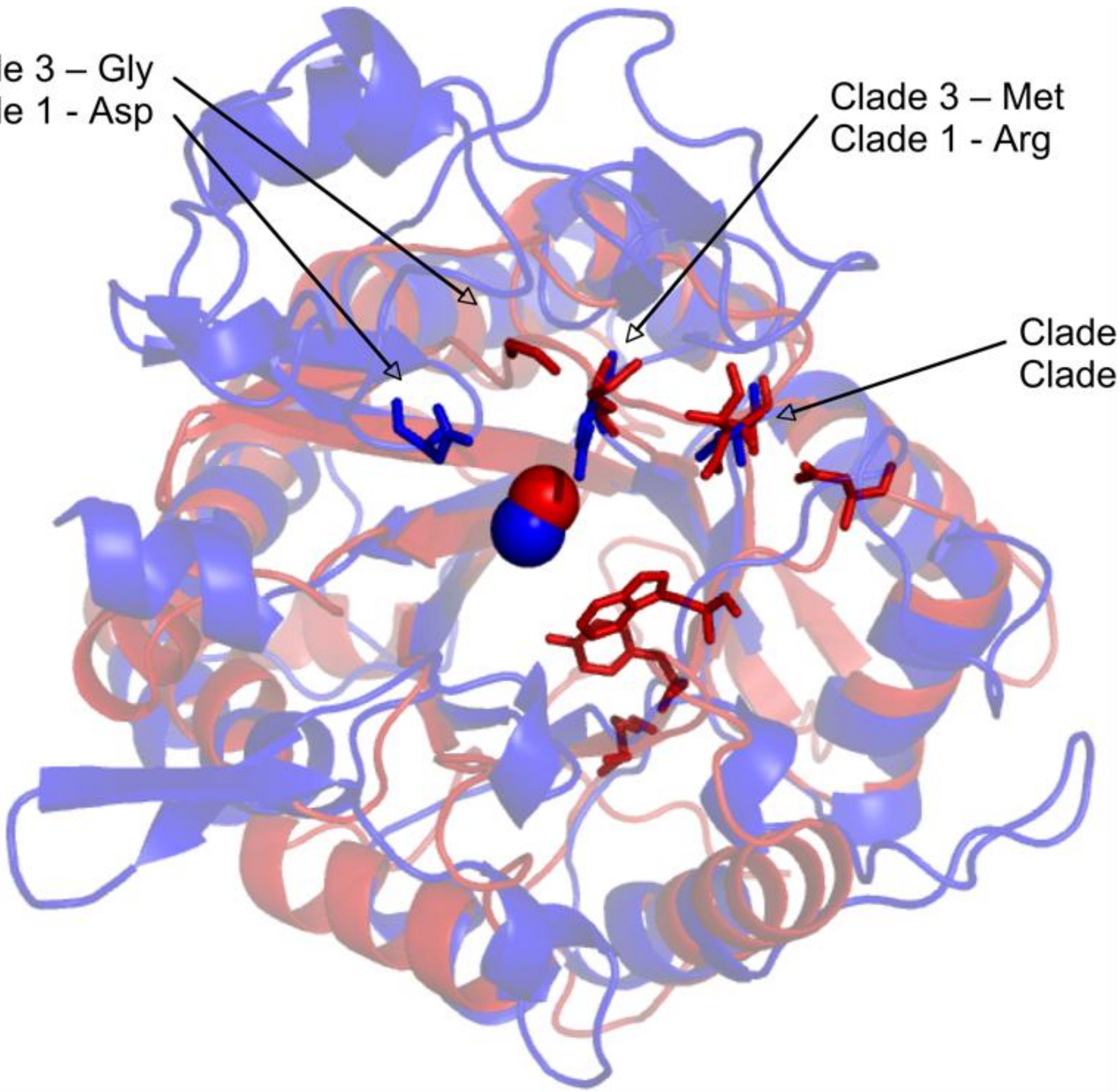


1-phosphtidyl-1D-myo-inositol



Sphingomyelin

Phosphatidylinositol-Phosphodiesterase Superfamily



Legend
Spider venom (Clade 3) – red
Eukaryote (Clade 1) - blue

Enzyme Domains & Superfamilies

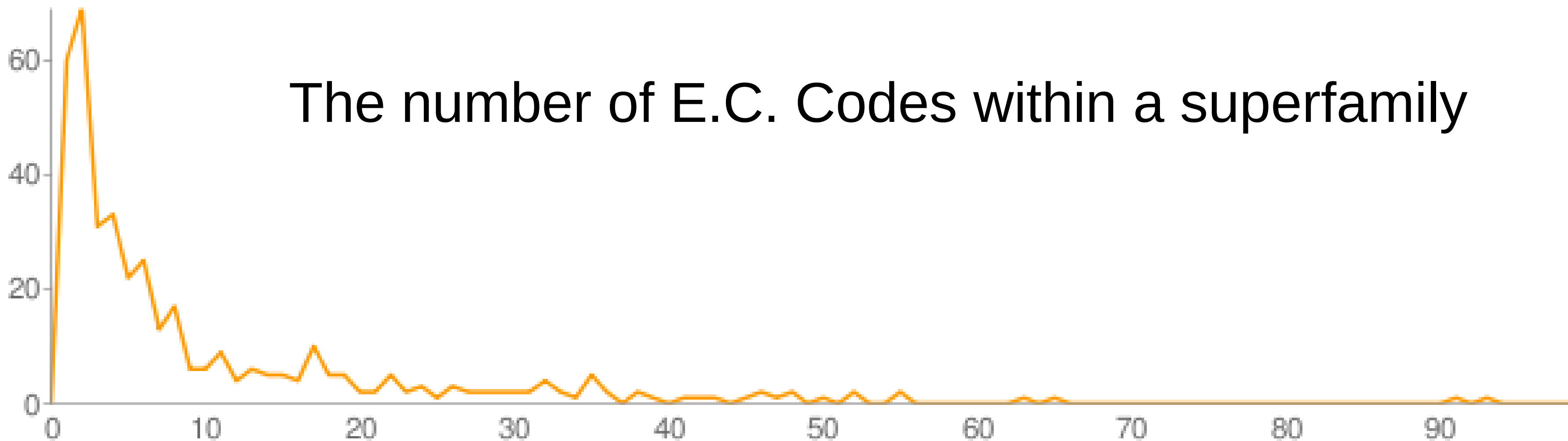
To test we started with an analysis of 6 superfamilies
(based on SFLD database from Babbitt group):

Haloacid dehalogenase
Terpene Cyclases
Amidohydrolase
Crotonase
Enolase
Vicinal Oxygen Chelate

Now we have processed 276 Superfamilies

The superfamilies were chosen using MACiE to identify domains with known catalytic residues.

Data Overview



Changes in enzyme function:-

- Which changes in enzyme function are observed?
- At which level of E.C. Code?
- How do we represent these changes?

E.C. Exchange Matrix

E.C. Changes Using Phylogenetic Trees

EMBL-EBI

Percentage of changes (total number of counts)

	E.C. 1	E.C. 2	E.C. 3	E.C. 4	E.C. 5	E.C. 6
E.C. 1	36.3% (1212)	1.3% (42)	0.8% (28)	0.6% (19)	0.7% (24)	0.2% (7)
E.C. 2		12.6% (421)	1.6% (52)	1.6% (52)	0.5% (18)	0.5% (16)
E.C. 3			25.1% (837)	0.8% (27)	0.5% (17)	0.2% (8)
E.C. 4				9.4% (314)	1.0% (33)	0.4% (13)
E.C. 5					2.4% (79)	0.1% (4)
E.C. 6						3.4% (112)

Total Number within class changes 2967 (89%)

Total Number between class changes 360 (11%)

CONCLUSIONS

- New functions emerge by local domain evolution and domain fusions

- Evolution of enzyme function occurs within most superfamilies
  - Changes within a class dominate – ie changes of specificity
  - Changes between EC primary classes do occur, but much more rarely – some changes are more common than expected

