Introduction to protein structure analysis and prediction

Mónica Chagoyen monica.chagoyen@cnb.csic.es

Protein sequence analysis and prediction service

Centro Nacional de Biotecnologia (CNB-CSIC)

24-26 October 2011

Course organization and contents

Day 1:
The protein structure universe, resources and visualization

Day 2:
Structural alignment, classification and 1D prediction

Day 3:
3D structure prediction

The protein universe

There are four levels of protein structure

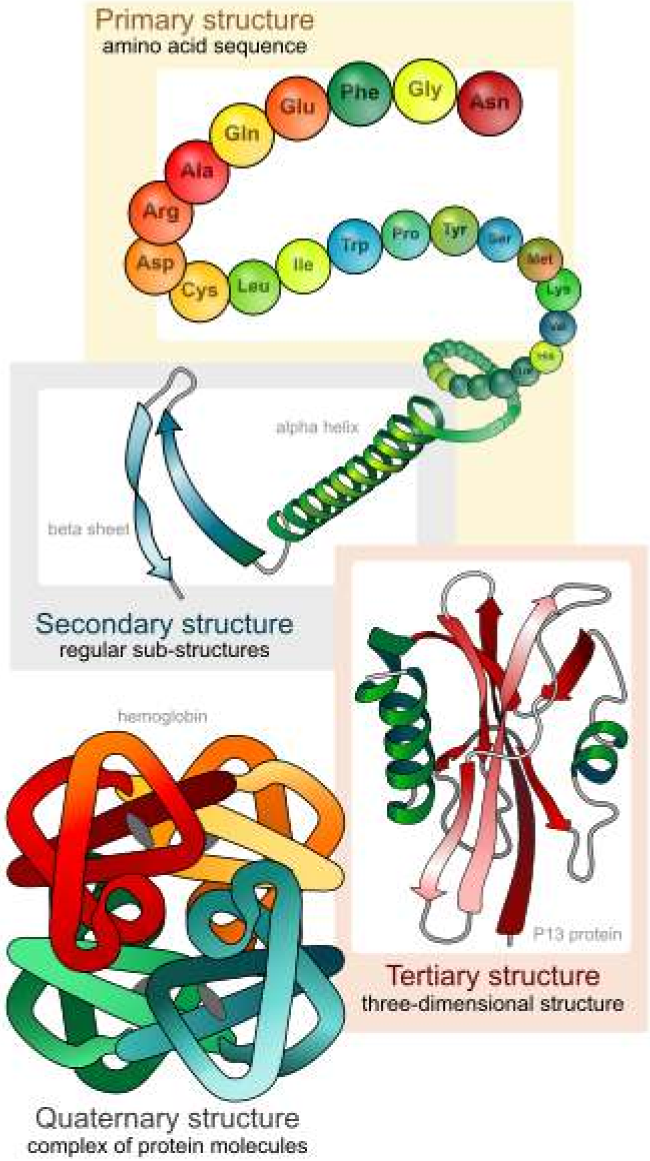
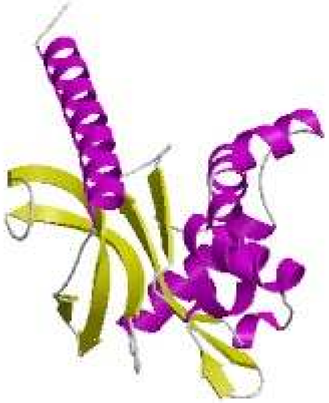
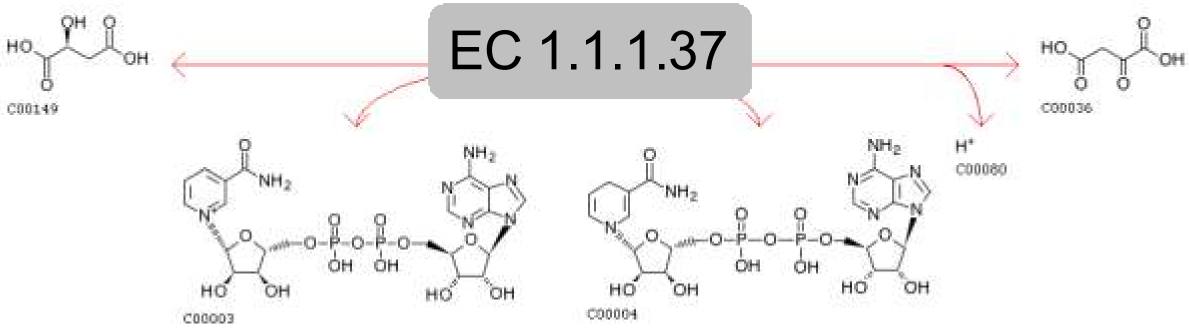


Image: wikipedia

The sequence-structure-function paradigm

LTRLDHNRAKAQIALKLGVTSDDVKNVIIWGNHSSTQYPDVNHAKVKLQAKEVGVYEAVKDDSWLKGEFITTVQQRGAAVIKARKLSSAM
SAAKAICDHVRDIWFGTPEGEFVSMGIISDGNSYGVPDDLLYSFPVTIKDKTWKIVEGLPINDFSREKMDLTAKELAEEKETAFEFLSSA





EC 1.1.1.37

Known protein sequences

UniProt Release 2011_09	Swiss-Prot	532,146
	TrEMBL	16,886,838
		(redundant)

UniRef100 Release 2011_09 consists of 13,992,000 entries

Protein structures are archived in the Protein Data Bank (PDB) since 1971

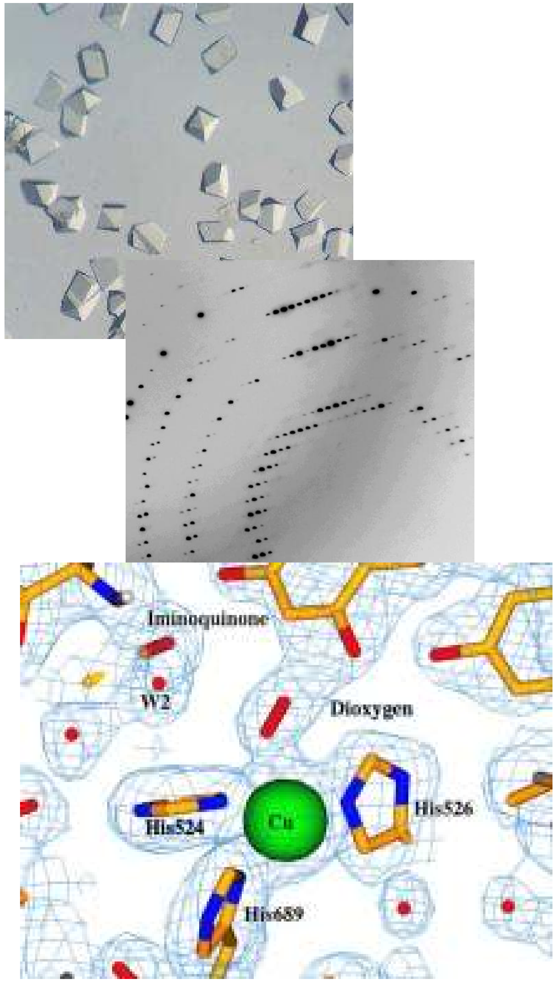
www.wwpdb.org

The mission of the wwPDB is to maintain a single Protein Data Bank archive of macromolecular structural data that is freely and publicly available to the global community.

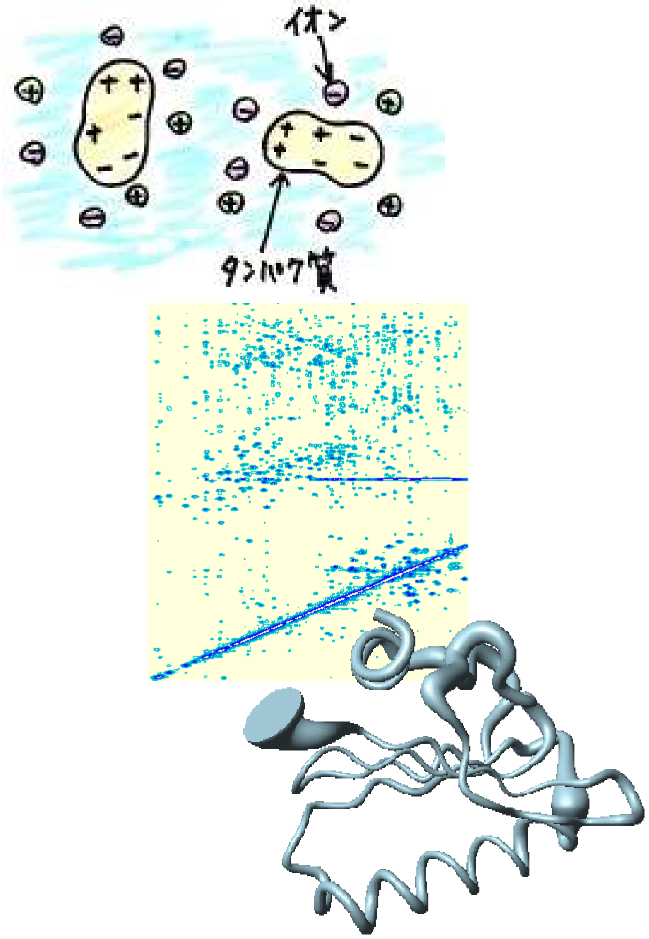
RCSB (USA)

PDBe (Europe)

PDBj (Japan)

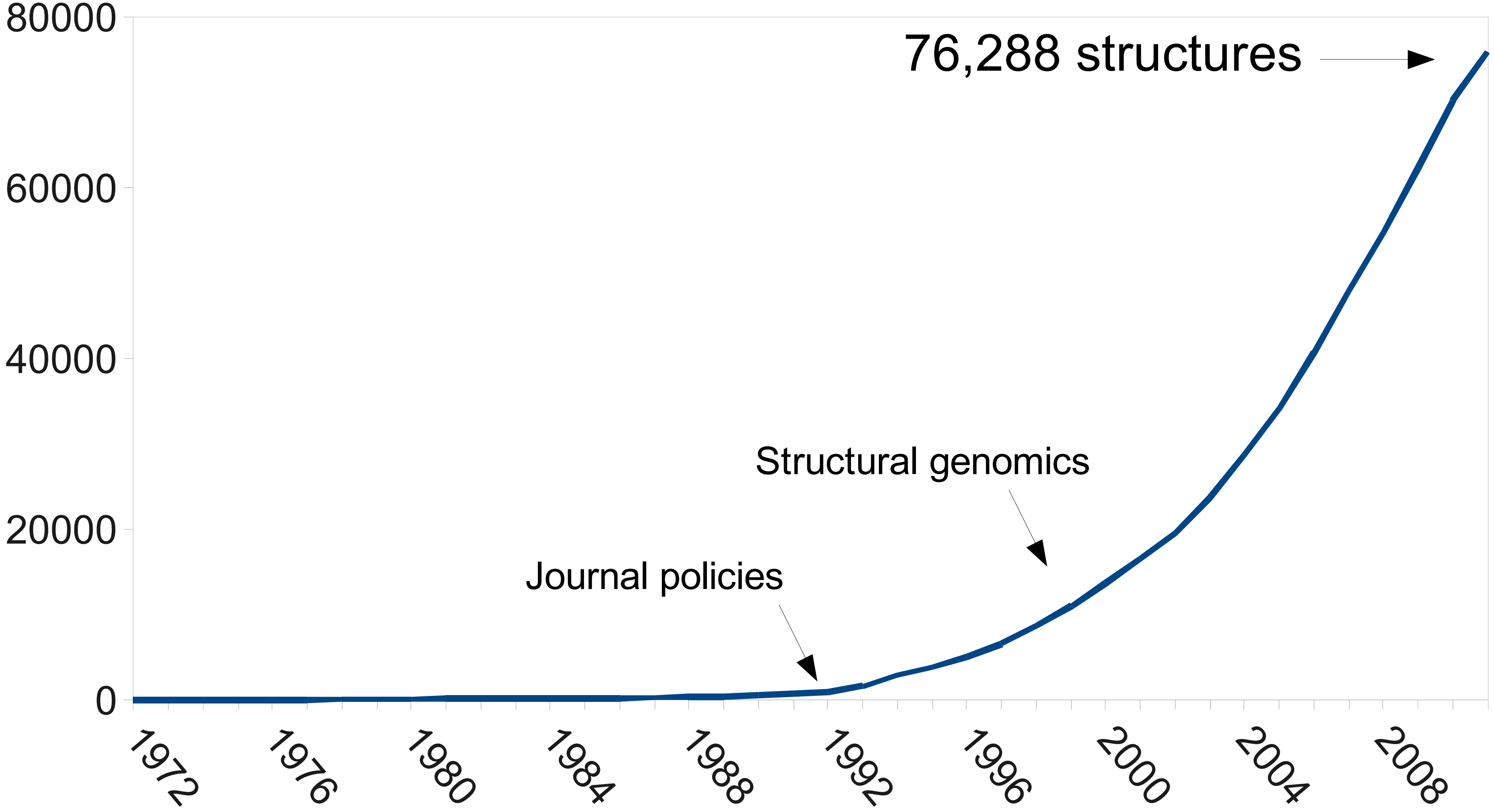


87% X-ray crystallography



12% NMR spectroscopy

Total number of PDB structures per year



76,288 structures

Structural genomics

Journal policies

43,535

(non-redundant, 100% identity)

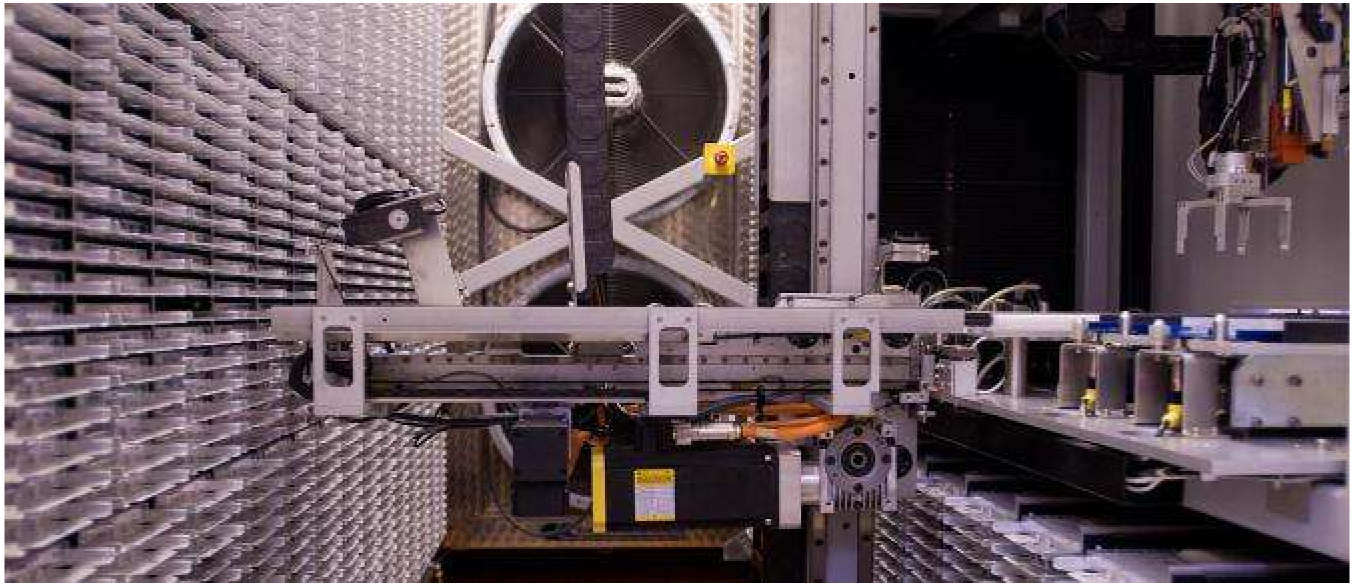
Structural genomics

Large scale determination and analysis of three-dimensional structures

To determine by **experimental methods** a **representative set** of macromolecular structures, including medically important human proteins and proteins from important pathogens and model organisms.

