Bioinformatique M2: Lecture 3

P. Derreumaux

III. From protein structure to function

1. What is function ?

→ Biochemical or molecular function refers to the particular catalytic function (activity), binding properties or conformational change

→ At higher level, the metabolic or signal transduction pathway, in which a protein participates the cellular function. This is always context-dependent with respect to tissue, organ and taxon

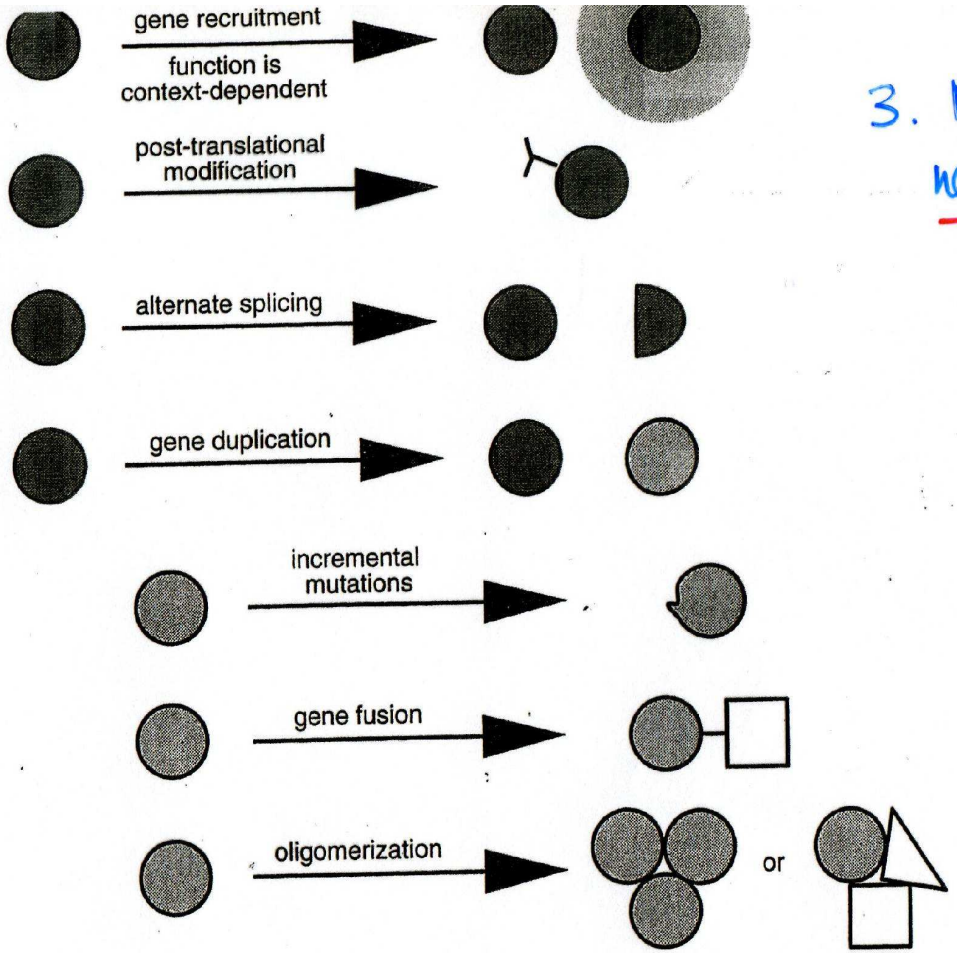
Here 'function' largely refers to biochemical aspects.

2. Challenges of inferring function from structure

Table 10.1 One structure, many functions and one function, many structures paradox.

One structure: many functions	One function: many structures
Four-wheeled van	**Telling a story**
Ambulance	Book
Ice-cream stall	CD
α/β hydrolase fold[a]	**Glycosyl hydrolase**
Triacylglycerol lipase	α/α toroid
Cholesterol esterase	Concanavalin A-like 2-layer β sandwich
Dienelactone hydrolase	Double psi β-barrel
Haloalkane dehalogenase	6-bladed β-propeller
Serine carboxypeptidase	$(\beta\alpha)_8$ or TIM barrel
Non-heme chloroperoxidase	Cellulase-like β/α-barrel
Neurotactin (cell-cell adhesion)	Orthogonal α-bundle

[a]Despite their differences in function, enzymes of the α/β hydrolase fold nevertheless have similar catalytic triads in their active-sites.

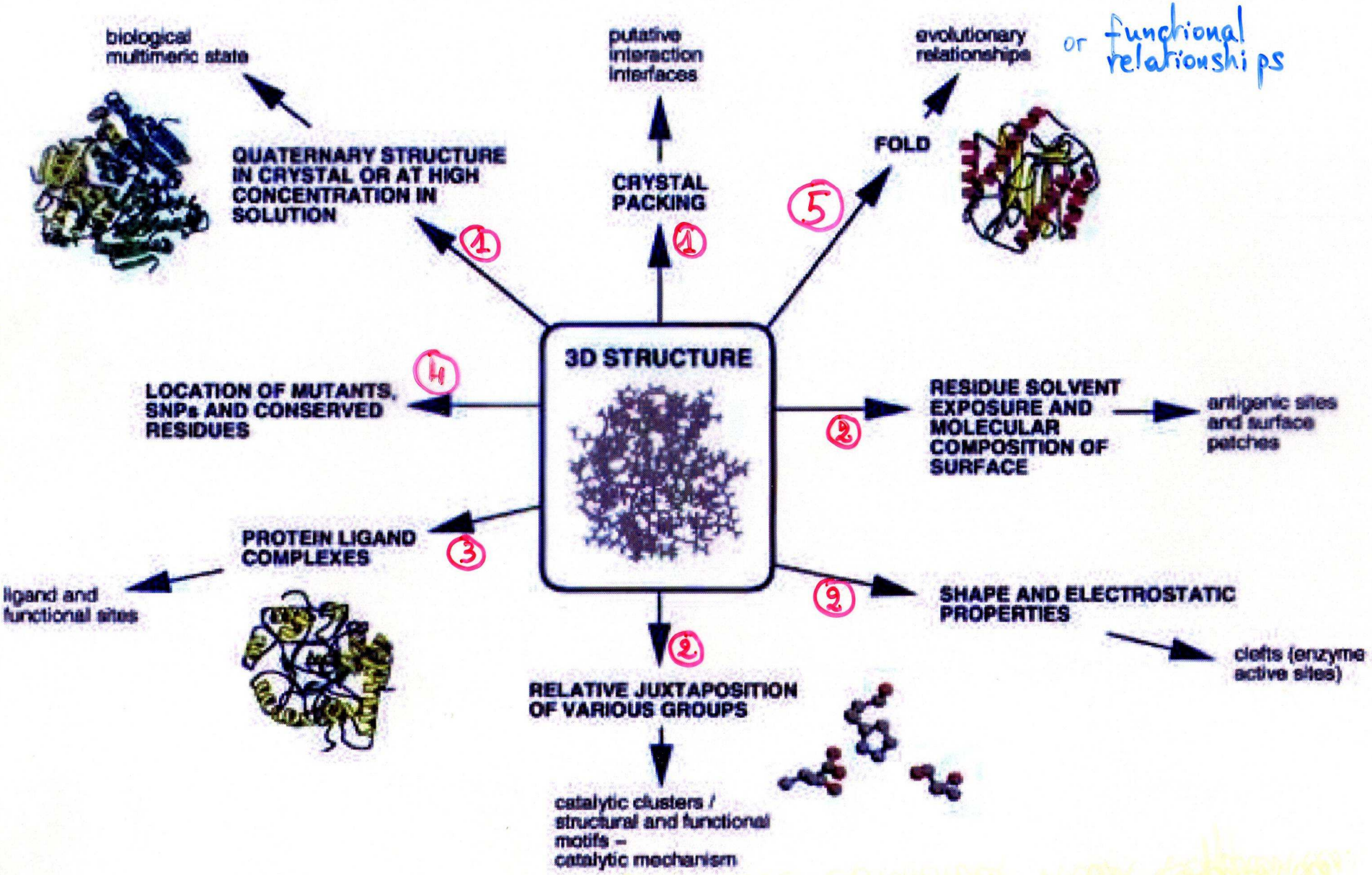


3. Mechanisms by which new functions are created.

Figure 10.1

The mechanisms by which new functions are created. In practice, new functions often evolve by a combination of mechanisms. In particular, gene duplication provides two identical copies of the same gene and one, free of functional constraints, can assume a new biological role through incremental mutations, gene fusion or oligomerization. Note however, that a change in oligomerization state without prior duplication or mutation events can provide a route to an alternative gene function and this represents a method by which proteins 'moonlight'.