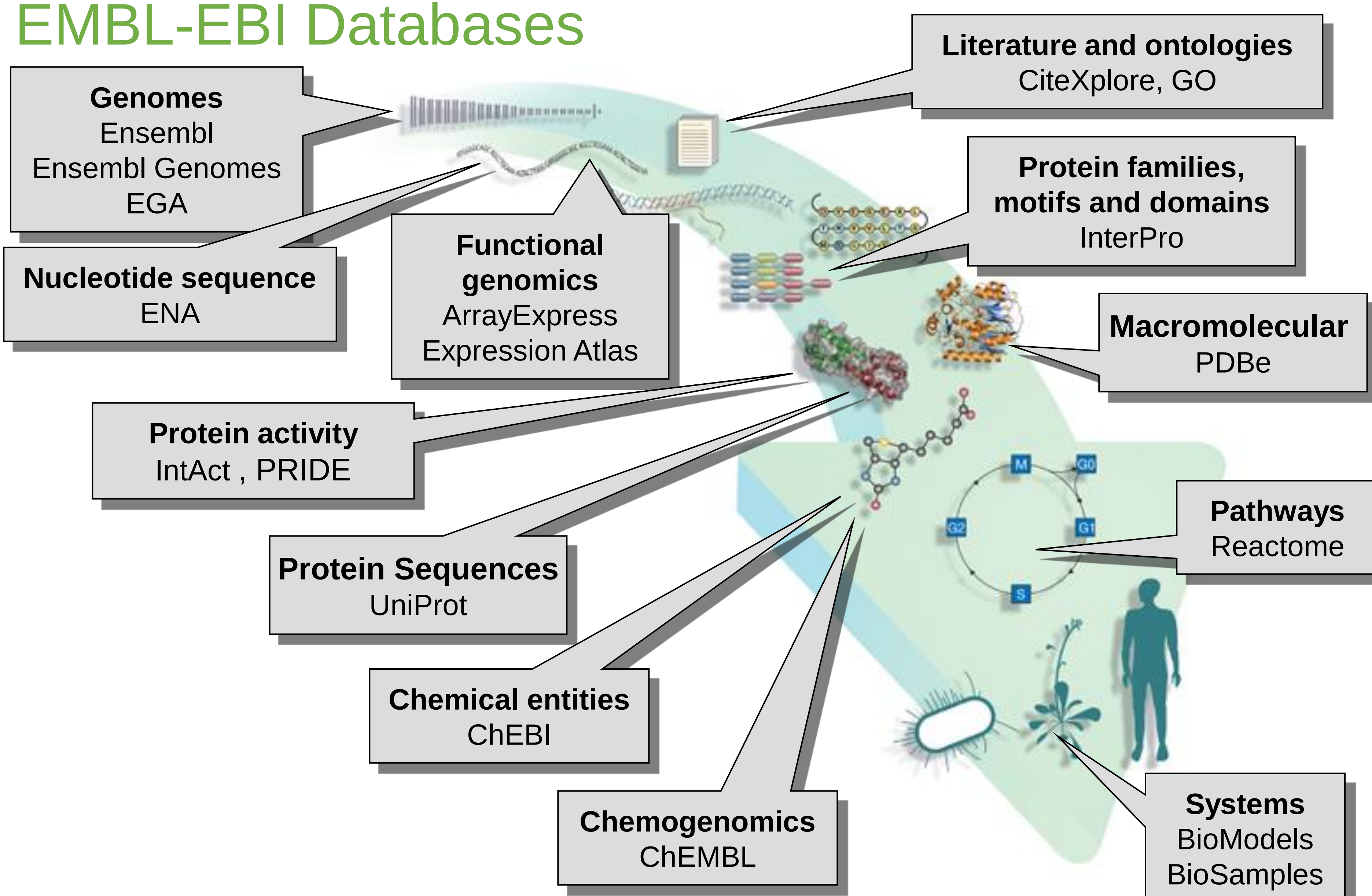
- Small number of families cover majority of reactions
  - Small no. of primordial enzymes sufficient for life?

- Most changes in reaction chemistry are observed in very distantly related enzymes (ancient changes?)
  - Changes in specificity at leaves of trees
  - Changes in reaction chemistry at ‘root’ of trees

Challenges for the PDB (from Gerard)

- Growth
  - Number, size, complexity of entries
  - Hybrid, low-resolution methods
  - From molecular to cellular structural biology
  - User base!
- Validation
- Integration
- From structural biology archive to biomedical resource
  - Best-practice models *versus* published models
  - New ways of accessing and using structural information