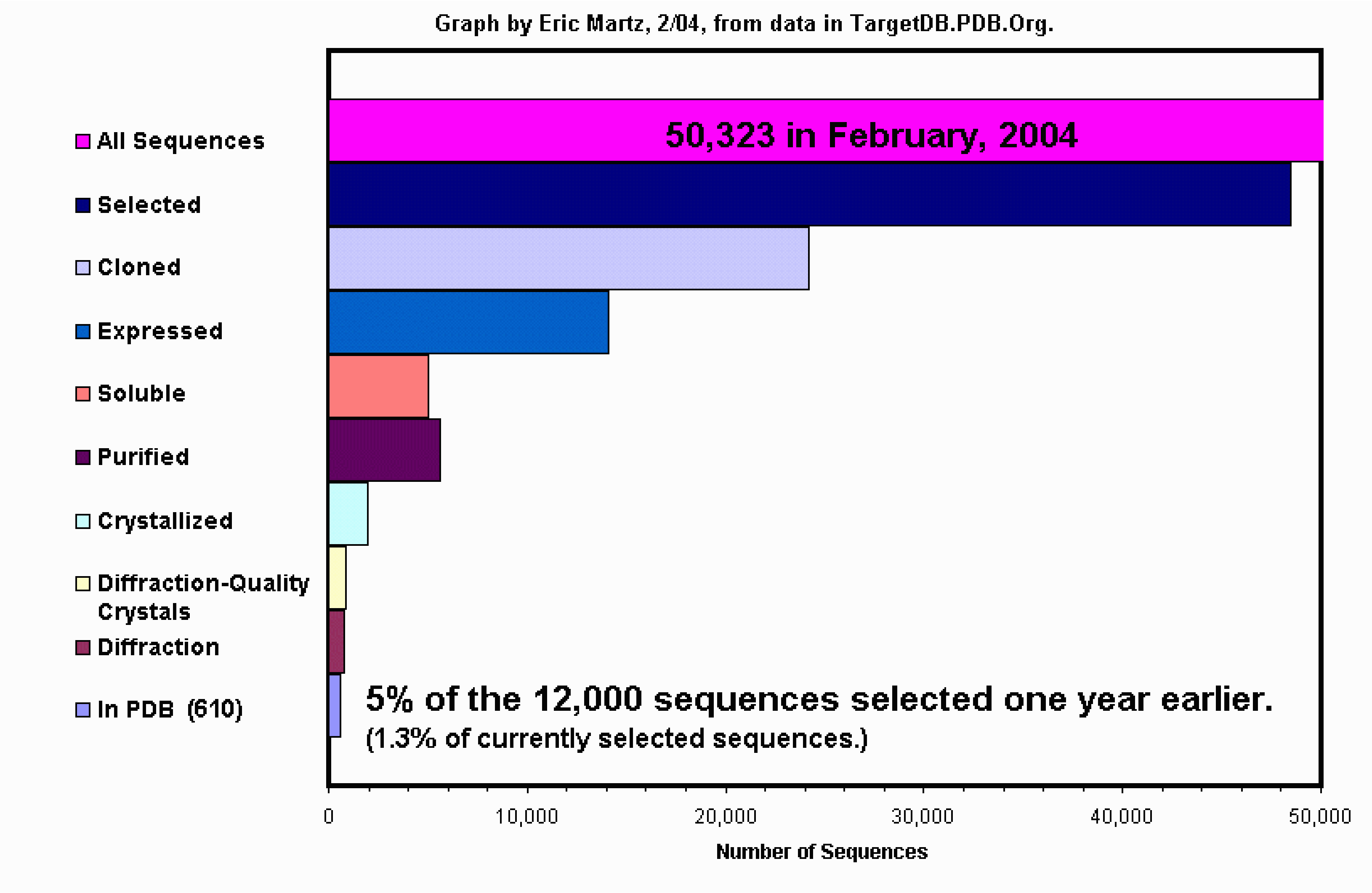
To provide **models** based on sequence similarity to significantly extend the coverage of structure space.

To derive **functional information** from these structures by experimental and computational methods.



Text source: www.isgo.org
Image: Instruct project

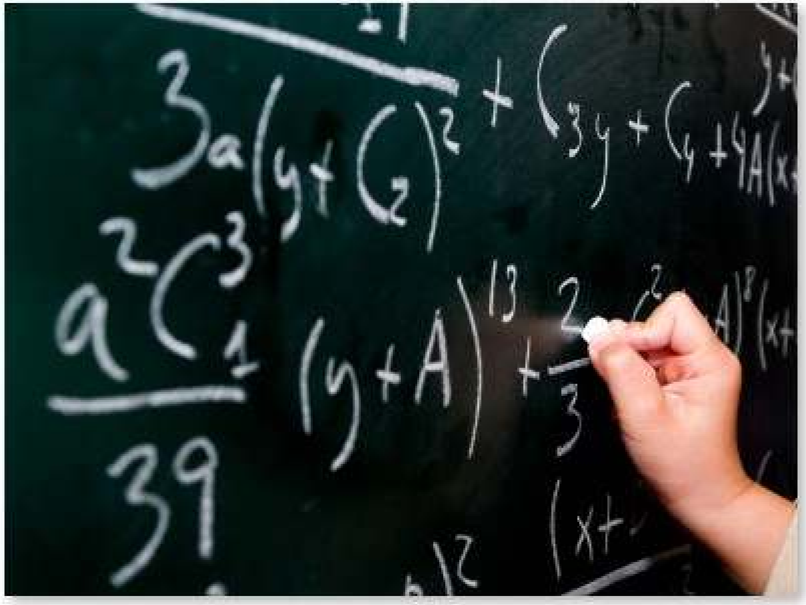
Success rate at each step towards a structure



Sequence – structure relationships

What is the chance that your "favorite" protein is in PDB?

PDB
43,535



UniProt
13,992,000

Known structures for 0.3% of known protein sequences

Some proteins are similar due to evolutionary and/or functional reasons

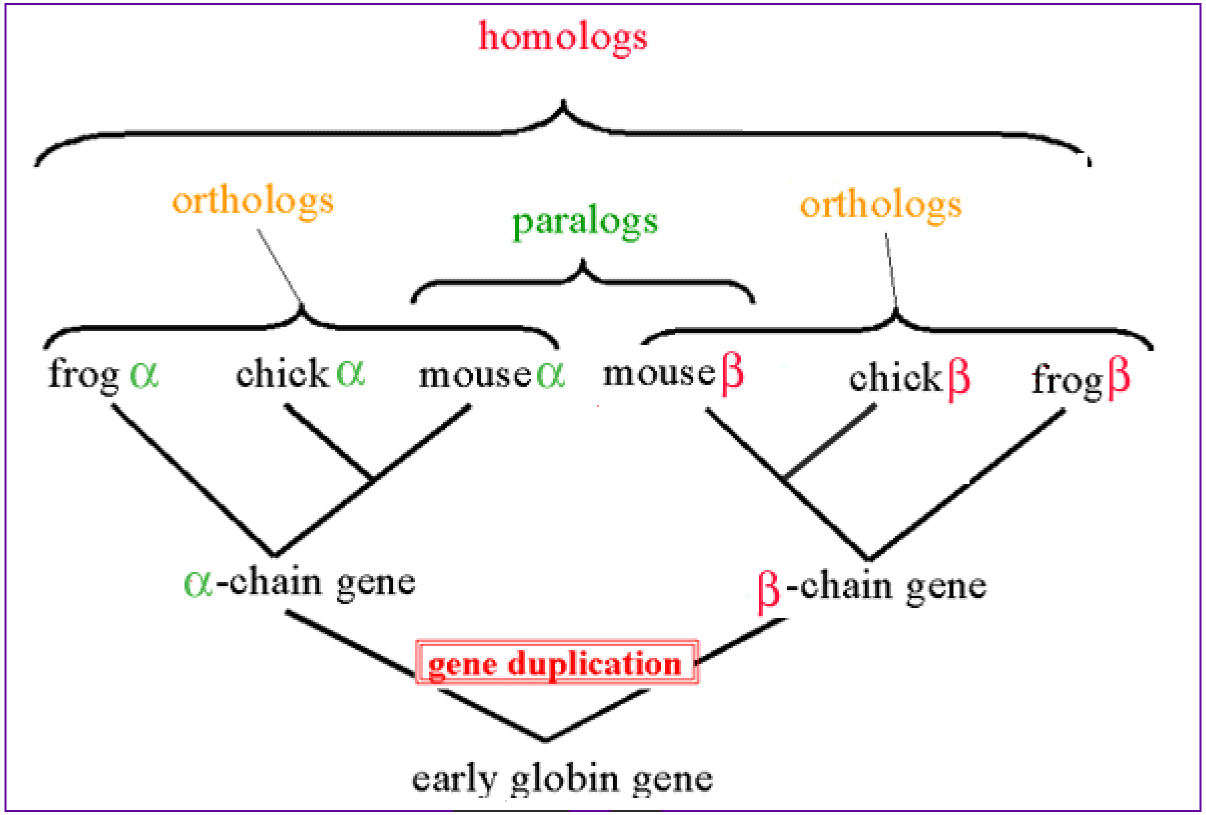
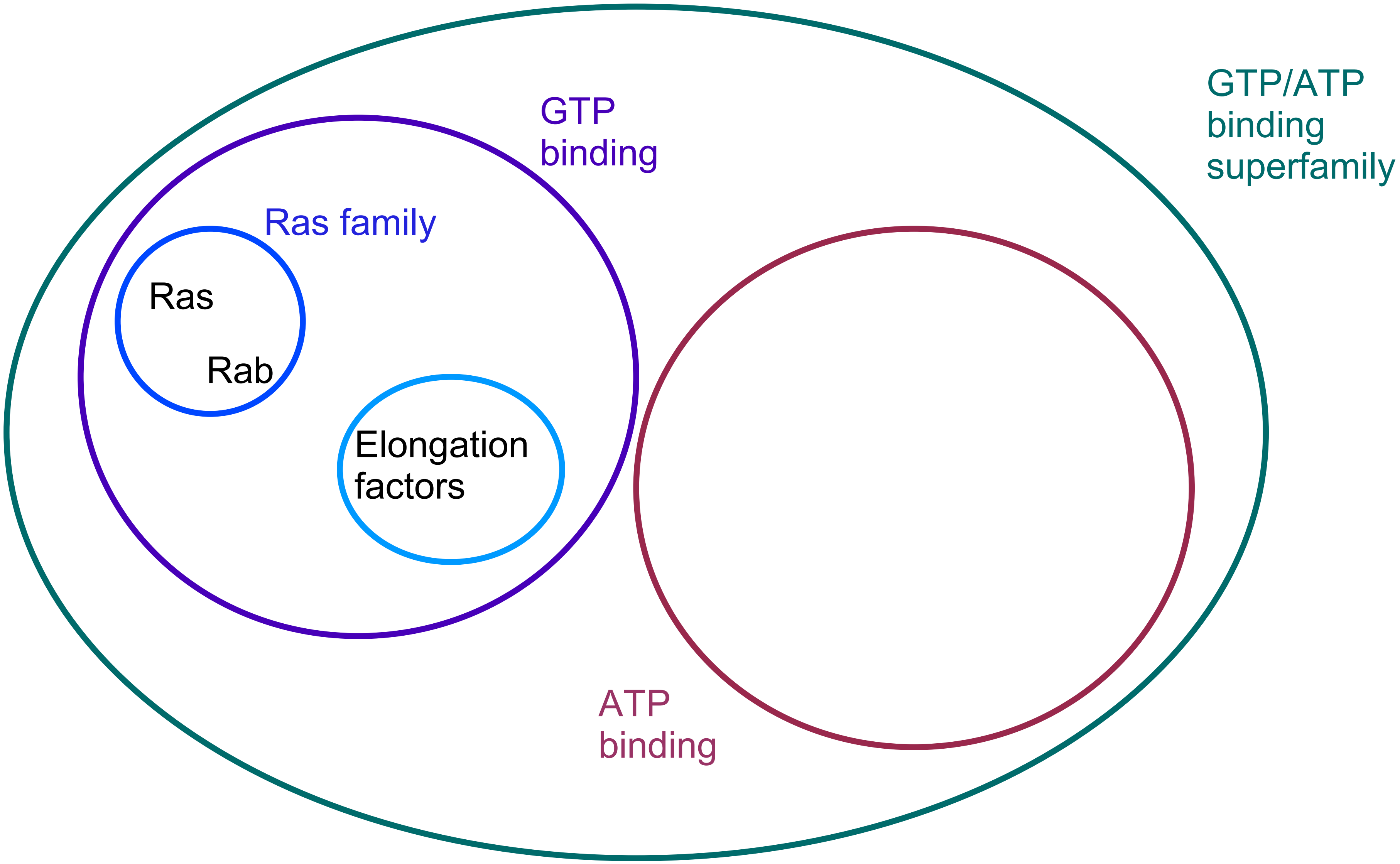


Image: http://www.stanford.edu/group/pandegroup/folding/education/h.html

Protein families and superfamilies



Proteins are composed of domains

Domain = Structural / Evolutionary / Functional unit

Structural
Evolutionary
Functional

Usually described from

sequences

e.g.

Pfam

SMART

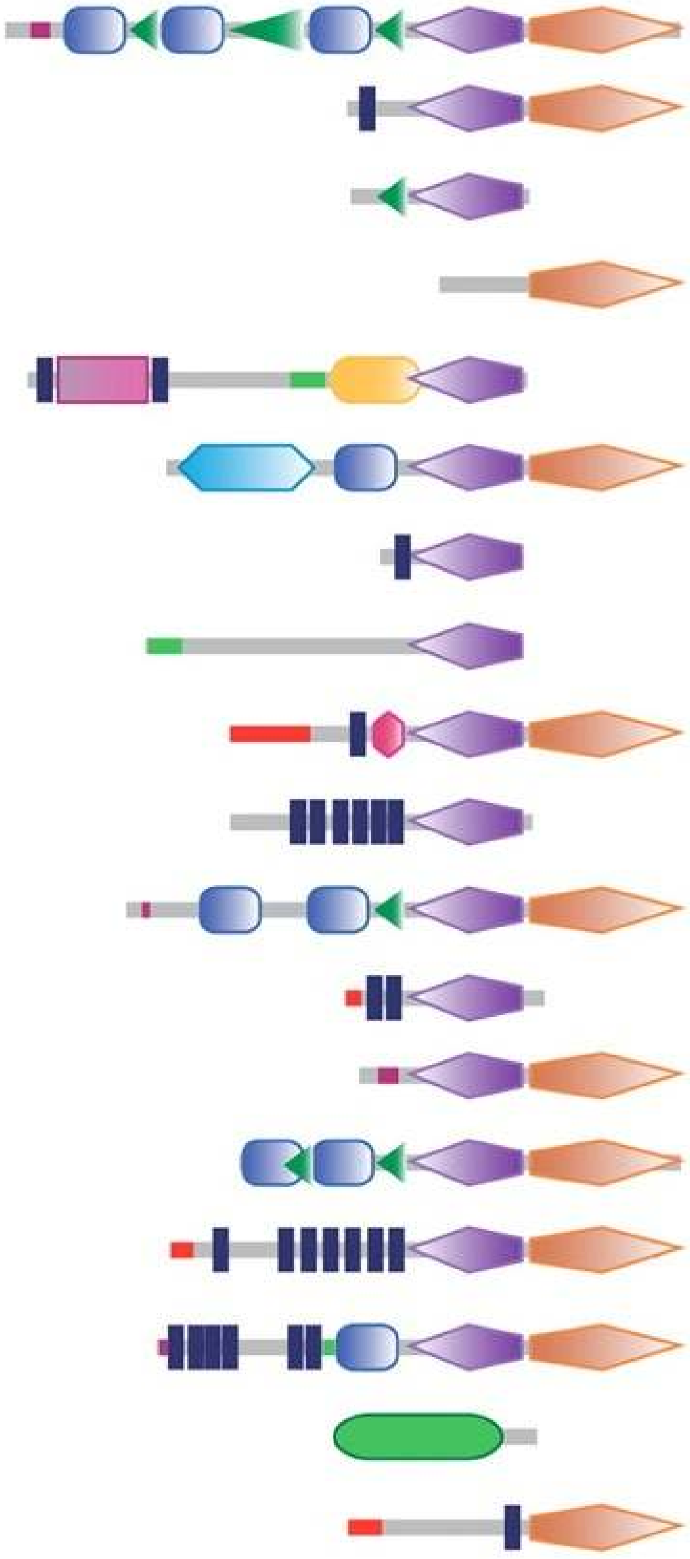
structures

e.g.