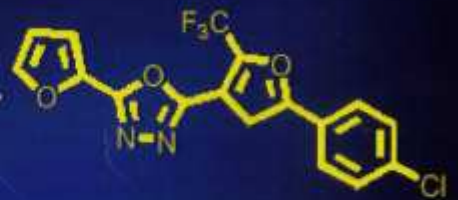
Structure to Function



Target-structure based methods

Protein-Based Design

?



Docking (rigid, flexible)

2. de novo Building

Scoring

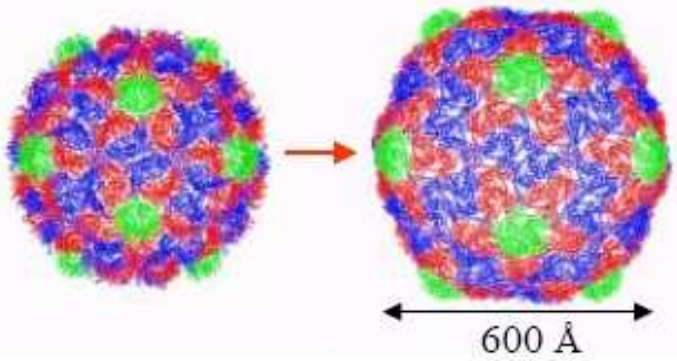
IC_{50}, K_i ???

Large-scale conformational changes

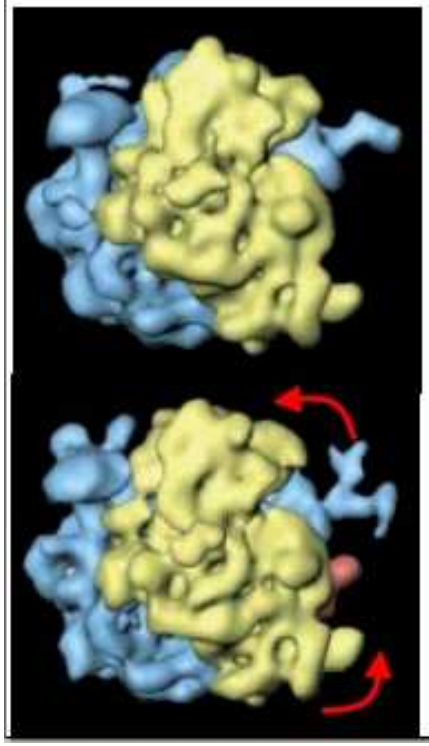
X-ray crystallography ⇨ atomic level

Cryo electron microscopy (cryo-EM) ⇨ low-resolution (from 7Å), shape information

Virus maturation

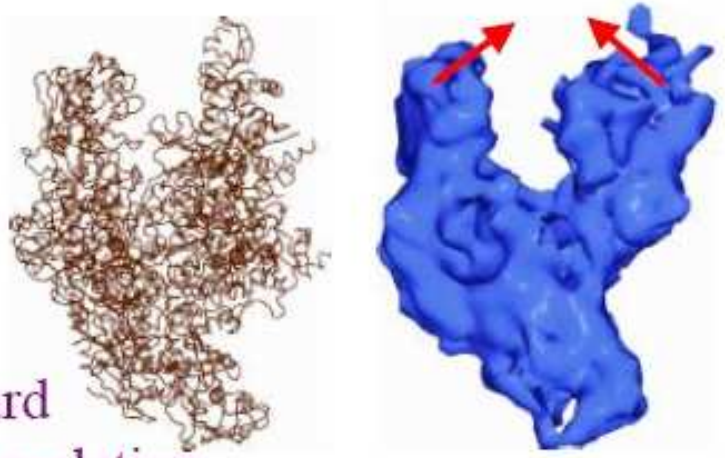


Protein synthesis: ribosome



Functional motions

Transcription - Replication



Time scale (> μs) not accessible from standard molecular dynamics simulation

RNA polymerase

Repositories for data on 3D structures of Biological Macromolecules

PDB: Protein Data Bank : 3D structures of biological macromolecules
[http://www.rcsb.org/pdb/]

MMDB: Entrez (NCBI) structure database (no models)
[http://www.ncbi.nlm.nih.gov:80/Structure/MMDB/mmdb.shtml]

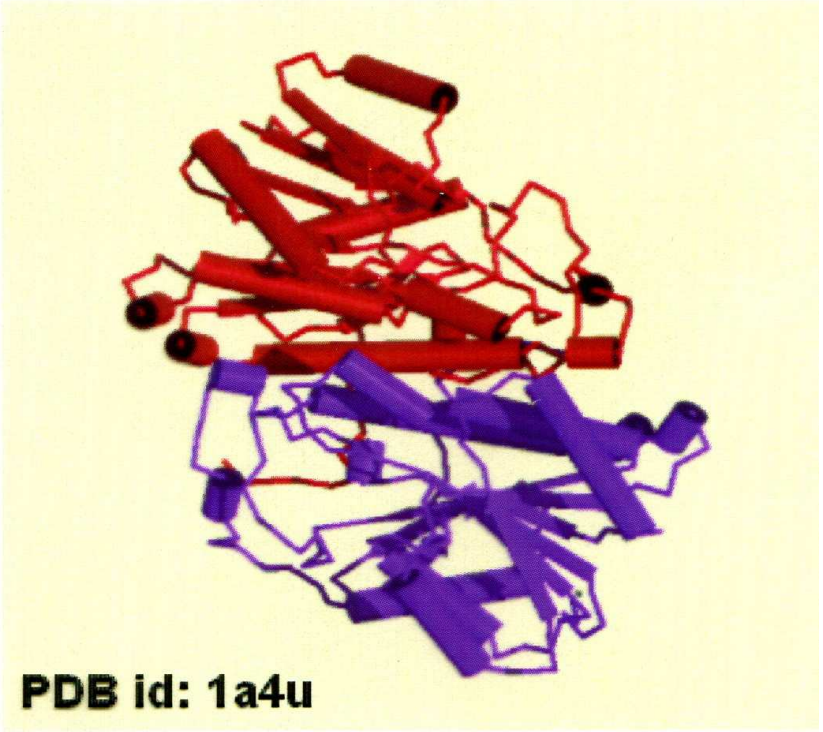
BioMagResBank: data on 3D structures determined by NMR
[http://www.bmrb.wisc.edu/]

CSD: Cambridge small molecule database
[http://www.ccdc.cam.ac.uk/]

http://www.expasy.ch/alinks.html#Proteins

Structure of ADH

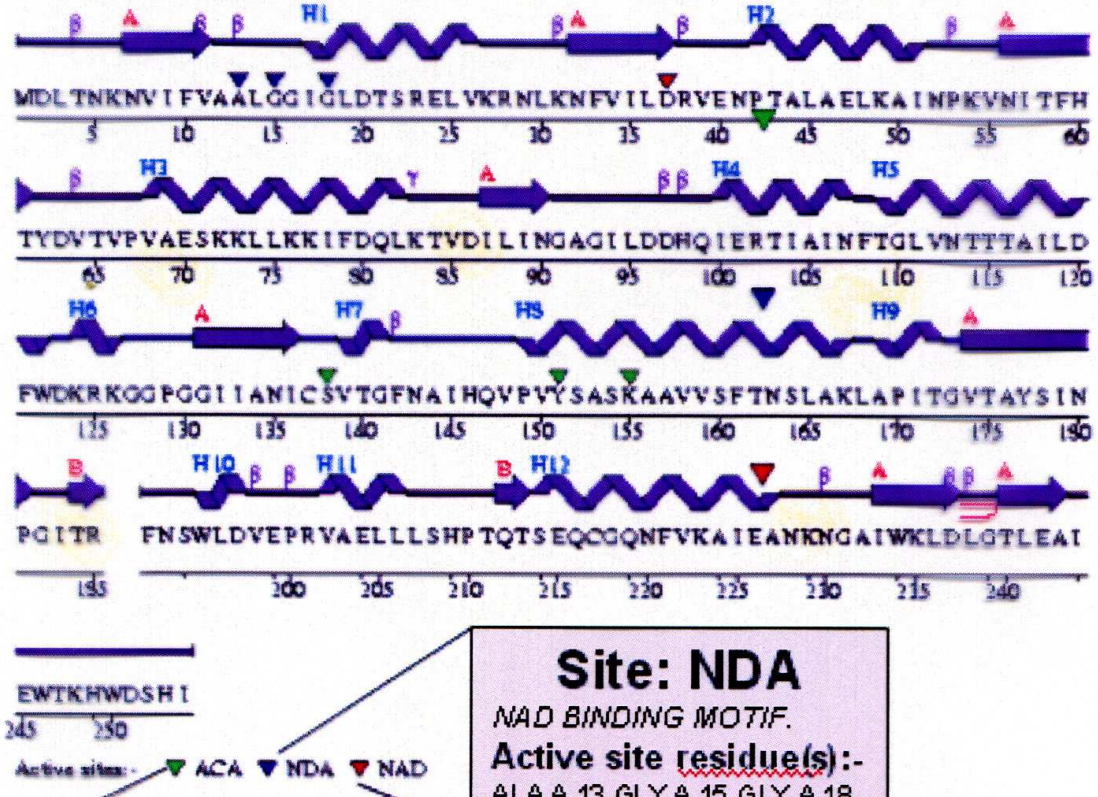
PDB id: 1a4u



PDB id: 1a4u

SOURCE: http://www.biochem.ucl.ac.uk/bsm/pdbsum/1a4u/main.html

Two-dimensional structure of ADH



Site: NDA

NAD BINDING MOTIF.