EMBL-EBI Databases



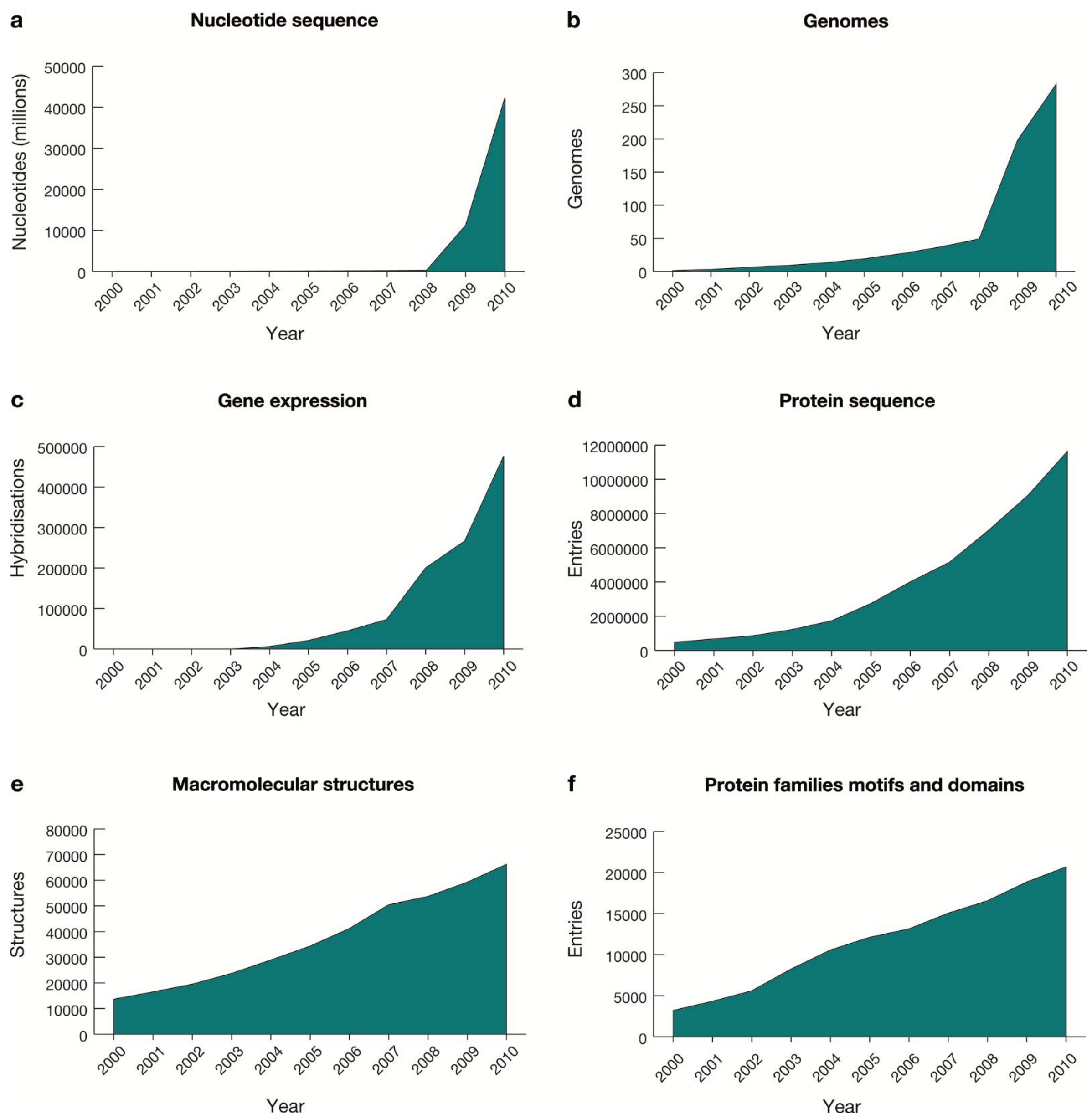
Growth of EBI Databases 2000-2010*

All resources are growing rapidly

Data doubling every 5 months

12 petabytes data storage

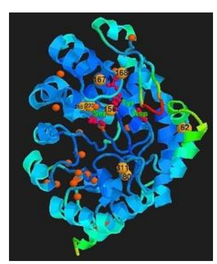
CHALLENGE:
DATA => KNOWLEDGE



More Data

- Structural data:
  - More data
  - RNA
  - Membrane proteins
  - Protein complexes
  - FEL Data (Dynamics)
- Other data
- Integration of data
- ??