SCOP

CATH

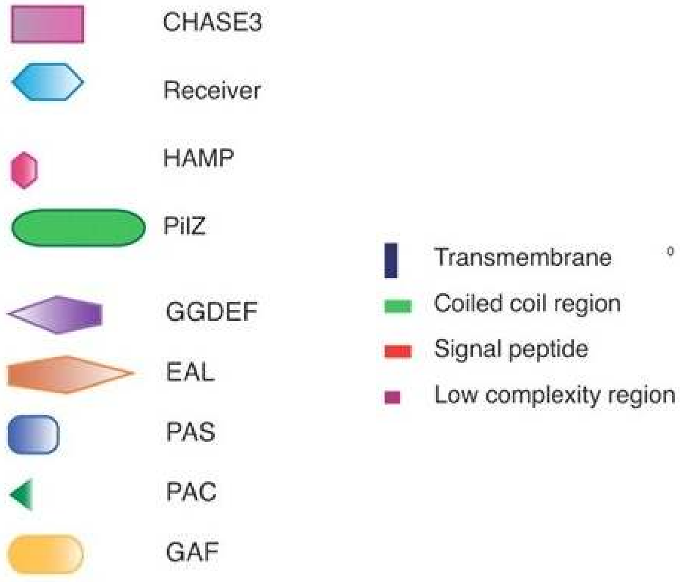
PROTEIN STRUCTURE CLASSIFICATION



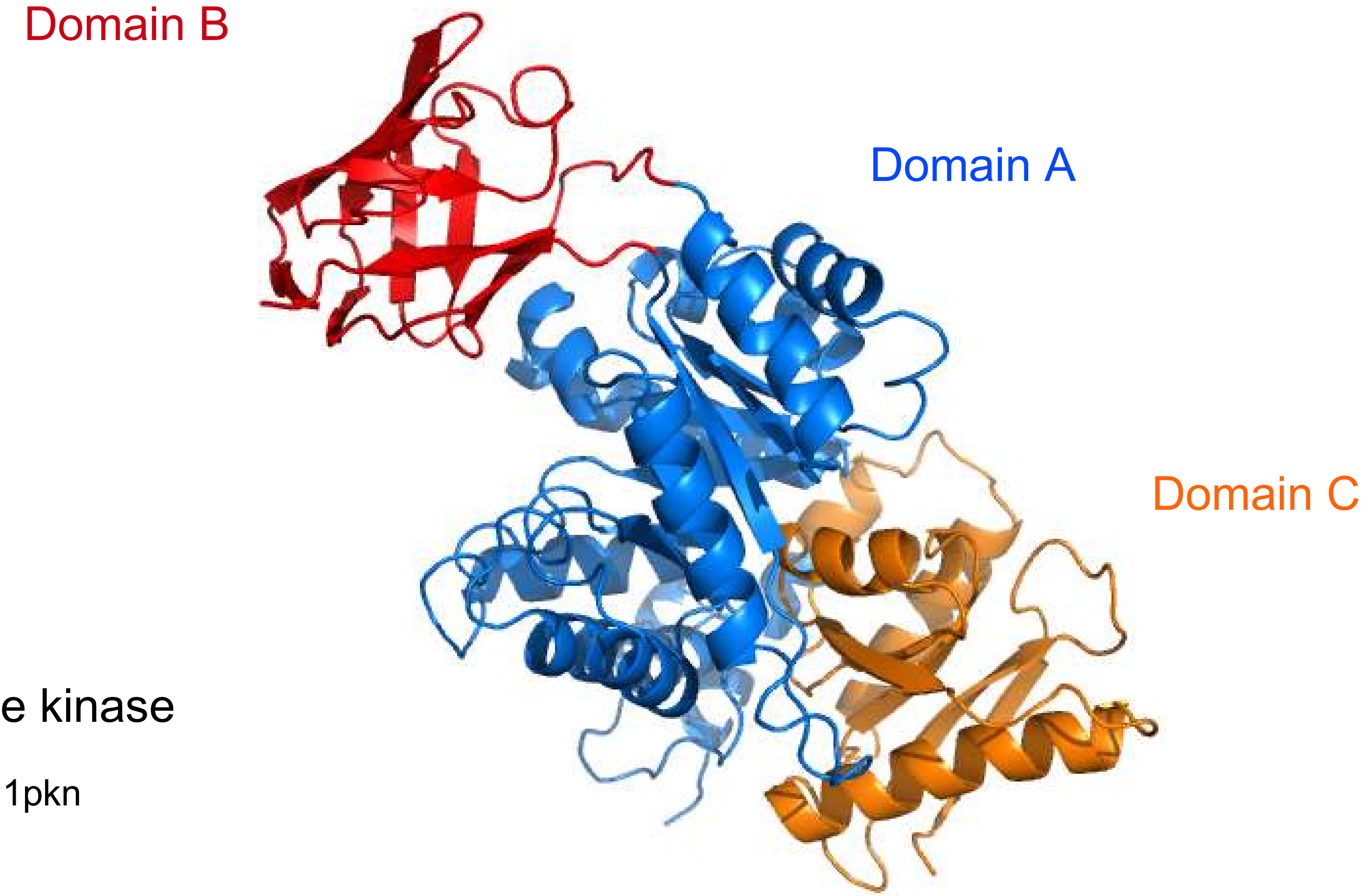
Domain organization

of predicted GGDEF/EAL/PilZ domain proteins in the *Legionella pneumophila Philadelphia-1* genome

Levi, A., M. Folcher, U. Jenal, and H. A. Shuman. 2011. Cyclic diguanylate signaling proteins control intracellular growth of Legionella pneumophila. mBio 2(1):e00316-10

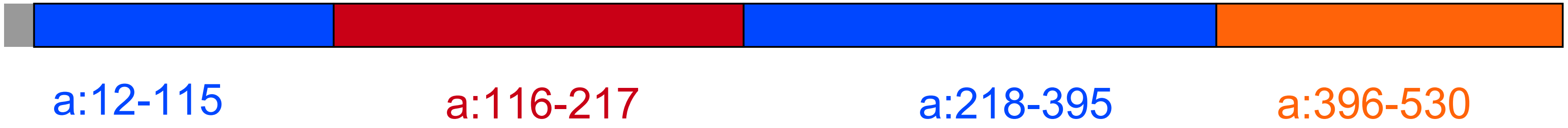


Structural domains

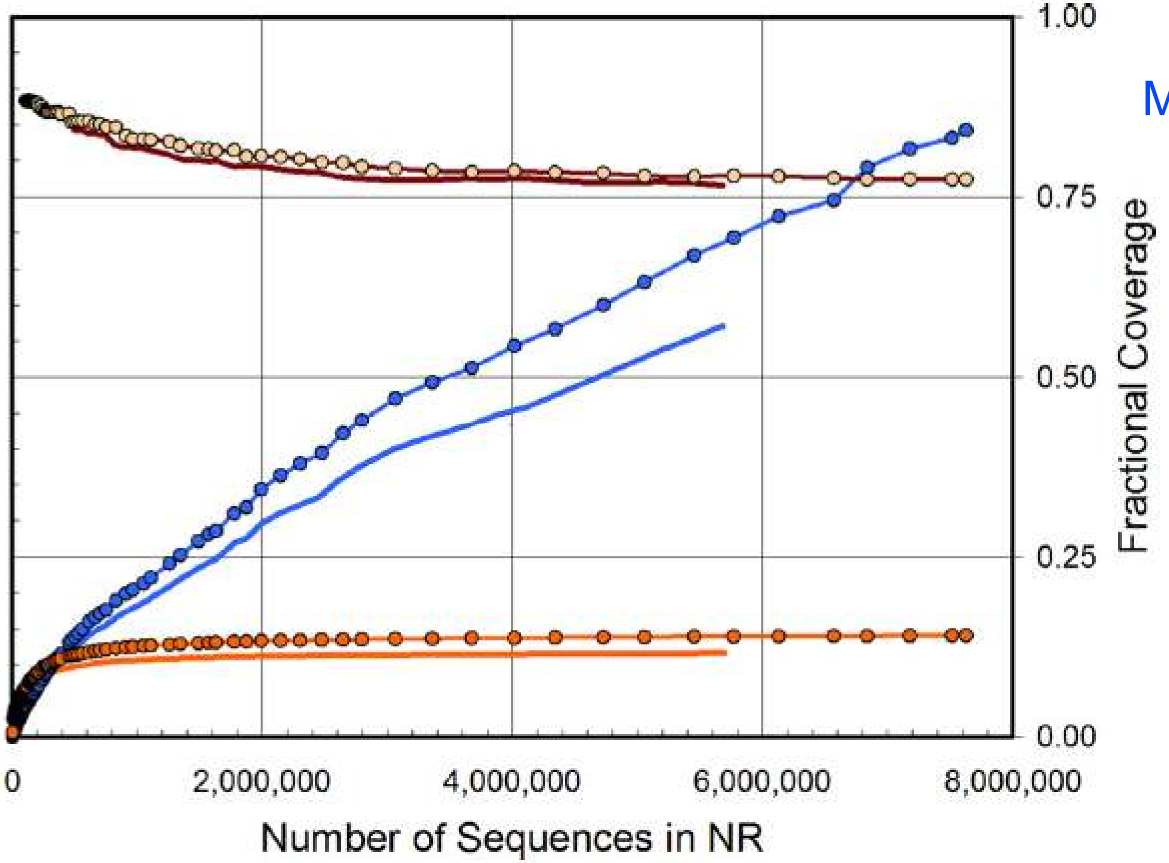


Pyruvate kinase

PDB:1pkn



How many families?



Multiple domain families

Single domain families

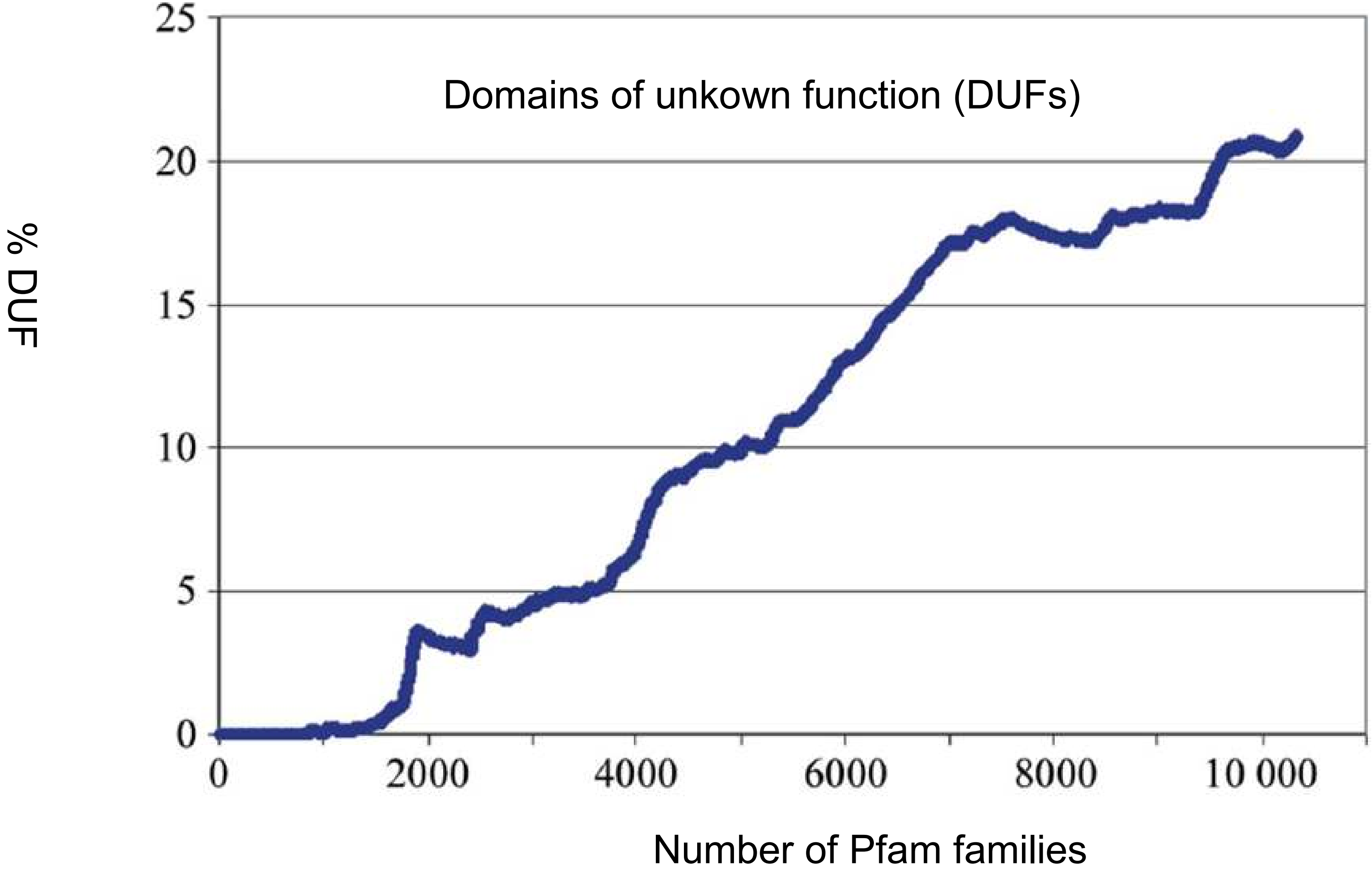
M. Levitt (2009) Nature of the protein universe. *Proc Acad Sci USA* 106:11079