Active site residue(s):-

ALA A 13 GLY A 15 GLY A 18

1a4u: Chain A - Active sites

Site: ACA

ACTIVE SITE, CATALYTIC TRIAD.

Active site residue(s):-

SER A 138 TYR A 151 LYS A 155

Site: NAD

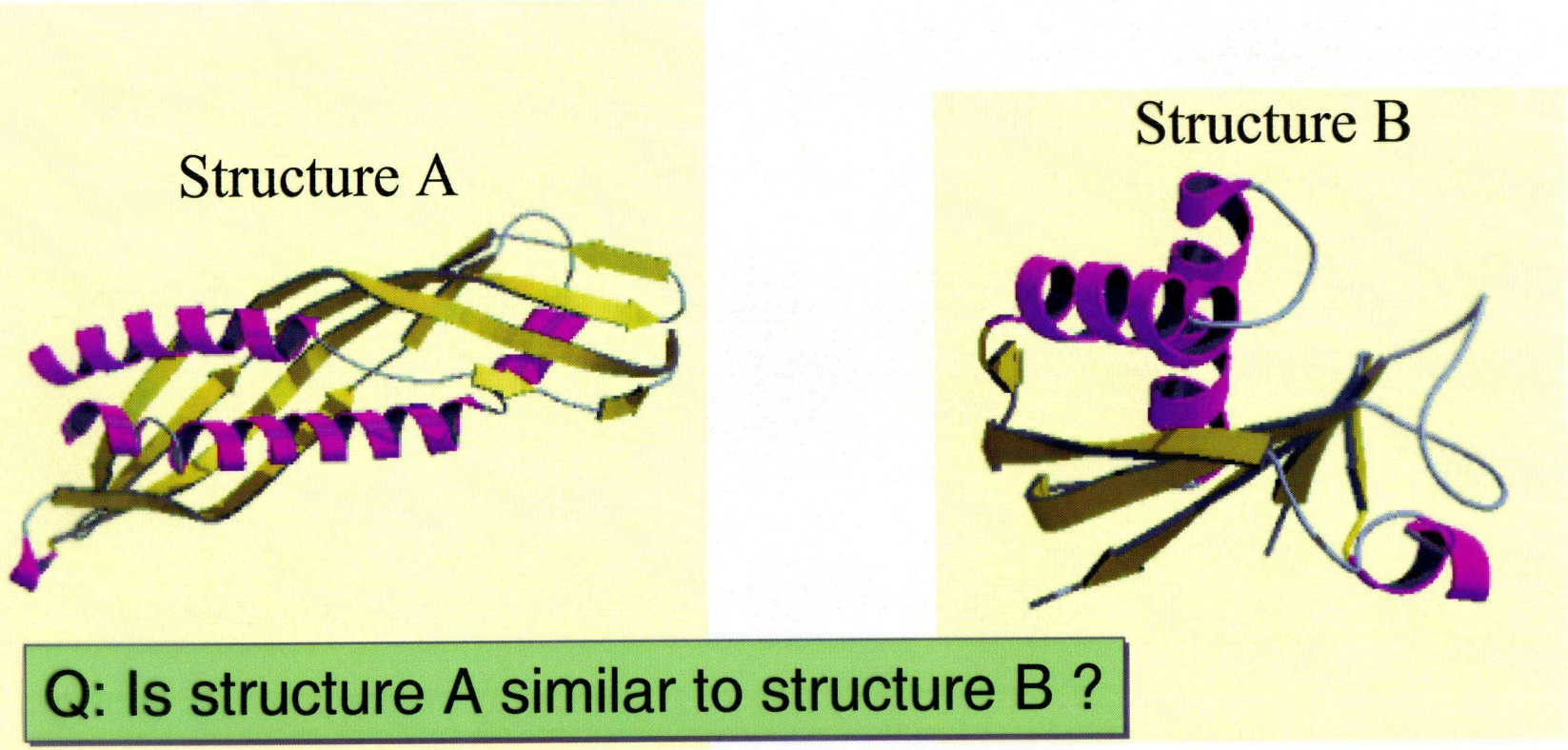
NAD/NADP SELECTIVITY AMINO ACID. SITE

Active site residue(s):-

ASP A 37

Because structures are more conserved than sequences.

Structure comparisons and structure-structure alignments



Q: Is structure A similar to structure B ?

A: from structure alignments

Structural Alignment Mathematics

- Find rotation matrix R and translation vector T for which:

$$B = R\ A + T$$

- No known deterministic algorithm
- NP hard!
- Existence of gaps and insertions
- Measure used to quantify difference (rmsd → E-value)

Similarity Measure

Root Mean Square Deviation (RMSD)

$$R_{ms} = \sqrt{\sum_{i=1}^{n} \frac{(C_{Ai} - C_{Bi})^2}{n}}$$

Distance between aligned atoms

The number of aligned atoms

- Penalizes worst fitting atoms
- Contributions of individual atoms not discernable

Classifications structurales et évolutive des Protéines

- Les 2 principales classifications :
  - "Structural Classification of Proteins " (SCOP)
    http://scop.mrc-lmb.cam.ac.uk/scop/
  - CATH (Class Architecture Topology Homology)
    http://www.biochem.ucl.ac.uk/bsm/cath/
- Autres classifications :
  - FSSP
    http://www2.ebi.ac.uk/dali/fssp/
  - VAST
    http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html
  - 3 Dee
    http://circinus.ebi.ac.uk:8080/3Dee/help/help_intro.html
  - HOMSTRAD
    http://www-cryst.bioc.cam.ac.uk/~homstrad/

THE MAJOR LEVELS IN SCOP HIERARCHY

1. <u>FAMILY</u>: clear evolutionarily relationship.

Two criteria: all proteins with sequence identities > 30%; proteins with lower sequence identities, but whose functions and structures are very similar, e.g. globins with sequence identities of 15%

THE MAJOR LEVELS IN SCOP HIERARCHY

2. <u>SUPERFAMILY</u>: probable common evolutionarily origin.

proteins with low sequence identities, but whose functions and structures suggest a common origin, e.g. actin, the ATPase domain of the heat-shock protein and hexokinase form a superfamily.

THE MAJOR LEVELS IN SCOP HIERARCHY

3. <u>FOLD</u>: major structural similarity.

Proteins have same major secondary structures in same arrangement and with the same topological connections. Fold similarity between two proteins probably arises from physics and chemistry laws, but a common evolutionary origin cannot be totally excluded.