NGS Data
Human Variation Data
Links to disease phenotypes
HT Cell Biology
HT Light microscopy
EM Tomography
DNA in 3D
Large protein machines
?

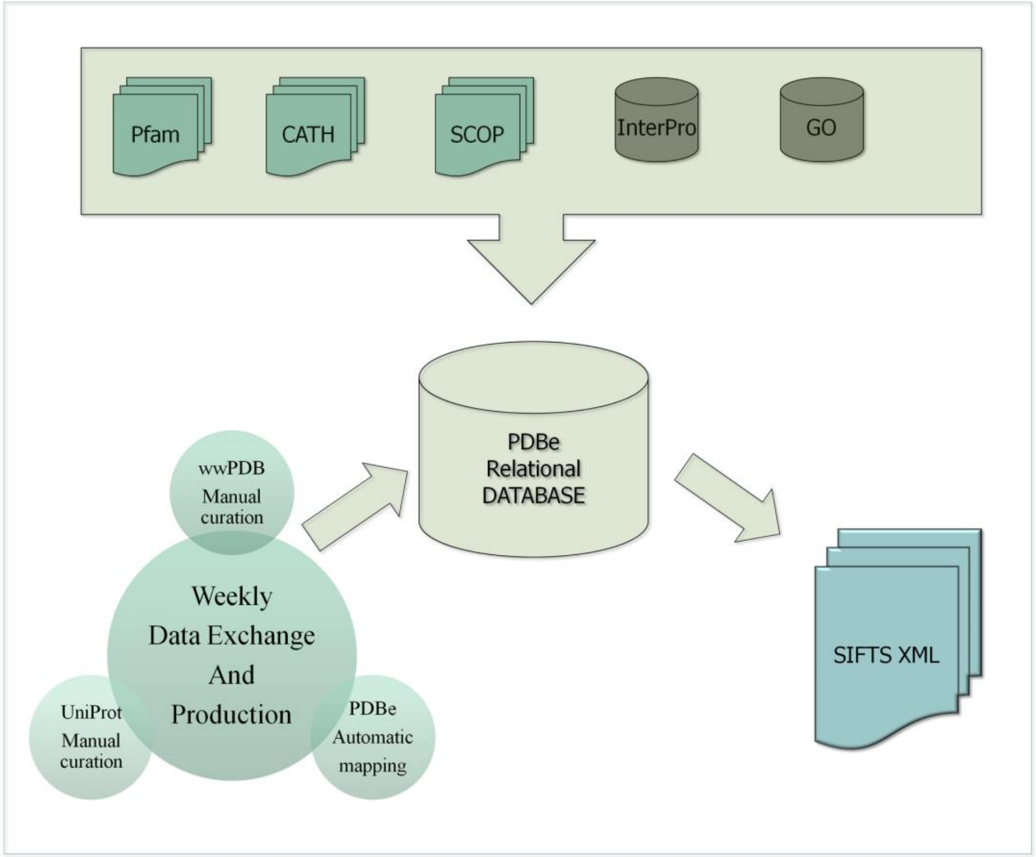


Uroporphyrinogen decarboxylase (1uro)

Heme biosynthesis pathway

Porphyria cutanea tarda

Data Integration: PDB ⟷ Sequences
SIFTS



Used by:

- wwPDB
-UniProt
-Pfam
-PDBe
-RSCB
-SCOP
-CATH
-PDBsum
-…

PLEA FOR MORE FUNCTIONAL DATA IN PDB TO FACILITATE KNOWLEDGE EXTRACTION:
Capturing knowledge learnt from structure into the PDB, using agreed standards, vocabularies and ontologies:

- **Simple things**:
  - Experimental protocols
  - Function of protein
  - Function of ligand
    eg inhibitor/crystallisation aid
  - Functional highlights of structure – biological consequences
  - Role of dynamic movement
  - Relationship to other structures in PDB
- **More complex**:
  - Protein localisation
  - Catalytic site for enzyme
  - Binding site for receptor
  - Mechanism of enzyme
  - Effects of Mutations
  - Interaction partners/pathway context
  - Disease relationships

THANKS to

- All Structural Biologists, who deposit in PDB
- Original Founders of PDB
- Current and past leaders of PDB
- All staff of wwPDB