22% of known protein sequences do not match any known domain family

M. Levitt (2009) Nature of the protein universe. *Proc Acad Sci USA* 106:11079

Image: www.spacetelescope.org

Functional coverage



Source:xfam.wordpress.com

Structural coverage

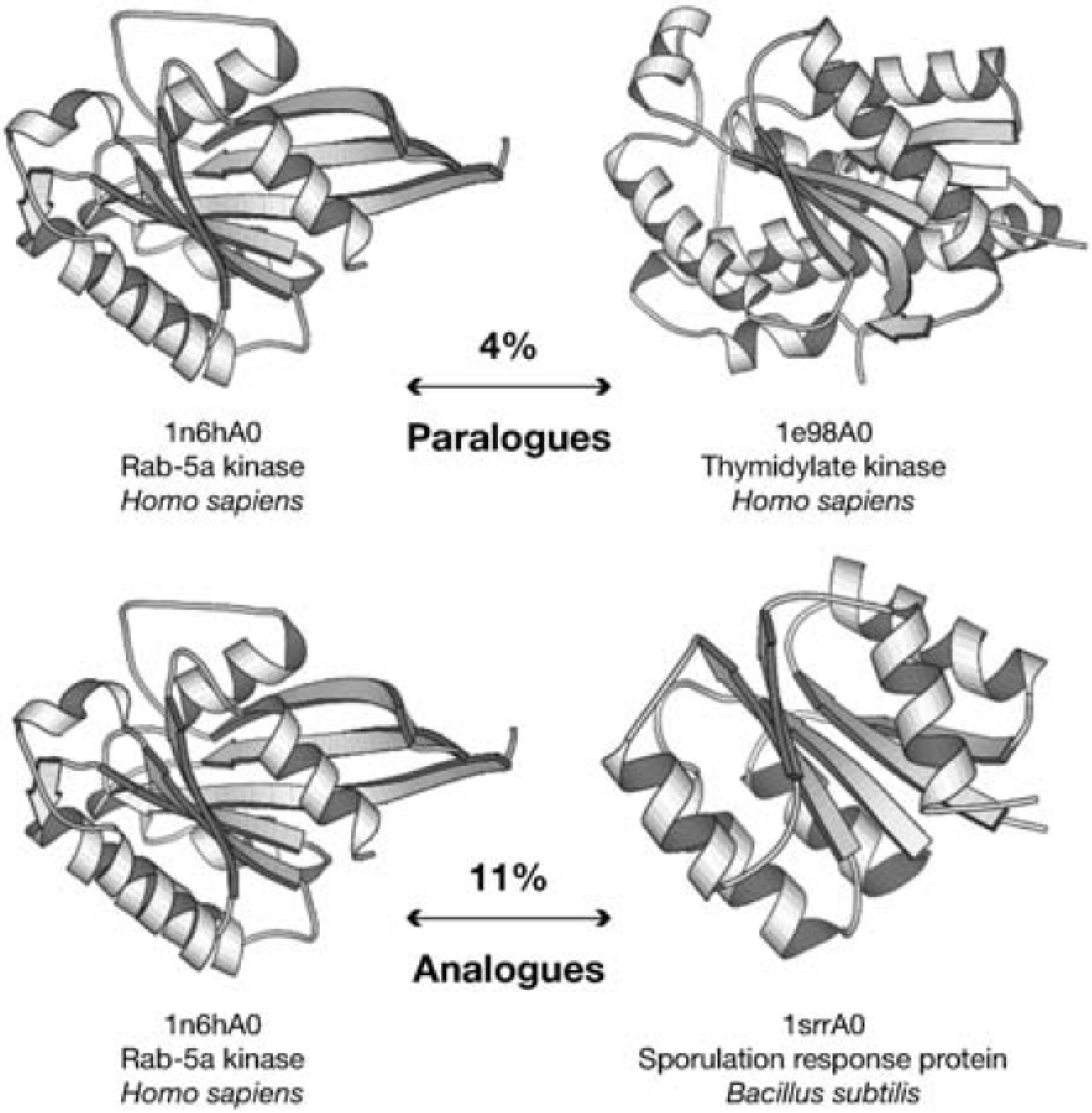
Known structures for 0.3% of known protein sequences

25% of single domain families have at least one structure solved

Expected 85% coverage by 2050

M. Levitt (2009) Nature of the protein universe. *Proc Acad Sci USA* 106:11079

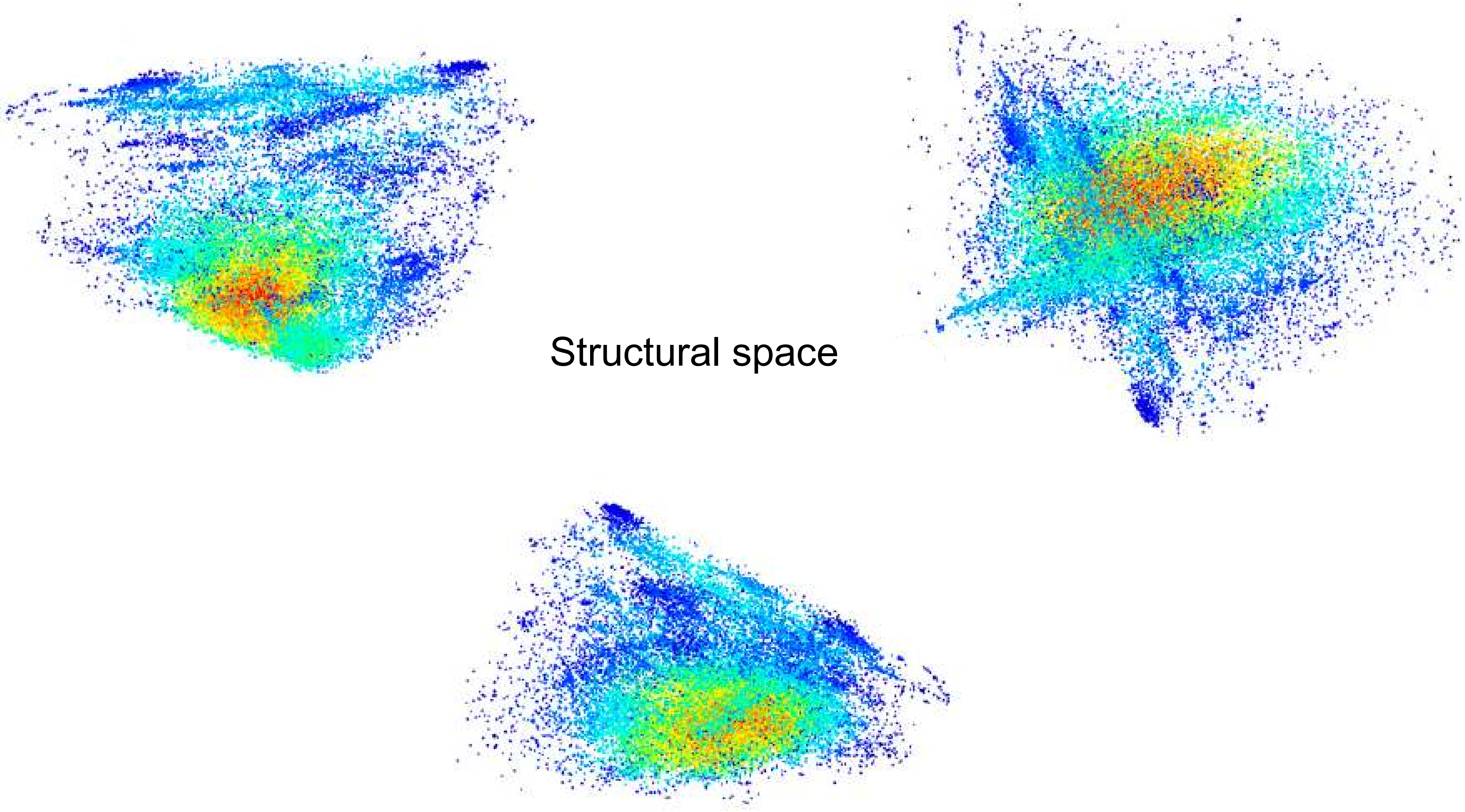
In distantly related proteins, structure is more conserved than sequence



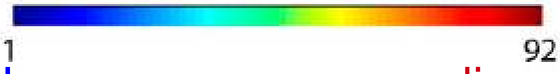
CA Orengo, JM Thorton. Protein families and their evolution – a structural perspective. *Annu. Rev. Biochem.* 2005.74:867-900

Structure – function relationships

Structure space has a core of high functional diversity



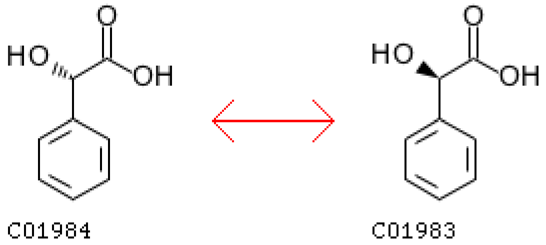
Functional diversity scale



M Osadchy and R Kolodny
Maps of protein structure space reveal a fundamental relationship between protein structure and function
Proc Nat Acad USA 2011, 108:12301-12306

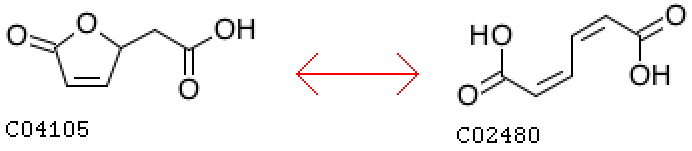
David J. Neidhart, George L. Kenyon, John A. Gerlt & Gregory A. Petsko
Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous
Nature 347, 692 - 694 (18 October 1990); doi:10.1038/347692a0

Mandelate racemase
EC 5.1.2.2



(S)-Mandelate <=> (R)-Mandelate

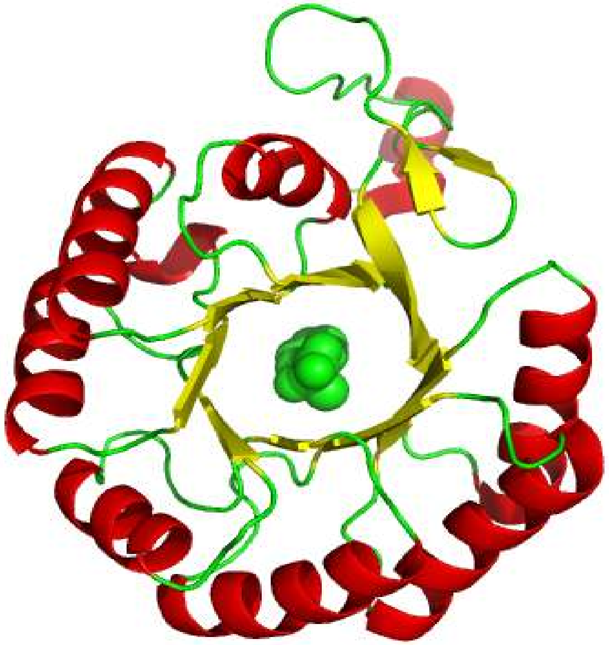
Muconate lactonizing enzyme
EC 5.5.1.1



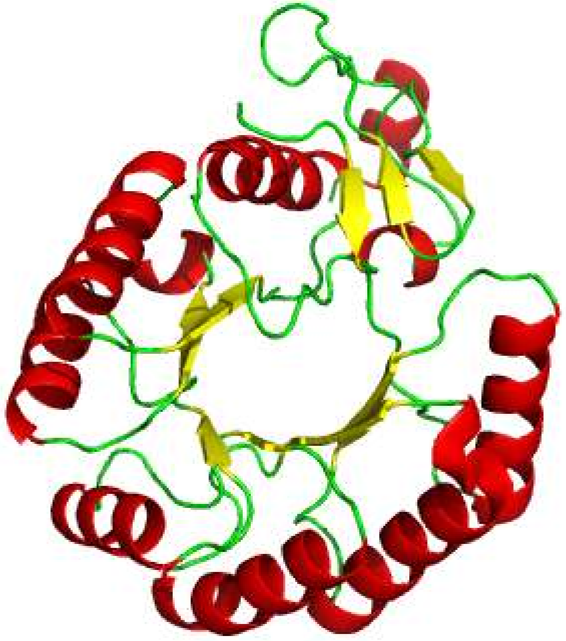
2,5-Dihydro-5-oxofuran-2-acetate <=> cis,cis-Muconate

Reaction and figures: KEGG

David J. Neidhart, George L. Kenyon, John A. Gerlt & Gregory A. Petsko
Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous
Nature 347, 692 - 694 (18 October 1990); doi:10.1038/347692a0



TIM barrel



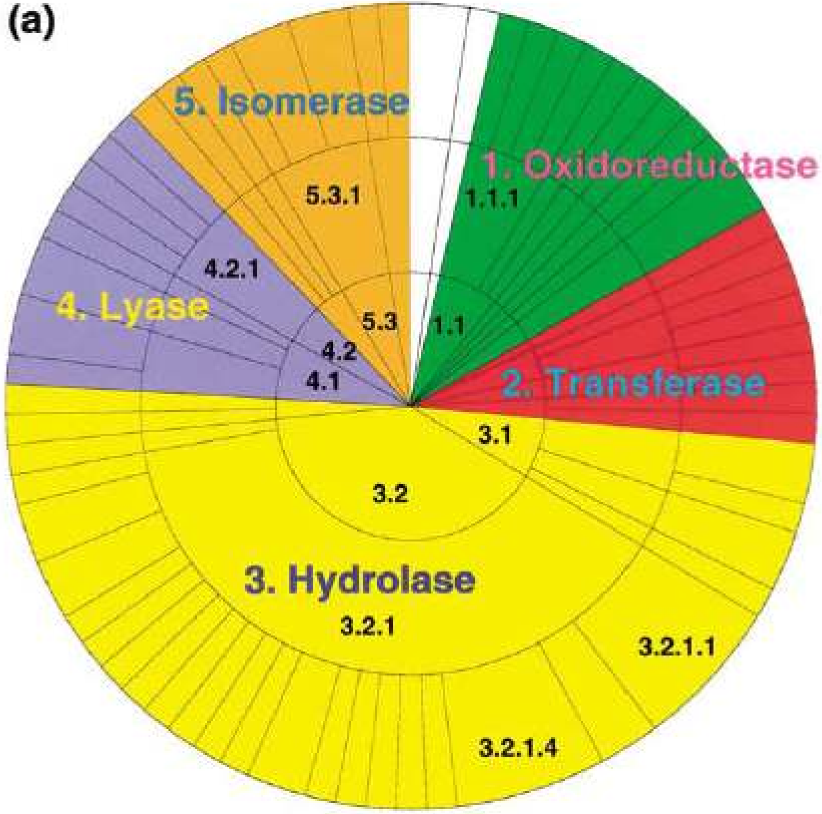
1mra a:133-359

Mandelate racemase
EC 5.1.2.2

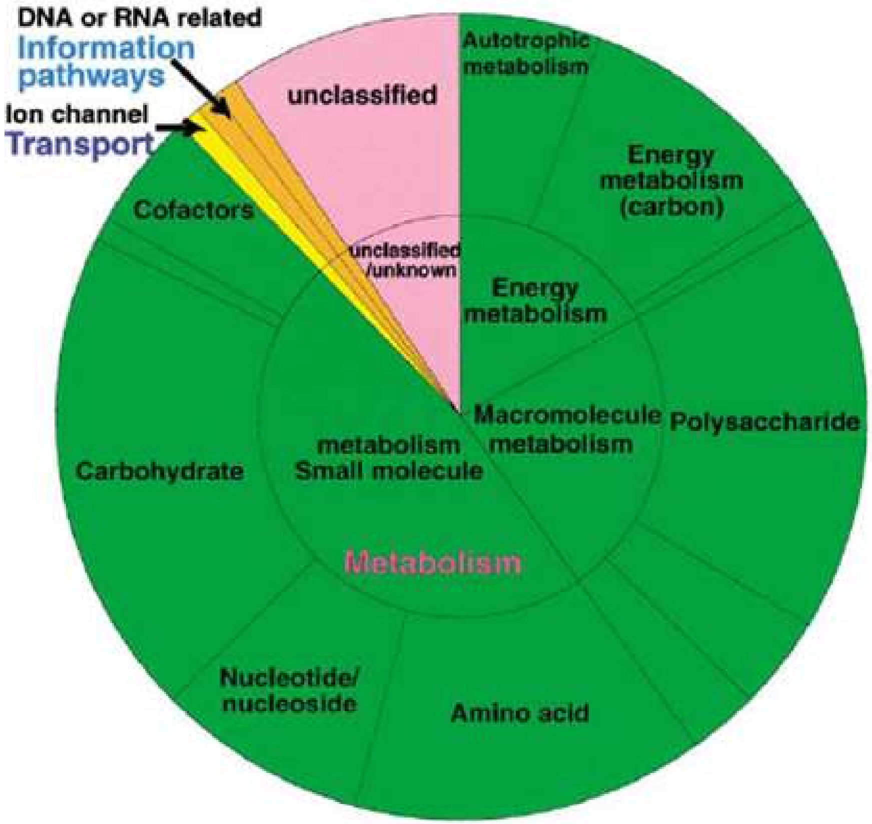
1bkh a:131-372

Muconate lactonizing enzyme
EC 5.5.1.1

TIM barrel functions



Enzymatic



Non-Enzymatic

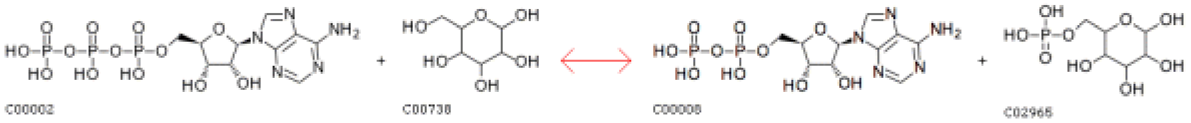
Nozomi Nagano, Christine A. Orengo and Janet M. Thornton
One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions
J. Mol. Biol. (2002) 321, 741–765

P Bork, C Sander, A Valencia
Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases.
Protein Sci. 1993 Jan;2(1):31-40.