THE MAJOR LEVELS IN SCOP HIERARCHY

4. CLASS: secondary structure identification.

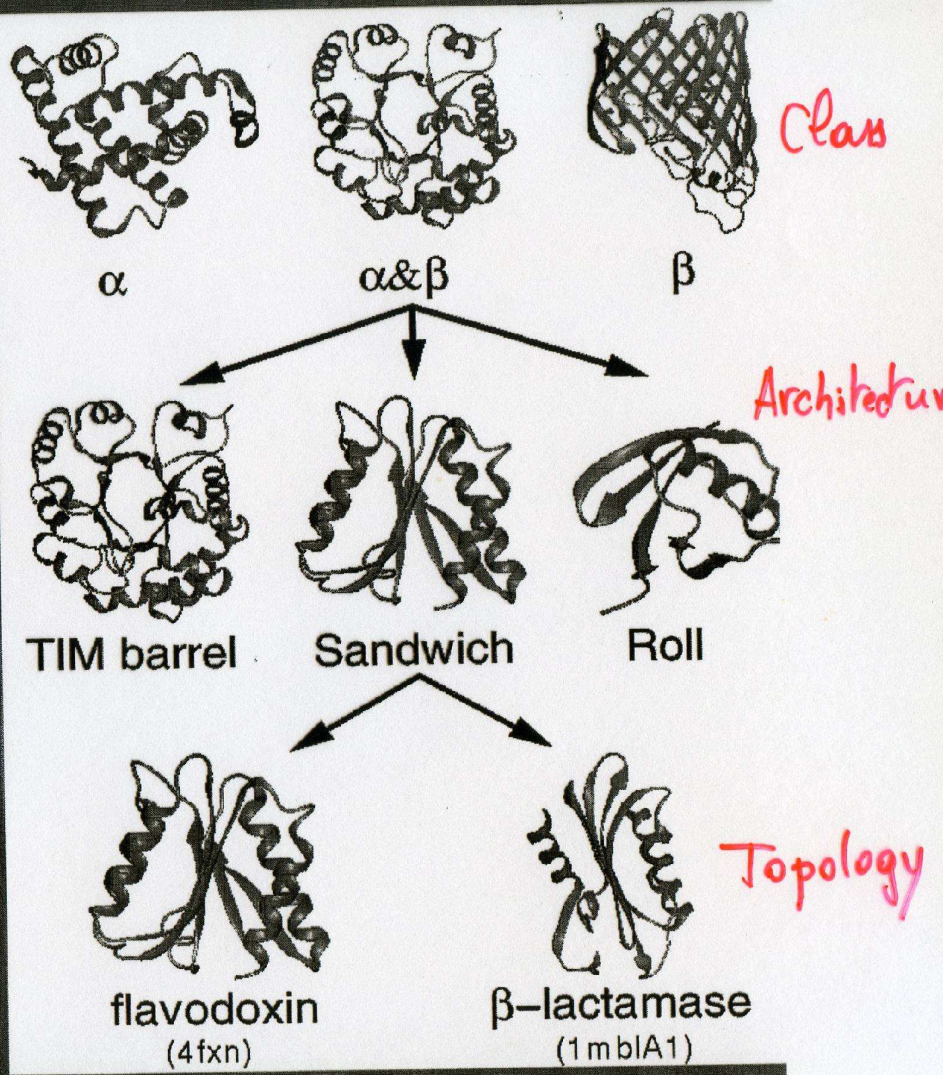
all alpha, all beta, alpha/beta, alpha+beta, multi-domain,

membrane and cell surface, and small proteins.

Classification of Protein Structures

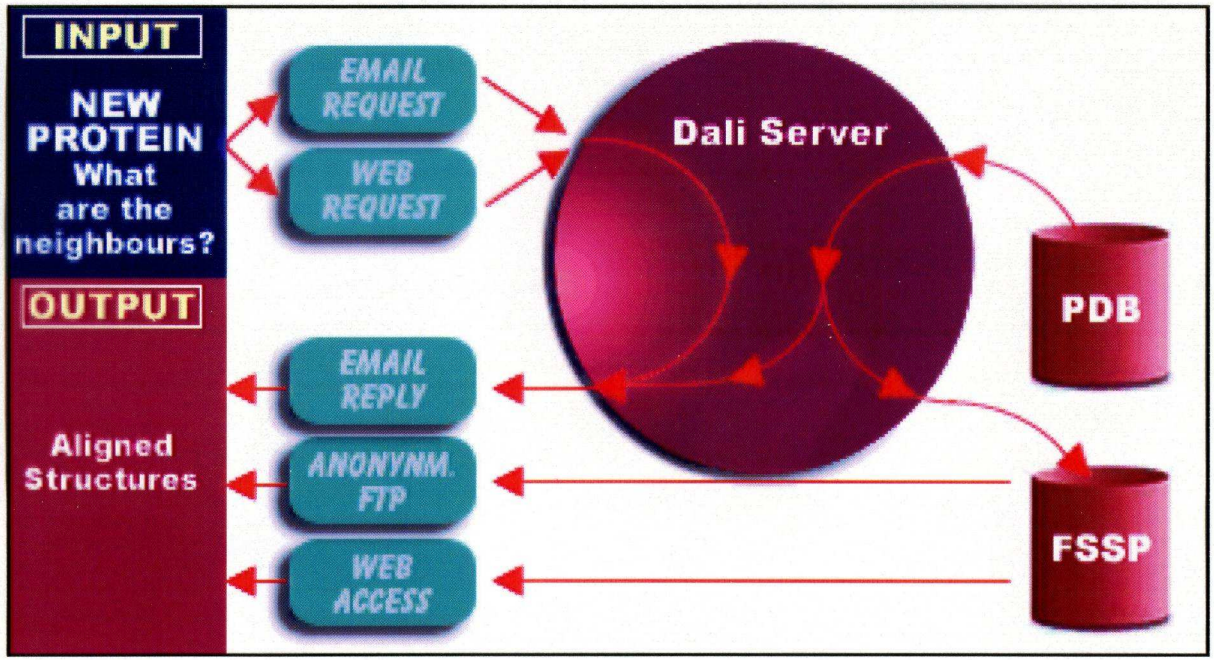
CATH

- Class
  - Similar secondary structure content
  - All α, all β, α/β, etc
- Fold (Architecture)
  - Major structural similarity
  - SSE's in similar arrangement
- Superfamily (Topology)
  - Probable common ancestry
- Family
  - Clear evolutionary relationship
  - Sequence similarity usually > 25%



FSSP

- **F**old classification based on **S**tructure-Structure alignment of **P**roteins (FSSP)
- Uses DALI (**D**istance **Ali**gnment tool)



Proc Natl Acad Sci U S A. 2009 Oct 13;106(41):17377-82. Epub 2009 Sep 24.

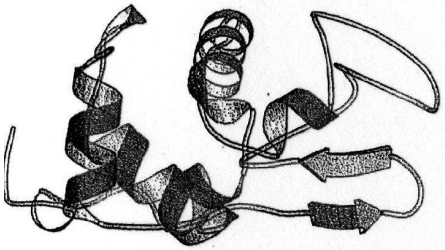
Structural relationships among proteins with different global topologies and their implications for function annotation strategies.

Petrey D, Fischer M, Honig B.

Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Columbia University, 1130 St. Nicholas Avenue, Room 815, New York, NY 10032, USA.

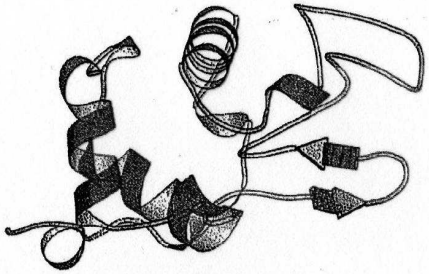
It has become increasingly apparent that geometric relationships often exist between regions of two proteins that have quite different global topologies or folds. In this article, we examine whether such relationships can be used to infer a functional connection between the two proteins in question. We find, by considering a number of examples involving metal and cation binding, sugar binding, and aromatic group binding, that geometrically similar protein fragments can share related functions, even if they have been classified as belonging to different folds and topologies. Thus, the use of classifications inevitably limits the number of functional inferences that can be obtained from the comparative analysis of protein structures. In contrast, the development of interactive computational tools that recognize the "continuous" nature of protein structure/function space, by increasing the number of potentially meaningful relationships that are considered, may offer a dramatic enhancement in the ability to extract information from protein structure databases. We introduce the MarkUs server, that embodies this strategy and that is designed for a user interested in developing and validating specific functional hypotheses.