Hexokinase
EC 2.7.1.1

D-hexose

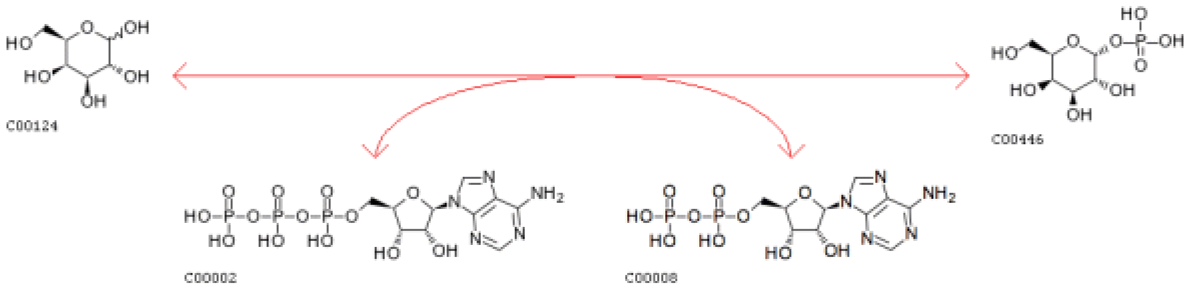
D-Hexose 1-phosphate



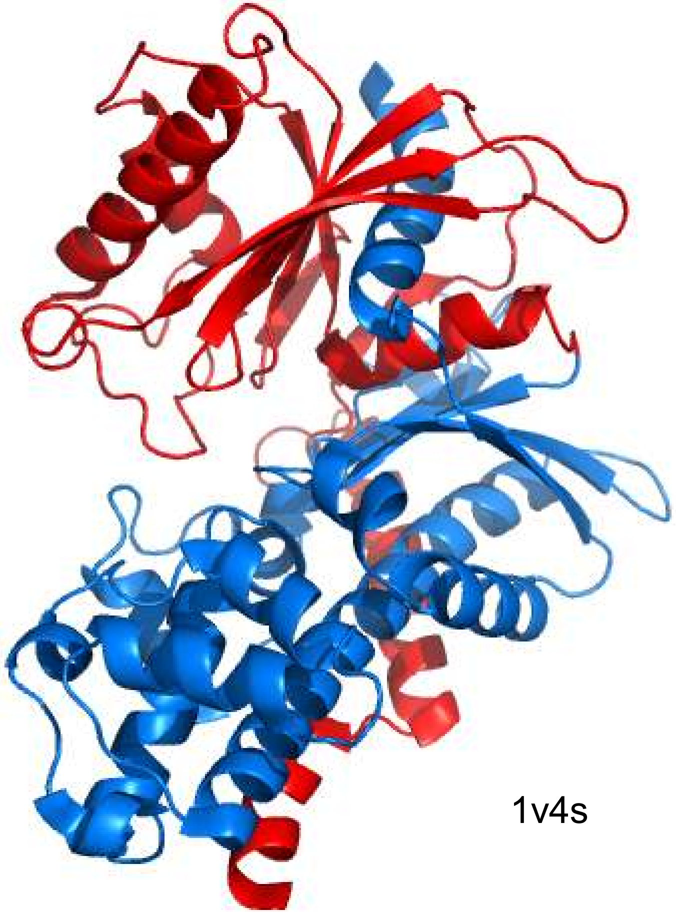
D-Galactose

alpha-D-Galactose 1-phosphate

Galactokinase
EC 2.7.1.6



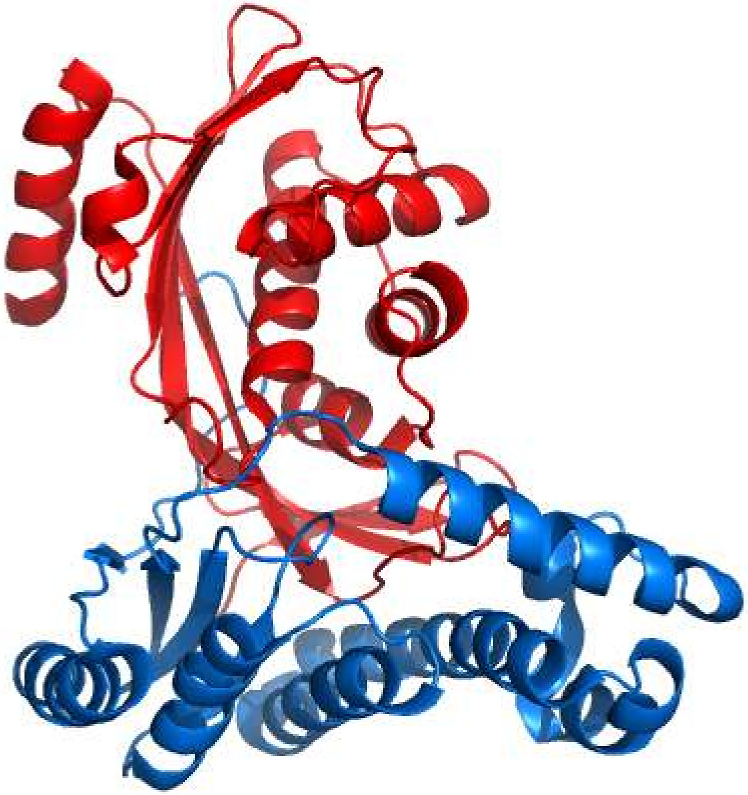
Hexokinase



1v4s

Actin-like ATPase domain

Galactokinase

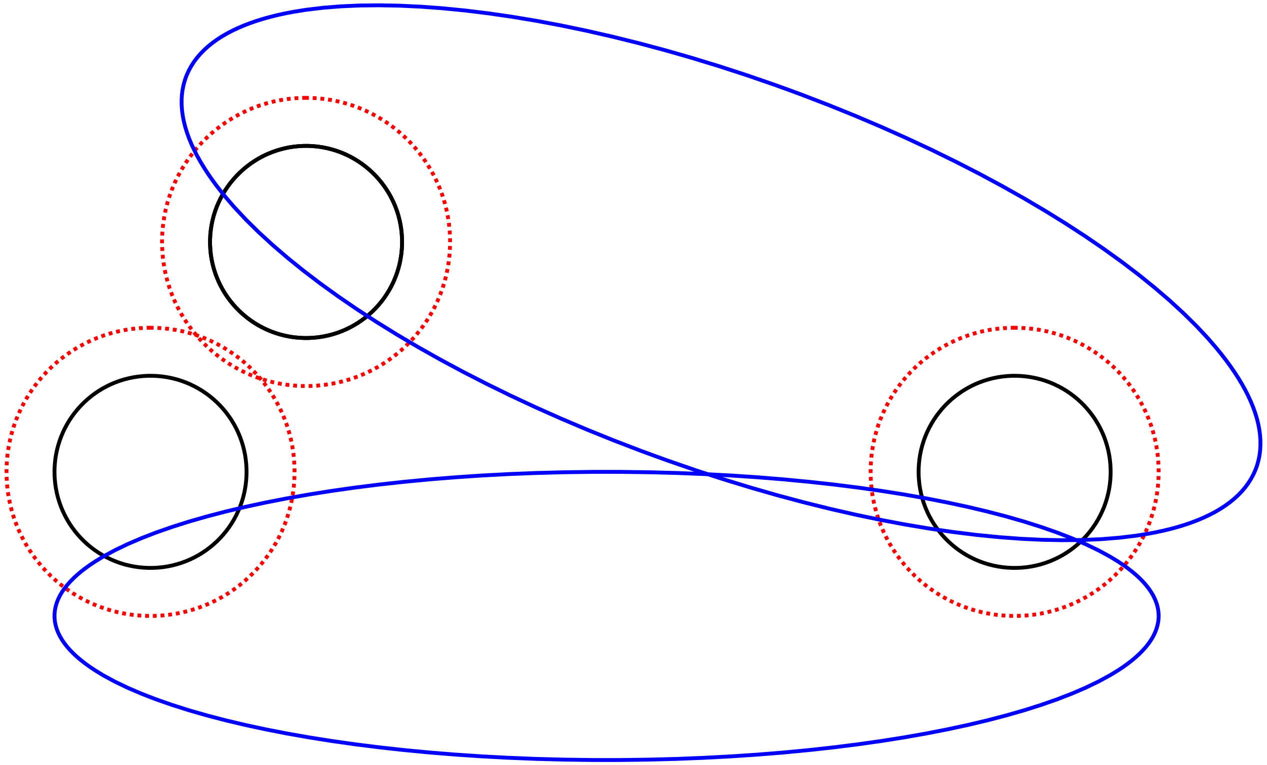


1pie

Ribosomal protein S5 domain 2-like

GHMP Kinase, C-term domain

Relationships among sequence, structure and function



Similar sequences

Similar structures

Similar functions

Adpated from A.M. Lesk. *Introduction to protein science*.

Structural resources

PDB sites

RCSB (USA)	www.pdb.org	RCSB PDB PROTEIN DATA BANK
PDBe (Europe)	www.pdbe.org	EMBL-EBI PDBe PROTEIN DATA BANK EUROPE
PDBj (Japan)	www.pdbj.org	PDBj Protein Data Bank Japan

Other PDB portal sites

PDBsum	http://www.ebi.ac.uk/pdbsum
OCA PDB	http://oca.weizmann.ac.il/oca-docs/oca-home.html
Jena Library of Biological Macromolecules	http://www.imb-jena.de/IMAGE.html
NCBI	http://www.ncbi.nlm.nih.gov/sites/entrez?db=structure

PDB contents

3D atomic coordinates

Biochemical composition

Experimental process

Additional experimental data

- Structure factors (X-ray crystallography)
- Restraints and chemical shifts (NMR)

From: http://swissmodel.expasy.org/course/text/chapte11.htm

The PDB file

Header section

```
HEADER    OXIDOREDUCTASE(NAD(A)-CHOH(D))          12-APR-89   4MDH      4MDH   3
COMPND    CYTOPLASMIC MALATE DEHYDROGENASE (E.C.1.1.1.37)               4MDH   4
SOURCE    PORCINE (SUS $SCROFA) HEART                                   4MDH   5
AUTHOR    J.J.BIRKTOFT,L.J.BANASZAK                                     4MDH   6
REVDAT   3   15-APR-92 4MDHB   3       ATOM                             4MDHB  1
REVDAT   2   15-JAN-90 4MDHA   1       JRNL                             4MDHA  1
REVDAT   1   19-APR-89 4MDH    0                                        4MDH   7
SPRSDE     19-APR-89 4MDH      2MDH                                     4MDH   8
JRNL        AUTH   J.J.BIRKTOFT,G.RHODES,L.J.BANASZAK                   4MDH   9
JRNL        TITL   REFINED CRYSTAL STRUCTURE OF CYTOPLASMIC MALATE      4MDHA  2
JRNL        TITL 2 DEHYDROGENASE AT 2.5-*ANGSTROMS RESOLUTION           4MDHA  3
JRNL        REF    BIOCHEMISTRY                  V.  28  6065 1989      4MDHA  4
JRNL        REFN   ASTM BICHAW  US ISSN 0006-2960                  033  4MDHA  5
REMARK   1                                                              4MDH  14
REMARK   1 REFERENCE 1                                                  4MDH  15
REMARK   1  AUTH   J.J.BIRKTOFT,Z.FU,G.E.CARNAHAN,G.RHODES,             4MDH  16
REMARK   1  AUTH 2 S.L.RODERICK,L.J.BANASZAK                            4MDH  17
REMARK   1  TITL   COMPARISON OF THE MOLECULAR STRUCTURES OF            4MDH  18
REMARK   1  TITL 2 CYTOPLASMIC AND MITOCHONDRIAL MALATE DEHYDROGENASE   4MDH  19
REMARK   1  REF    TO BE PUBLISHED                                      4MDH  20
REMARK   1  REFN                                                   353  4MDH  21
```

Crystallographic data

```
CRYST1  139.200   86.600   58.800  90.00  90.00  90.00 P 21 21 2     8  4MDH 328
ORIGX1      1.000000  0.000000  0.000000        0.00000                 4MDH 329
ORIGX2      0.000000  1.000000  0.000000        0.00000                 4MDH 330
ORIGX3      0.000000  0.000000  1.000000        0.00000                 4MDH 331
SCALE1      0.007184  0.000000  0.000000        0.00000                 4MDH 332
SCALE2      0.000000  0.011547  0.000000        0.00000                 4MDH 333
SCALE3      0.000000  0.000000  0.017007        0.00000                 4MDH 334
MTRIX1   1 -0.865540  0.467810 -0.178880       55.21400    1            4MDH 335
MTRIX2   1  0.499790  0.829880 -0.248020       -1.79900    1            4MDH 336
MTRIX3   1  0.032420 -0.304070 -0.952100       89.13300    1            4MDH 337

                                   (...)
```

Sequence

```
SEQRES   1 A  334  ACE SER GLU PRO ILE ARG VAL LEU VAL THR GLY ALA ALA  4MDH 163
SEQRES   2 A  334  GLY GLN ILE ALA TYR SER LEU LEU TYR SER ILE GLY ASN  4MDH 164
SEQRES   3 A  334  GLY SER VAL PHE GLY LYS ASP GLN PRO ILE ILE LEU VAL  4MDH 165

                                (...)

SEQRES  24 A  334  VAL GLU GLY LEU PRO ILE ASN ASP PHE SER ARG GLU LYS  4MDH 186
SEQRES  25 A  334  MET ASP LEU THR ALA LYS GLU LEU ALA GLU GLU LYS GLU  4MDH 187
SEQRES  26 A  334  THR ALA PHE GLU PHE LEU SER SER ALA                  4MDH 188
SEQRES   1 B  334  ACE SER GLU PRO ILE ARG VAL LEU VAL THR GLY ALA ALA  4MDH 189
SEQRES   2 B  334  GLY GLN ILE ALA TYR SER LEU LEU TYR SER ILE GLY ASN  4MDH 190
SEQRES   3 B  334  GLY SER VAL PHE GLY LYS ASP GLN PRO ILE ILE LEU VAL  4MDH 191

                                (...)

SEQRES  24 B  334  VAL GLU GLY LEU PRO ILE ASN ASP PHE SER ARG GLU LYS  4MDH 212
SEQRES  25 B  334  MET ASP LEU THR ALA LYS GLU LEU ALA GLU GLU LYS GLU  4MDH 213
SEQRES  26 B  334  THR ALA PHE GLU PHE LEU SER SER ALA                  4MDH 214

                                (...)
```

Atomic coordinates

Chain

Atom id

X Y Z coordinates

B-factor

```
ATOM      1  C    ACE A    0      11.590    2.938   35.017  1.00 45.90         4MDHB   5
ATOM      2  O    ACE A    0      12.581    2.371   35.517  1.00 28.75         4MDHB   6
ATOM      3  CH3  ACE A    0      10.179    2.477   35.417  1.00 36.75         4MDHB   7
ATOM      4  N    SER A    1      11.648    3.946   34.081  1.00 49.10         4MDH  341
ATOM      5  CA   SER A    1      12.901    4.557   33.573  1.00 52.42         4MDH  342
ATOM      6  C    SER A    1      12.733    5.624   32.482  1.00 48.48         4MDH  343
ATOM      7  O    SER A    1      13.238    5.432   31.363  1.00 57.03         4MDH  344
ATOM      8  CB   SER A    1      13.990    3.553   33.162  1.00 41.45         4MDH  345
ATOM      9  OG   SER A    1      15.105    3.679   34.039  1.00 42.59         4MDH  346
ATOM     10  N    GLU A    2      12.073    6.774   32.772  1.00 37.72         4MDH  347
ATOM     11  CA   GLU A    2      11.948    7.788   31.721  1.00 20.88         4MDH  348
ATOM     12  C    GLU A    2      12.042    9.235   32.169  1.00 28.31         4MDH  349
ATOM     13  O    GLU A    2      11.285    9.654   33.030  1.00 14.56         4MDH  350
ATOM     14  CB   GLU A    2      10.925    7.482   30.621  1.00 18.66         4MDH  351
ATOM     15  CG   GLU A    2      10.188    8.729   30.102  1.00 39.41         4MDH  352
ATOM     16  CD   GLU A    2       8.693    8.532   30.110  1.00 55.62         4MDH  353
ATOM     17  OE1  GLU A    2       7.885    9.153   29.379  1.00 55.67         4MDH  354
ATOM     18  OE2  GLU A    2       8.352    7.589   30.997  1.00 68.00         4MDH  355
```

(...)