5. Complexities of protein structure, sequence and function.

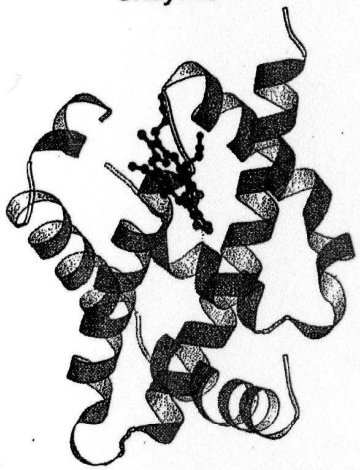


human lysozyme EC 3.2.1.17
enzyme

(b)
ENZYME/NON-ENZYME
40% seq ID
disruption of active-site



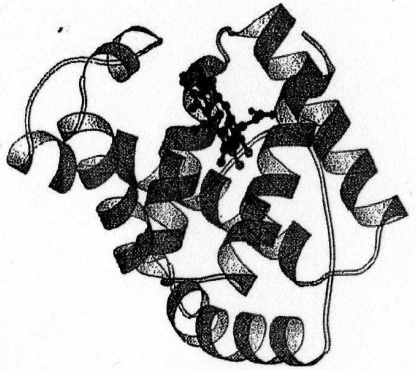
human α-lactalbumin
non-enzyme



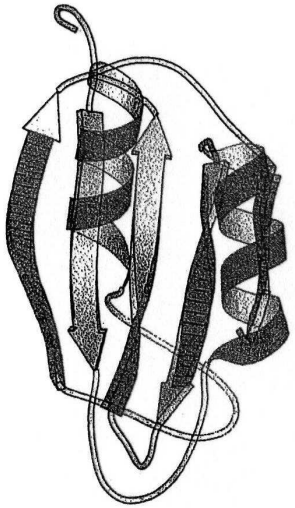
P. marinus hemoglobin

homologous relationships

(c)
IDENTICAL FUNCTIONS
8% seq ID



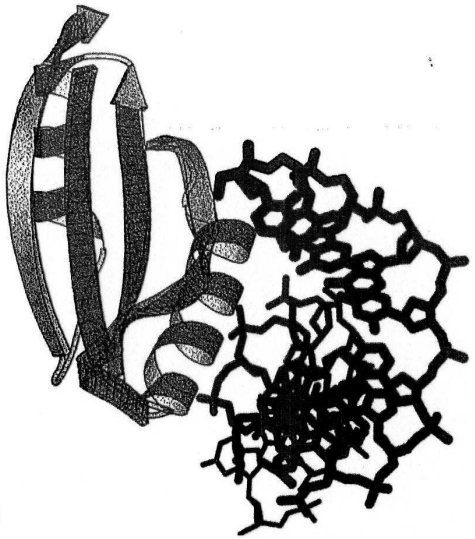
V. stercoraria hemoglobin



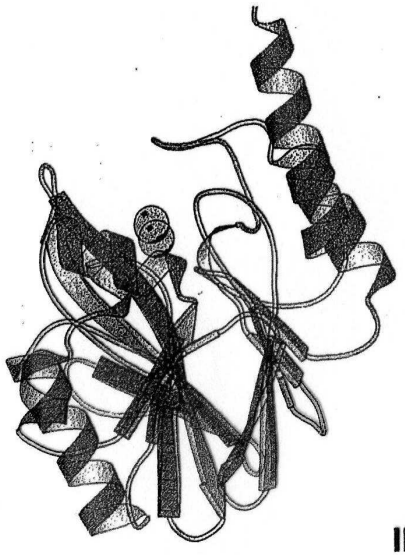
acylphosphatase
EC 3.6.1.7

(e)

SIMILAR FOLDS
DIFFERENT FUNCTIONS
no shared functional attributes



bovine papillomavirus-1 E2 transcription regulation protein, DNA-binding domain

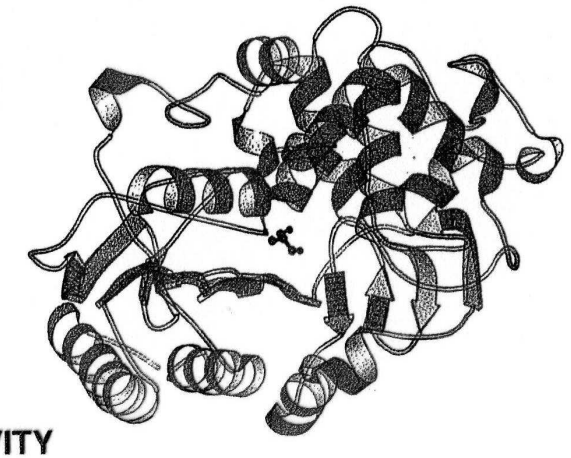


β-lactamase class B
EC 3.5.2.6
metal-dependent

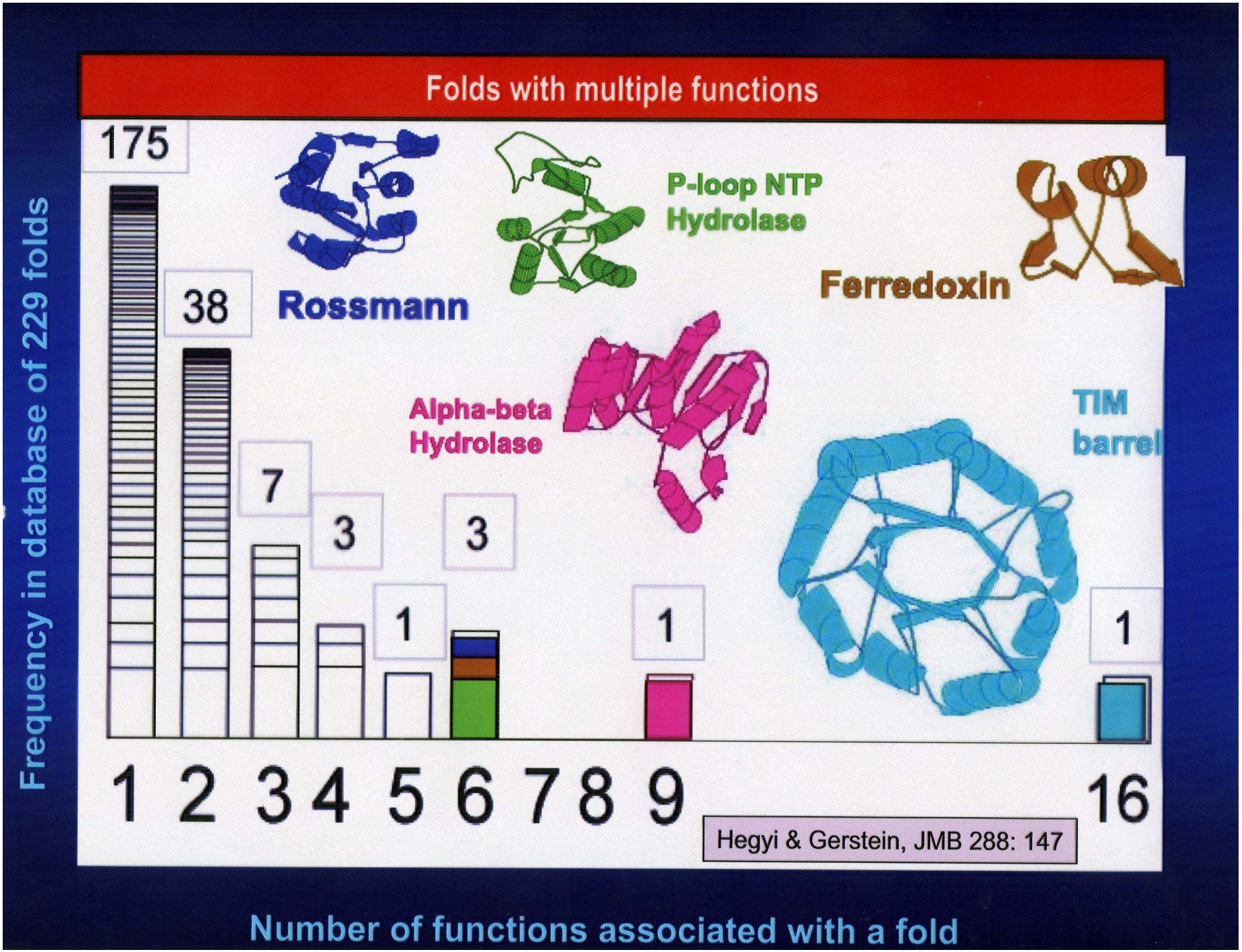
analogous relationships

(f)

DIFFERENT FOLDS
IDENTICAL ENZYME ACTIVITY
different active-sites



β-lactamase classes A, C, D
EC 3.5.2.6
catalytic Ser nucleophile


Folds with multiple functions
Frequency in database of 229 folds
175
38
7
3
3
1
1
1
Rossmann
P-loop NTP Hydrolase
Ferredoxin
Alpha-beta Hydrolase
TIM barrel
1 2 3 4 5 6 7 8 9 16
Hegyi & Gerstein, JMB 288: 147
Number of functions associated with a fold

. Enzyme Structure Database

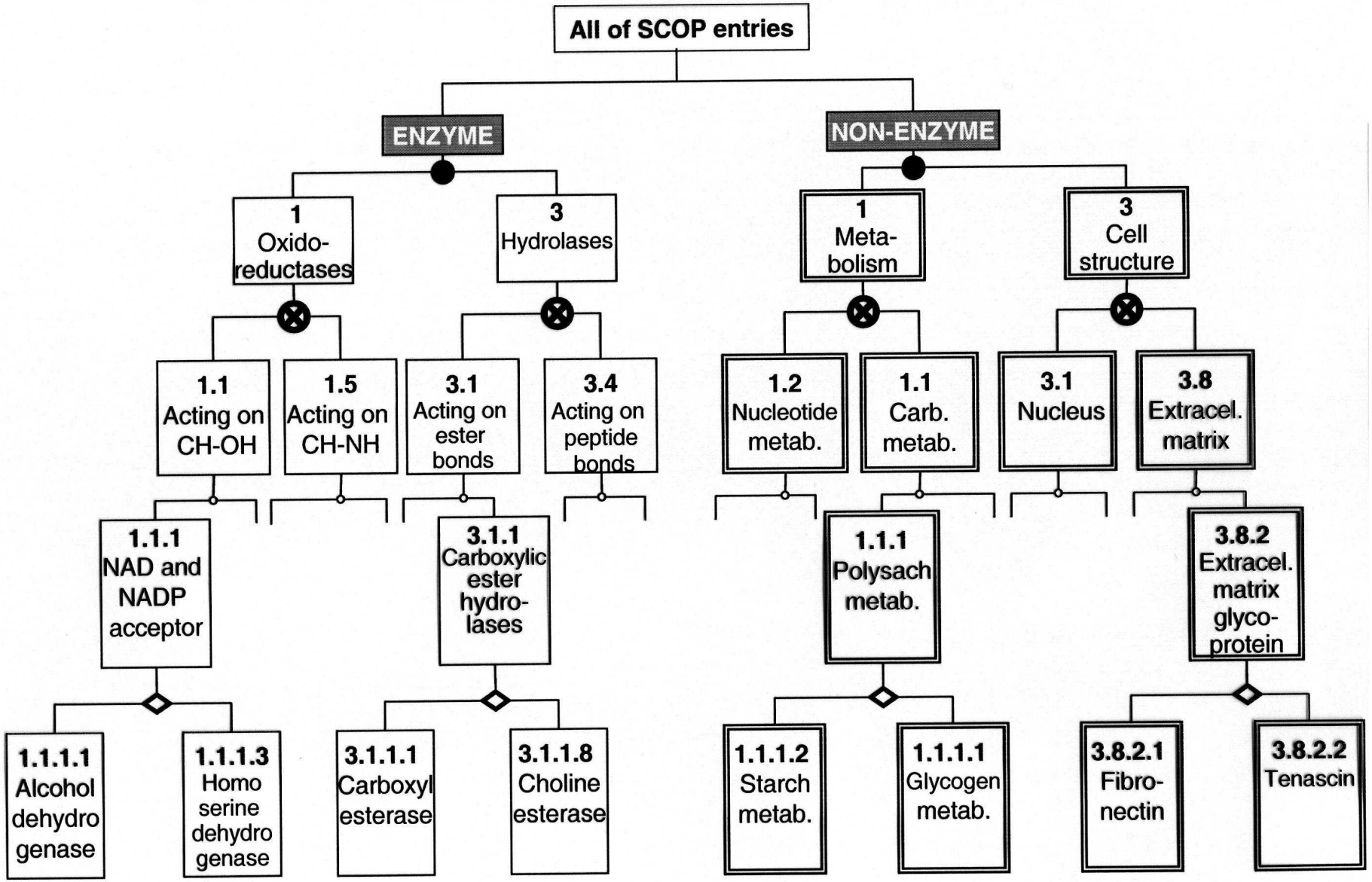
Protein Data Bank (the **PDB**).
PDB-enzyme entries: 10208 (as at 28 March 2003)
Separate PDB-files: 9873
Some files having more than one **E.C.** number associated with them.

Structure retrieval

* by E.C. numbers (v.30.0 release of the **ENZYME** Data Bank)

- **E.C.1.-.-.-** *Oxidoreductases*. [1733 PDB entries]
- **E.C.2.-.-.-** *Transferases*. [2261 PDB entries]
- **E.C.3.-.-.-** *Hydrolases*. [4637 PDB entries]
- **E.C.4.-.-.-** *Lyases*. [796 PDB entries]
- **E.C.5.-.-.-** *Isomerases*. [497 PDB entries]
- **E.C.6.-.-.-** *Ligases*. [284 PDB entries]

Hierarchy of Protein Functions

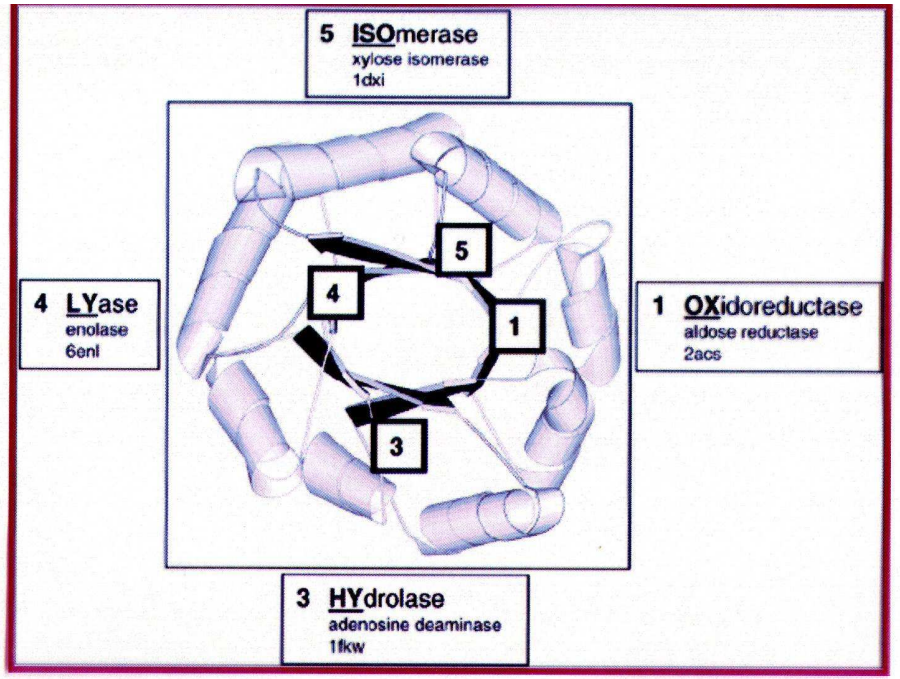


◇ **Precise functional similarity** ● **General similarity** ⊕ **Functional class similarity**