Residue

Secondary structure elements

```
HELIX    1 1BA GLY A   13   LEU A   20   1                                4MDH 226
HELIX    2 2BA LEU A   20   GLY A   26   1                                4MDH 227
HELIX    3  CA MET A   45   ALA A   60   1                                4MDH 228
                                   (...)

SHEET    1 S1A 6 LEU A  63   THR A  70   0                                4MDH 250
SHEET    2 S1A 6 PRO A  34   ASP A  41   1                                4MDH 251
SHEET    3 S1A 6 ILE A   4   GLY A  10   1                                4MDH 252
                                   (...)

TURN     1  T1 VAL A    8  ALA A  11                                     4MDH 274
TURN     2  T2 GLY A   10  GLY A  13                                     4MDH 275
TURN     3  T3 GLY A   26  PHE A  29                                     4MDH 276
                                   (...)
```

Heteroatoms

Description

```
HET    NAD  A   1      44     NAD CO- ENZYME                              4MDH 219
HET    SUL  A   2       5     SULFATE                                     4MDH 220
HET    NAD  B   1      44     NAD CO- ENZYME                              4MDH 221
HET    SUL  B   2       5     SULFATE                                     4MDH 222
FORMUL   3  NAD    2(C21 H28 N7 O14 P2)                                   4MDH 223
FORMUL   4  SUL    2(O4 S1)                                               4MDH 224
FORMUL   5  HOH   *471(H2 O1)                                             4MDH 225
                               (...)
```

Atomic coordinates

```
HETATM 5158  AP   NAD B   1      42.641  30.361  41.284  1.00 26.73      4MDH5495
HETATM 5159  AO1  NAD B   1      43.440  31.570  40.868  1.00 20.69      4MDH5496
HETATM 5160  AO2  NAD B   1      41.161  30.484  41.376  1.00 33.73      4MDH5497
HETATM 5161  AO5* NAD B   1      43.117  29.802  42.683  1.00 20.55      4MDH5498
HETATM 5162  AC5* NAD B   1      44.483  29.615  43.002  1.00 17.23      4MDH5499
                               (...)
HETATM 5202   S   SO4 B   2      44.842  24.424  31.662  1.00 72.77      4MDH5539
HETATM 5203   O1  SO4 B   2      45.916  23.890  32.631  1.00 31.43      4MDH5540
HETATM 5204   O2  SO4 B   2      44.065  23.296  30.916  1.00 26.35      4MDH5541
HETATM 5205   O3  SO4 B   2      45.570  25.307  30.620  1.00 52.53      4MDH5542
HETATM 5206   O4  SO4 B   2      43.834  25.257  32.482  1.00 47.91      4MDH5543
HETATM 5207   O   HOH     0      15.379   1.907   3.295  1.00 58.12      4MDH5544
HETATM 5208   O   HOH     1      58.861   0.984  17.024  1.00 37.58      4MDH5545
HETATM 5209   O   HOH     2      24.384   1.184  74.398  1.00 35.92      4MDH5546
                               (...)
```

Connectivity

```
CONECT   74   69   75                                                     4MDH6015
CONECT   77   76                                                          4MDH6016
CONECT   92   90   93                                                     4MDH6017
CONECT   99   98                                                          4MDH6018
                               (...)
```

Data quality

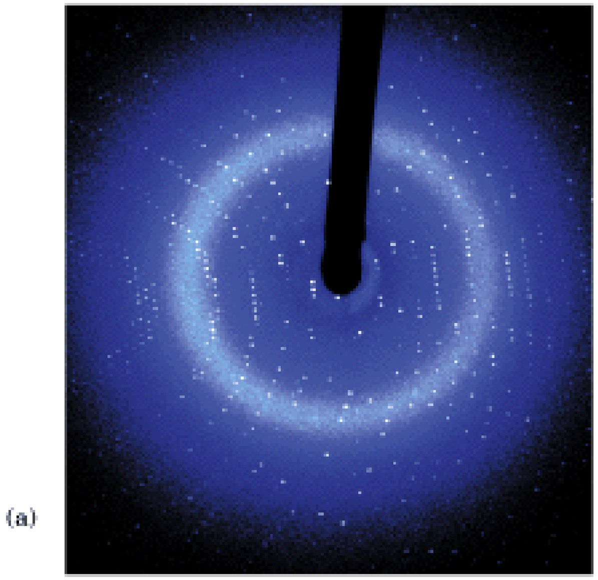
Protein structures are experimentally determined

They represent a *model* or explanation of experimental data

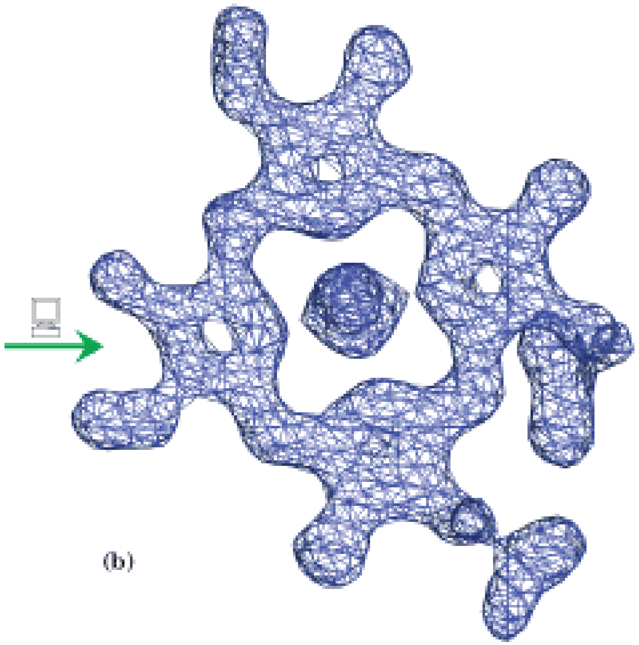
Any experiment, might have errors associated

Caution: not all structures are of equally high quality

X-ray crystallography experiment



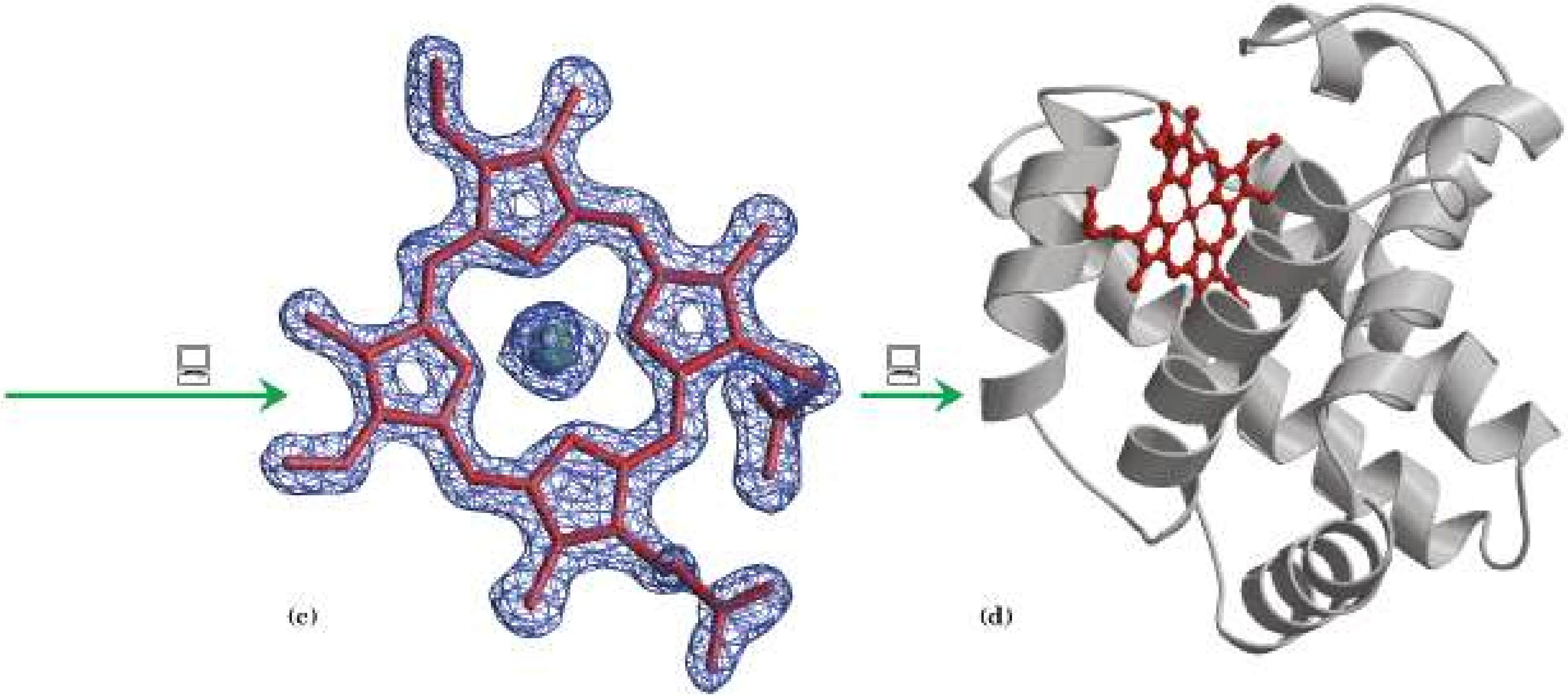
Amplitudes



Amplitudes
&
phases

Electron density maps are the primary result of crystallographic experiments

X-ray crystallography experiment



Atomic coordinates reflect an interpretation of the electron density

Validation & Structure Quality

EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

http://www.ebi.ac.uk/pdbe/resources/educationTabContent/presentations/StructureValidation.ppt