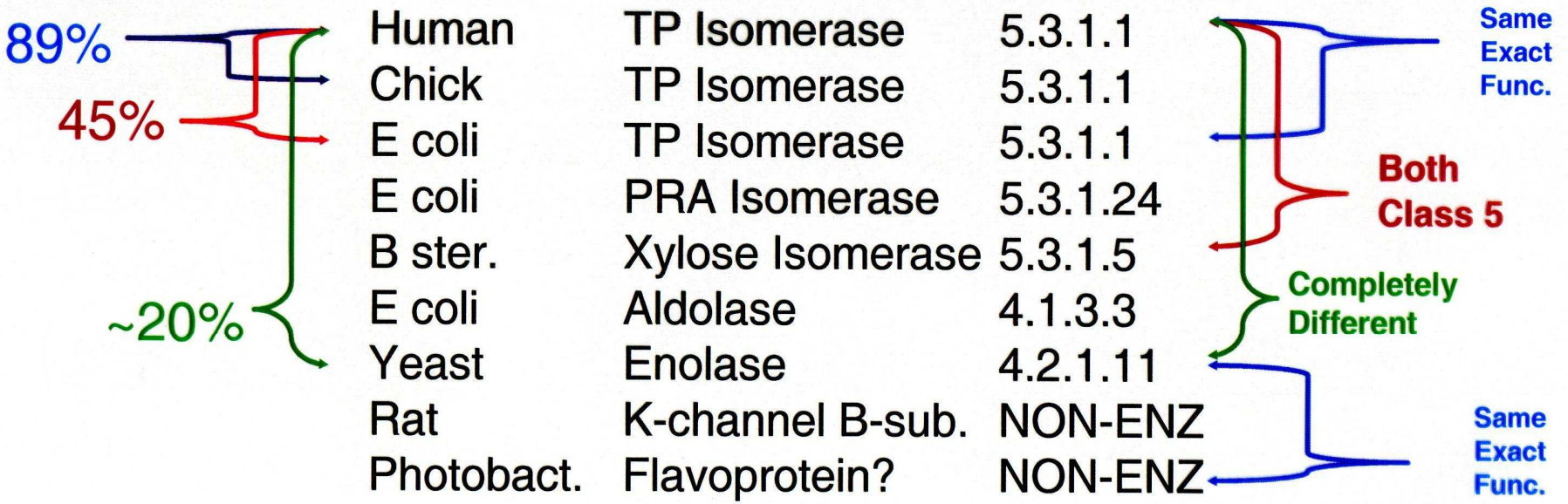
Fold-Function Combinations #2

Many Functions on the Same Fold -- e.g. the TIM-barrel

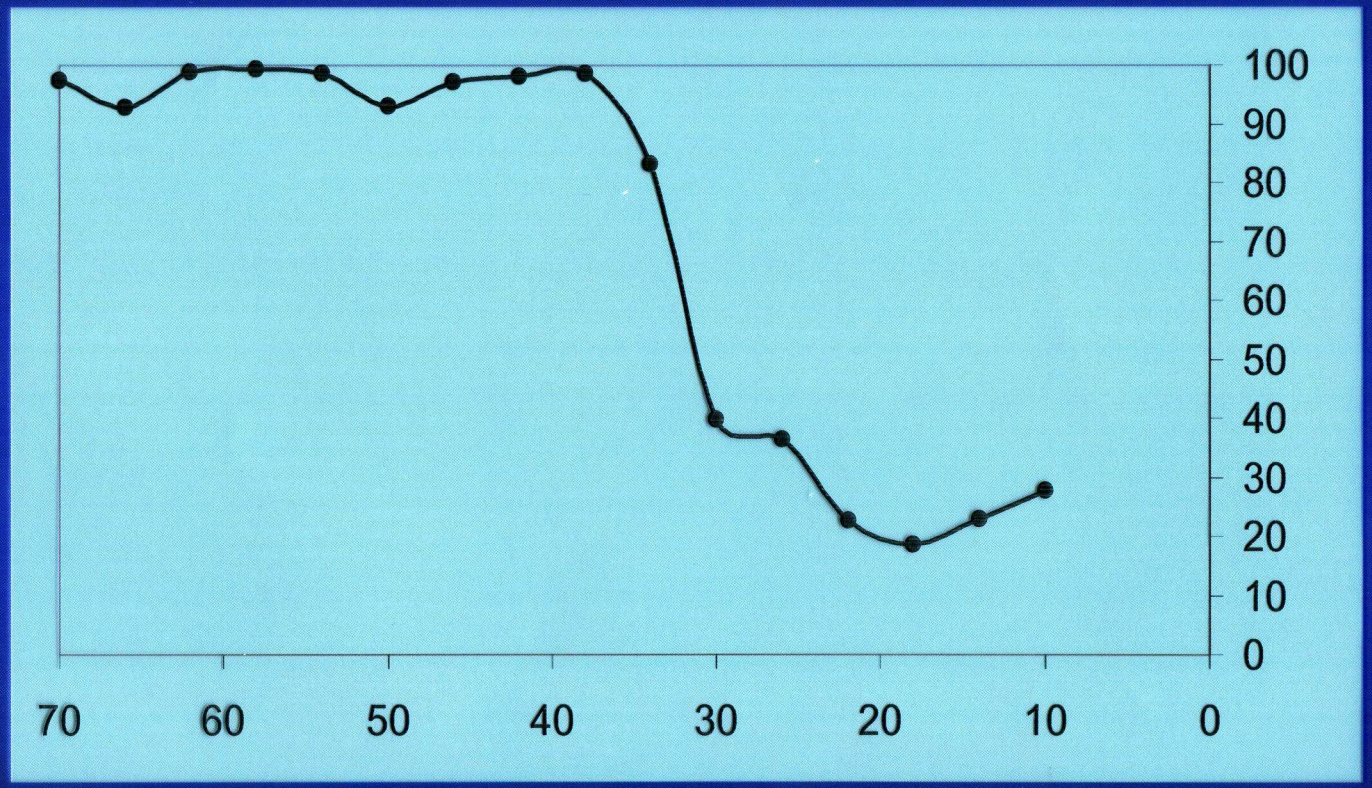
at what degree of divergence?
Sequence Diverg. (**%ID, P_{seq}**)
Structural Diverg. (**RMS, P_{str}**)
Functional Diverg. (**%SameFunc**)



Compare large number of pairs of sequences that have same fold but different functions.



Relationship of Similarity in Sequence to that in Function



Wilson et al. JMB 297: 233

JMB SCIENCE DIRECT® ELSEVIER

How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity?

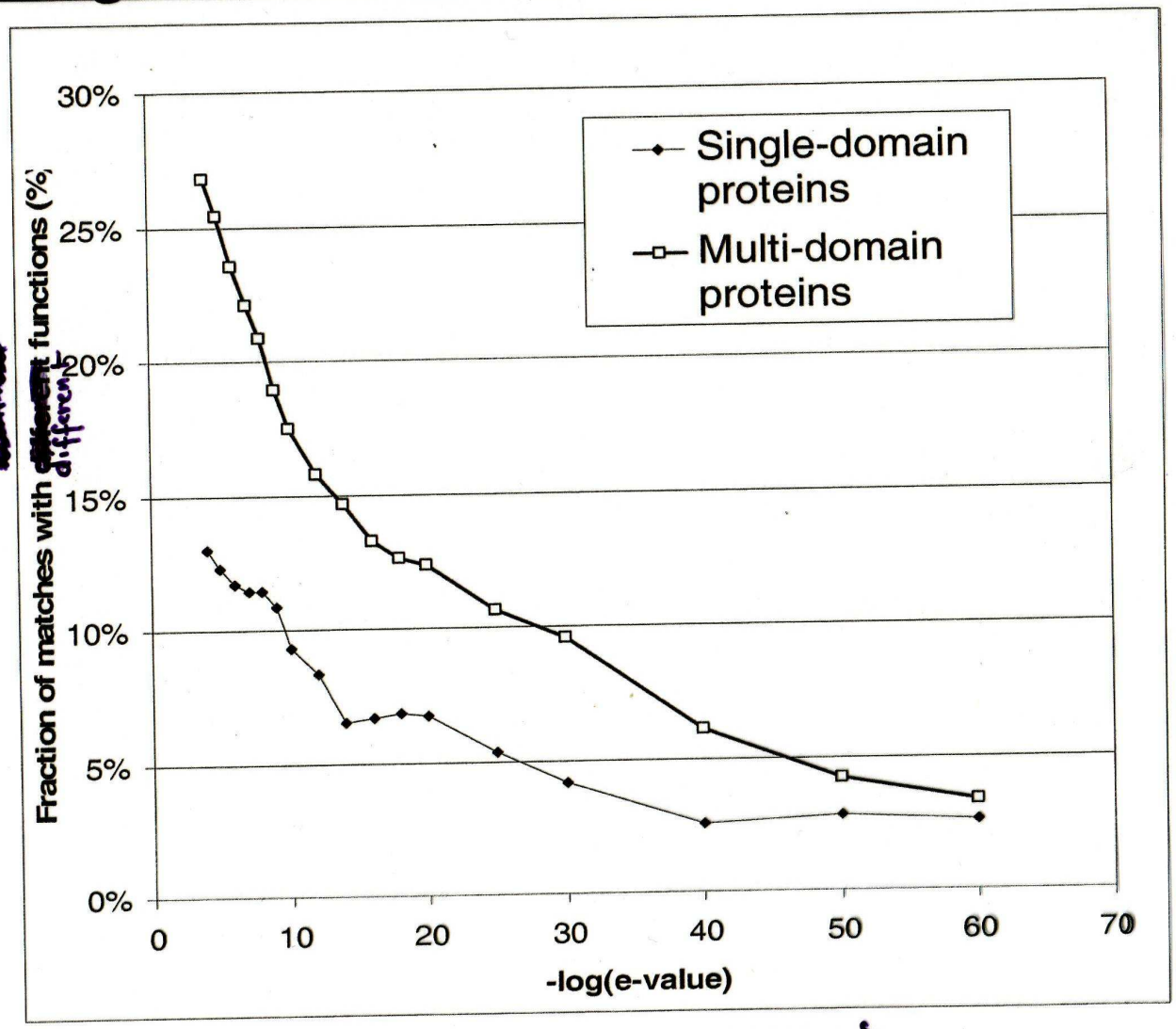
Weidong Tian[1,2] and Jeffrey Skolnick[1*]

[1]*Center of Excellence in Bioinformatics, University at Buffalo, The State University of New York, 901 Washington Street, Buffalo, NY 14203 USA*

[2]*Department of Biology Washington University in St Louis, One Brookings Drive, St Louis, MO 63130 USA*

Enzyme function conservation has been used to derive the threshold of sequence identity necessary to transfer function from a protein of known function to an unknown protein. Using pairwise sequence comparison, several studies suggested that when the sequence identity is above 40%, enzyme function is well conserved. In contrast, Rost argued that because of database bias, the results from such simple pairwise comparisons might be misleading. Thus, by grouping enzyme sequences into families based on sequence similarity and selecting representative sequences for comparison, he showed that enzyme function starts to diverge quickly when the sequence identity is below 70%. Here, we employ a strategy similar to Rost's to reduce the database bias; however, we classify enzyme families based not only on sequence similarity, but also on functional similarity, i.e. sequences in each family must have the same four digits or the same first three digits of the enzyme commission (EC) number. Furthermore, instead of selecting representative sequences for comparison, we calculate the function conservation of each enzyme family and then average the degree of enzyme function conservation across all enzyme families. Our analysis suggests that for functional transferability, 40% sequence identity can still be used as a confident threshold to transfer the first three digits of an EC number; however, to transfer all four digits of an EC number, above 60% sequence identity is needed to have at least 90% accuracy. Moreover, when PSI-BLAST is used, the magnitude of the

Multi-domain proteins have greater divergence in function with sequence



FUNCTION of NEW STRUCTURES (STRUCTURAL GENOMICS)

→ From sequence alignment (Swiss-Prot, PROSITE, Pfam, …)

→ From structure alignment (but dynamics can change the results, for example domain swapping)

→ A recent new approach: finding 3D functional sites: PROCAT, Rigor and PINTS.