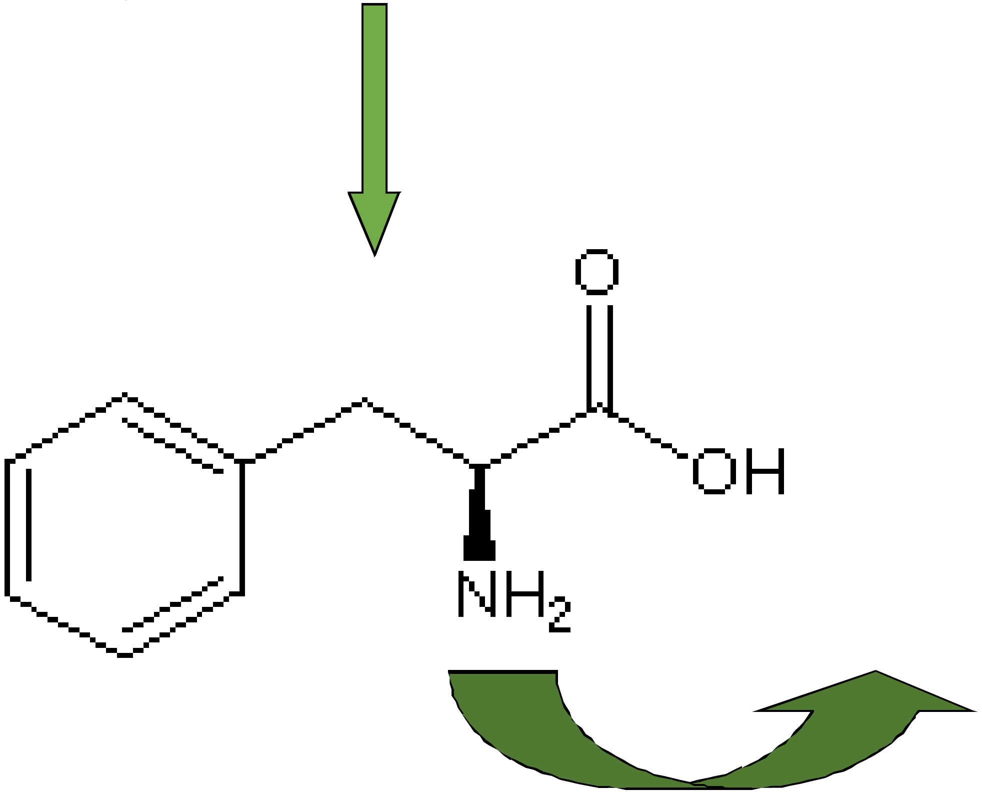
Errors in Structures

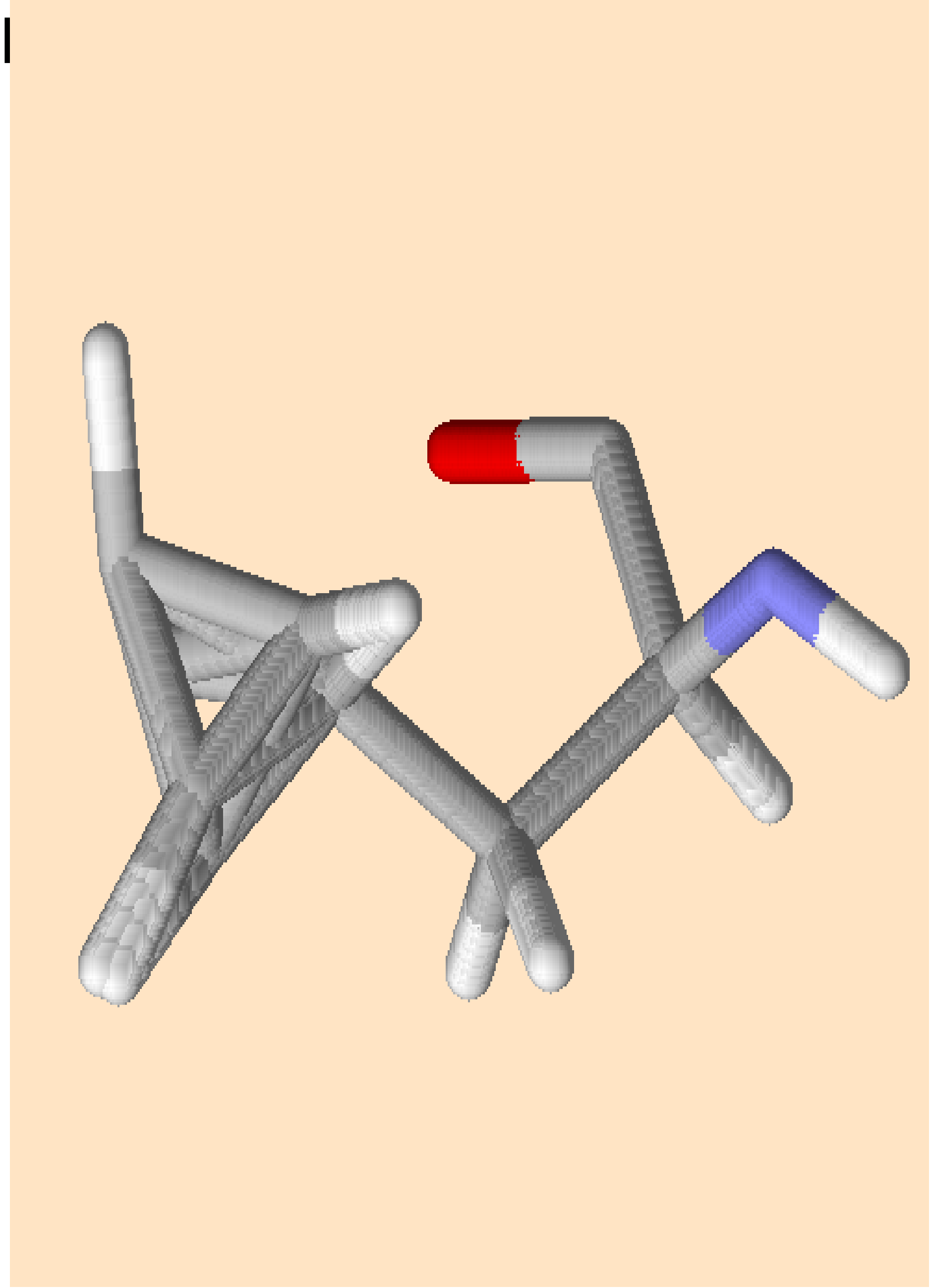
- Completely wrong
  - Wrong trace, incorrect fold of protein
  - Register errors, where trace of protein is not in keeping with sequence order.
- Partial errors
  - Incorrectly built loops.
  - Wrong residues built into the structure (i.e., Proline instead of Aspartic acid).
- Bad data quality
  - Bad geometry and stereochemistry.
  - Incorrect positioning of ligands etc due to lack of experimental evidence.

- FRAUD !!

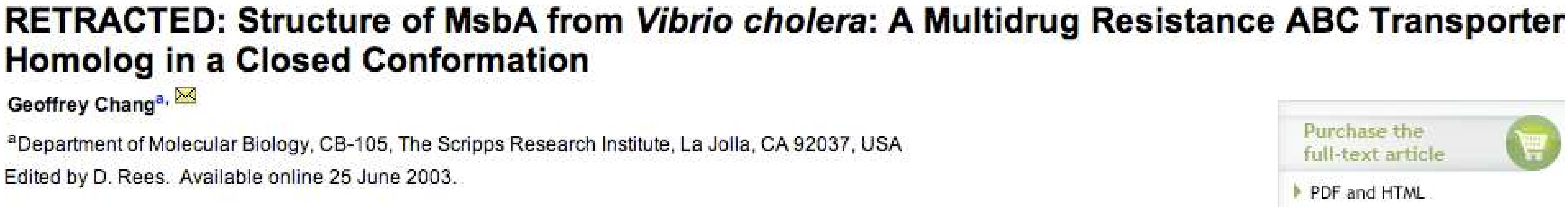
Geometry and Stereocl

This is supposed to be Phenylalanine and should look





Wrong Structures: Retracted !!



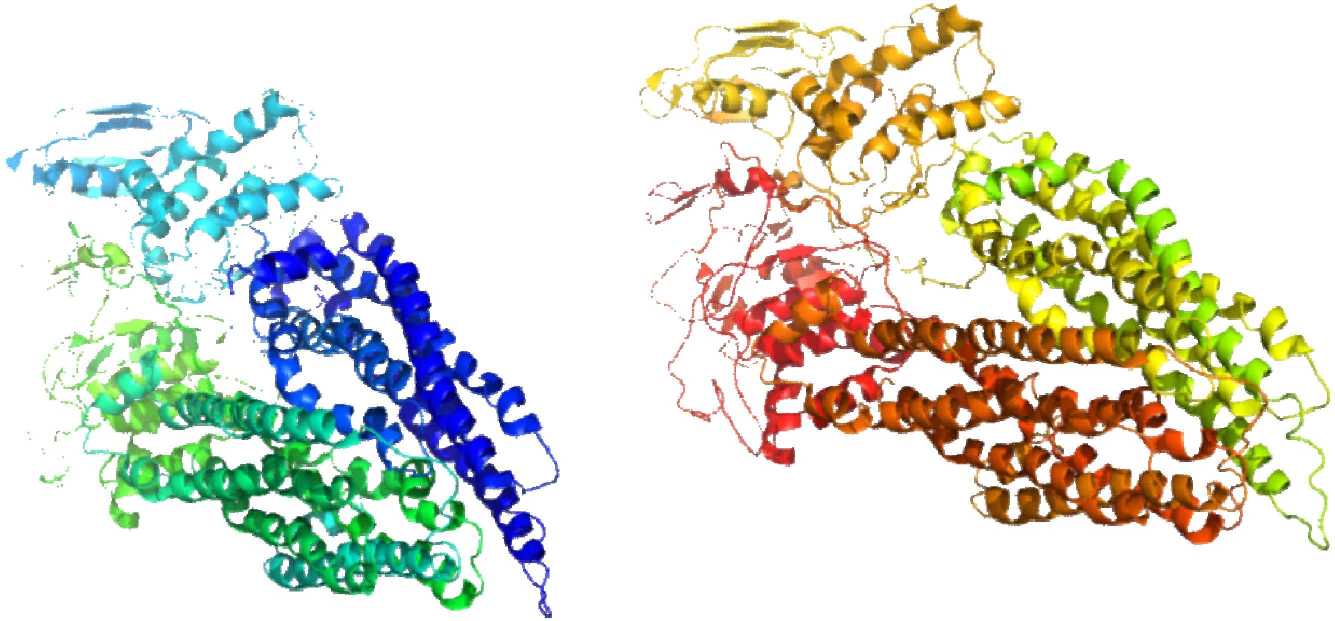
RETRACTED: Structure of MsbA from *Vibrio cholera*: A Multidrug Resistance ABC Transporter Homolog in a Closed Conformation

Geoffrey Chang[a]

[a]Department of Molecular Biology, CB-105, The Scripps Research Institute, La Jolla, CA 92037, USA

Edited by D. Rees. Available online 25 June 2003.

“were incorrect in both the hand of the structure and the topology. Thus, the biological interpretations based on the inverted models for MsbA are invalid.”



1PF4

WORLDWIDE PDB PROTEIN DATA BANK

PDBe PROTEIN DATA BANK EUROPE

Ground rules for Bioinformatics

- Don't always believe what programs tell you

  they're often misleading & sometimes wrong!
- Don't always believe what databases tell you

  they're often misleading & sometimes wrong!
- Don't always believe what lecturers tell you

  they're often misleading & sometimes wrong!
- In short, don't be a naive user
  - when computers are applied to biology, it is vital to understand the difference between mathematical & biological significance
  - computers do calculations quickly!

Computers don't do biology, You Do !

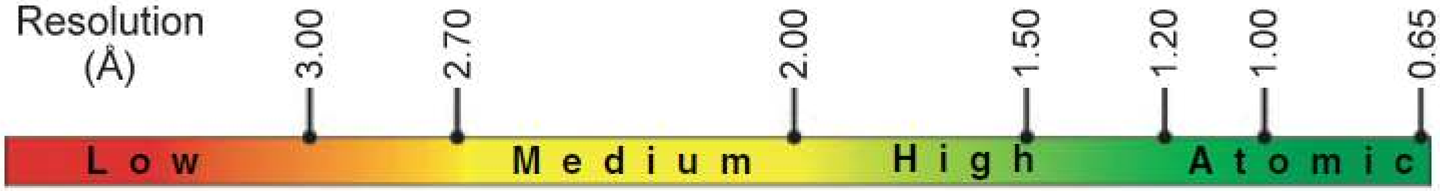
Quality indicators

Reported parameters (X-ray structures) → PDB

Stereochemical checks → Validation programs

A Wlodawer, *et al.*
Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures
FEBS Journal (2008) 275: 1–21

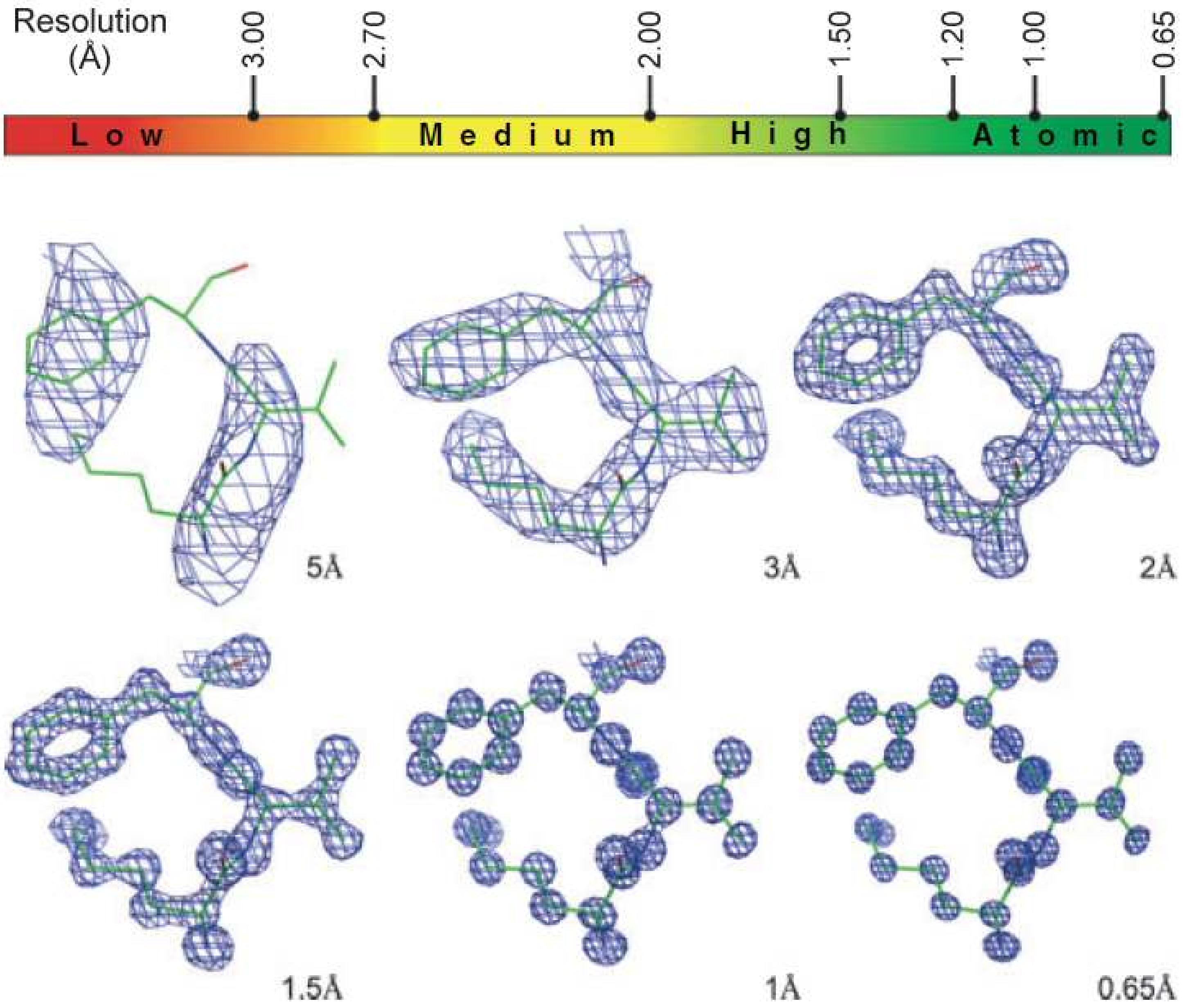
Resolution



minimum spacing of crystal lattice planes that still provide measurable diffraction of X-rays

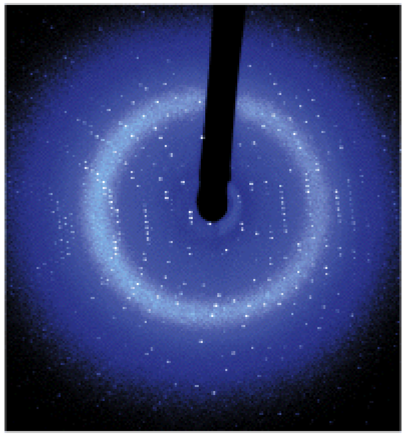
Wlodawer et al. (2008) FEBS Journal 275:1-21

Resolution

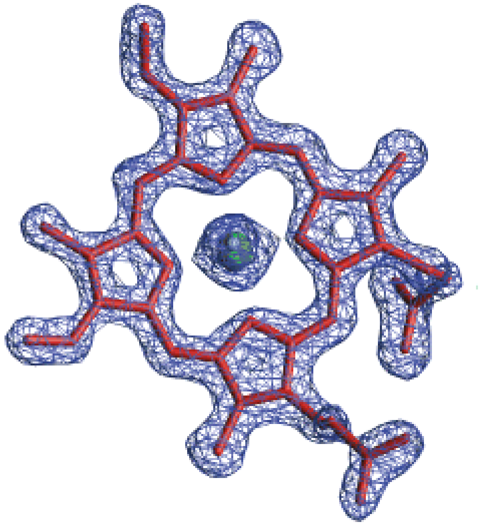


Wlodawer et al. (2008) FEBS Journal 275:1-21

R-factor and related measures



Agreement of factor amplitudes

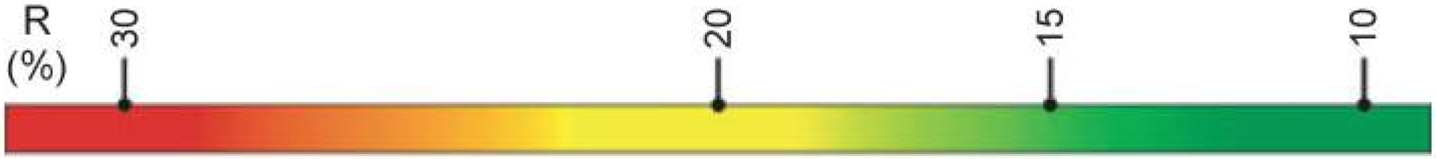


F_{obs}

F_{calc}

$$R = \Sigma|F_{obs} - F_{calc}| / \Sigma F_{obs}$$

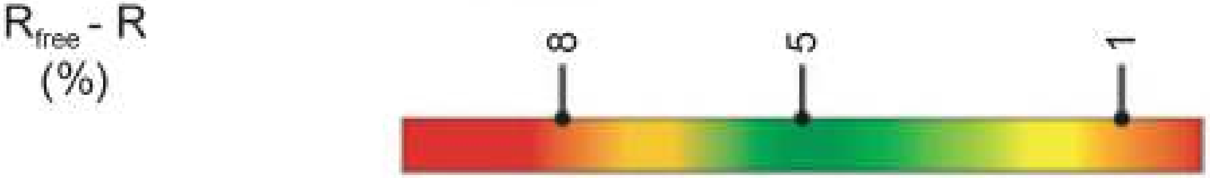
R-factor and related measures



Well-refined structures R < 20%

R_{free}

Uses only a small fraction of experimental data (5-10%)
Which is excluded from the refinement procedure



Wlodawer et al. (2008) FEBS Journal 275:1-21

Atomic B-factors

The B-factor (or temperature factor) is an indicator of thermal motion about an atom.