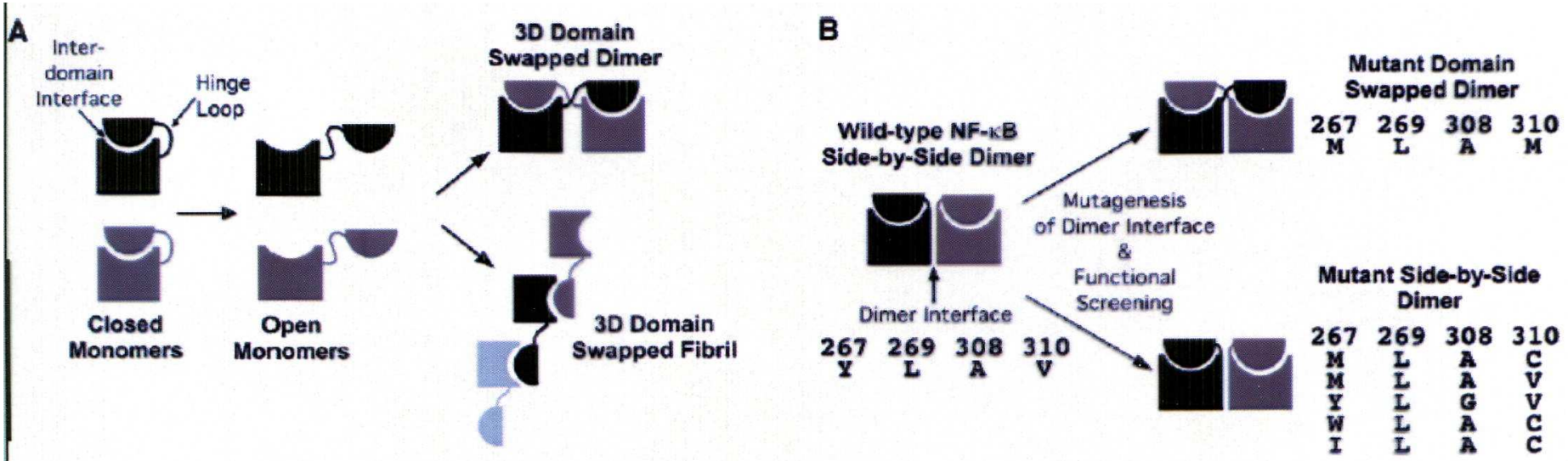


Figure 1. 3D Domain Swapping and the Structures of Wild-Type and Mutant NF-κB p50 Dimerization Domains

(A) Structures involved in 3D domain swapping. Closed monomers contain an interface between domains (square or semicircle) in one contiguous polypeptide chain. Each protein monomer (a single polypeptide chain) is shown in gray or black for clarity. Under some conditions (e.g., change in environment or mutations), open monomers can form (middle). Open monomers can be converted to closed-ended dimers or higher oligomers (top right) or open-ended fibrils (bottom right), in which interdomain interfaces are formed between two or more independent polypeptide chains.

(B) Structures of wild-type and mutant NF-κB p50 dimerization domains. Wild-type NF-κB forms a side-by-side dimer without 3D domain swapping. Each subunit in the homodimer (gray or black) has an Ig-like fold. Twenty-five randomly generated dimer interface mutants that retain DNA binding activity were identified (Hart et al., 2001). Of those structurally characterized (Chirgadze et al., 2004), five mutants form side-by-side dimers (bottom right), and one is a domain-swapped dimer (top right). The swapped domain in NF-κB comprises three β strands of the Ig-like fold.

PINTS (Patterns In Non-homologous Tertiary Structures) finds similar spatial arrangements of amino acids in protein structures that are close in space but not necessarily adjacent in sequence (**local structural patterns**). It is independent of sequence or overall fold similarity. Matches are evaluated by RMSD (root-mean-square deviation) and a statistical E-value.

PINTS is highly complementary to DALI, VAST, or other methods of whole structure comparison, for predicting functional details for proteins of known structure. PINTS can be used to resolve ambiguities that arise during fold comparisons (e.g. it can find the one TIM-barrel that matches a new structure best in terms of function), or suggest convergent similarities not apparent when comparing sequences or complete structures (e.g. such as catalytic triads).

PINTS lets one find answers to questions such as:

- Does my new structure contain a known active site / binding site?
- Which residues in my structure form the active site / binding site?
- Can I find an interesting pattern somewhere else in the PDB?
- Which structural patterns are shared between my two proteins?
- And many more...

PINTS is NOT a structural alignment (i.e. fold comparison) program. For this we recommend programs such as DALI or VAST. PINTS finds local structural similarities (current size limitatation is a diameter of 15 Å) in residue arrangement and requires some conservation in residue types (see below).

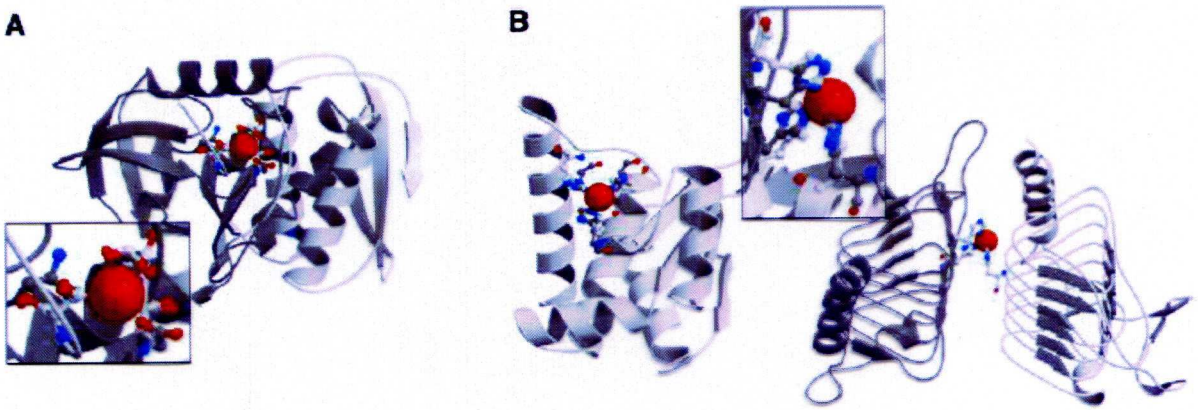
We currently provide six **databases** of whole **structures**...

1. SCOP Folds : one member of each protein fold (as defined by SCOP v. 1.61)
2. SCOP SFams : one member of each protein superfamily (as defined by SCOP v. 1.61)
3. SCOP Fams : one member of each protein family (as defined by SCOP v. 1.61)
4. SCOP P.Species : one structure for each protein (as defined by SCOP v. 1.61)
5. PDB_select 25 : protein chains with < 25% sequence identity (as defined by PDB_select, Apr. 2002)
6. PDB_select 90 : protein chains with < 90% sequence identity (as defined by PDB_Select, Apr. 2002)

... and five databases of **patterns**:

1. Ligand-binding Sites (NR): residues within 3.0 Å of a HETATM in the PDB (cleared and non-redundant: 500 entries)
2. Ligand-binding Sites (Red.): residues within 3.0 Å of a HETATM in the PDB (redundant: 13.000 entries)
3. SITE Annotations : residues from all annotated SITEs in the PDB
4. Conserved Residues: residues that are more than 80% conserved among homologs (>40% identity) - SOON
5. Surface Residues : residues from PDB_Select that are more than 25% exposed (using DSSP accessibility)

PINTS



Stark et al., Structure (2004) 12:1405