However, it should be pointed out that the B-factor is a mix of real thermal displacement, static disorder (multiple but defined conformations) and dynamic disorder (no defined conformation), and all the overlap between these definitions.

Expressed in $Å^2$ (2-100 range)

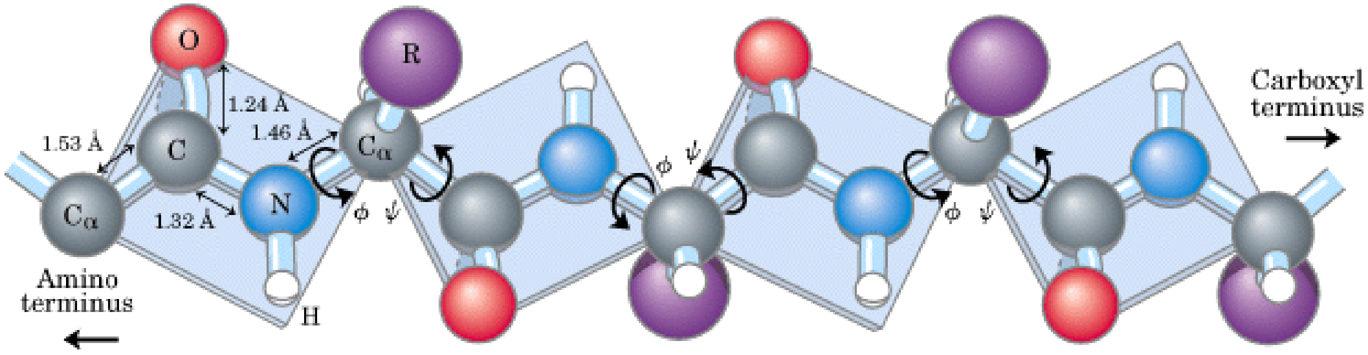
If one sees values systematically > 40 $Å^2$, the fragment may not be well defined at all.

Stereochemical checks

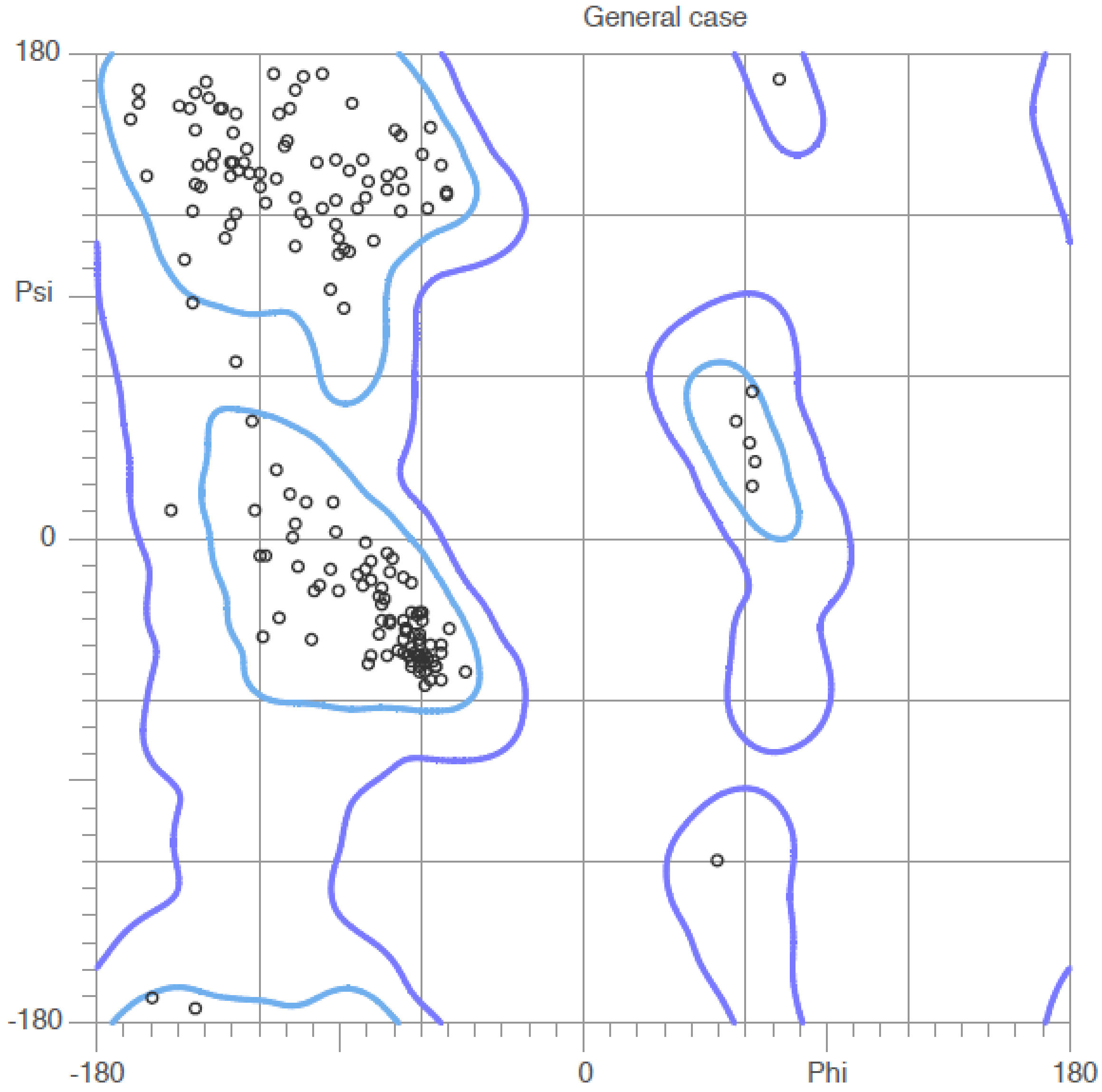
Ramachandran plot

Side-chain torsion angles

Bad contacts



MolProbity Ramachandran analysis for 2act



Validation software

Procheck
http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/

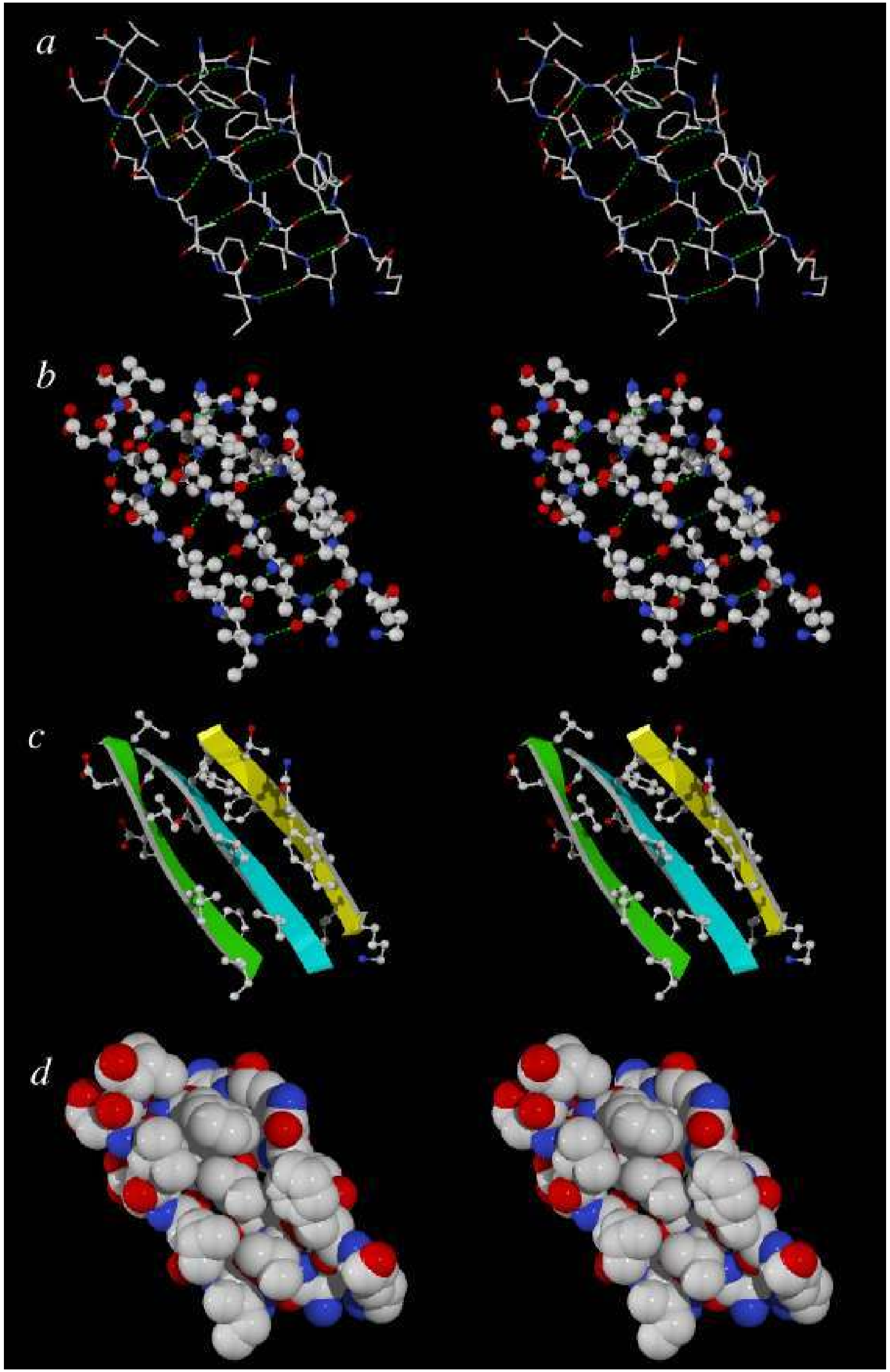
WHATCHECK
http://swift.cmbi.ru.nl/gv/whatcheck/

JCSG Validation
http://www.jcsg.org/scripts/prod/validation1.cgi

PDBeAnalysis
http://www.ebi.ac.uk/pdbe-as/pdbevalidate/

MolProbity
http://molprobity.biochem.duke.edu/

Visualization


a
b
c
d

Molecular visualization software

DeepView - Swiss
http://spdbv.vital-it.ch/

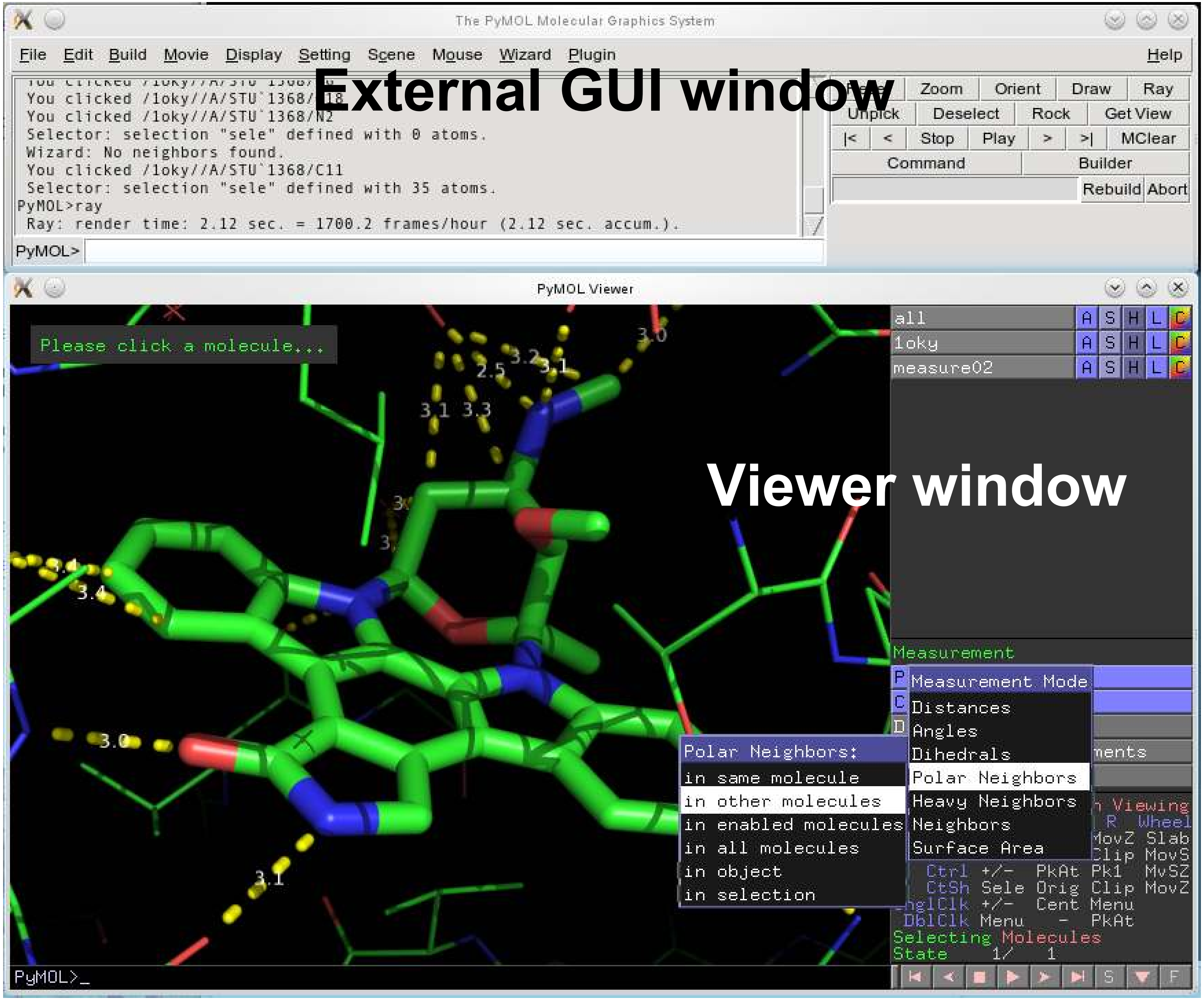
UCSF Chimera
http://plato.cgl.ucsf.edu/chimera/

Jmol
http://jmol.sourceforge.net/

Rasmol
http://www.rasmol.org/

Pymol
http://www.pymol.org/
Source code
http://sourceforge.net/projects/pymol/

VMD
http://www.ks.uiuc.edu/Research/vmd/



External GUI window

Viewer window

```
 You clicked /1oky//A/STU`1368/
 You clicked /1oky//A/STU`1368/N2
 Selector: selection "sele" defined with 0 atoms.
 Wizard: No neighbors found.
 You clicked /1oky//A/STU`1368/C11
 Selector: selection "sele" defined with 35 atoms.
PyMOL>ray
 Ray: render time: 2.12 sec. = 1700.2 frames/hour (2.12 sec. accum.).
PyMOL>
```

```
Please click a molecule...
```

```
PyMOL>_
```

To learn more:
PyMOL user manual

http://pymol.sourceforge.net/userman.pdf http://csbg.cnb.csic.es/Courses/Struct_2011/