Stucture and function

Arne Elofsson

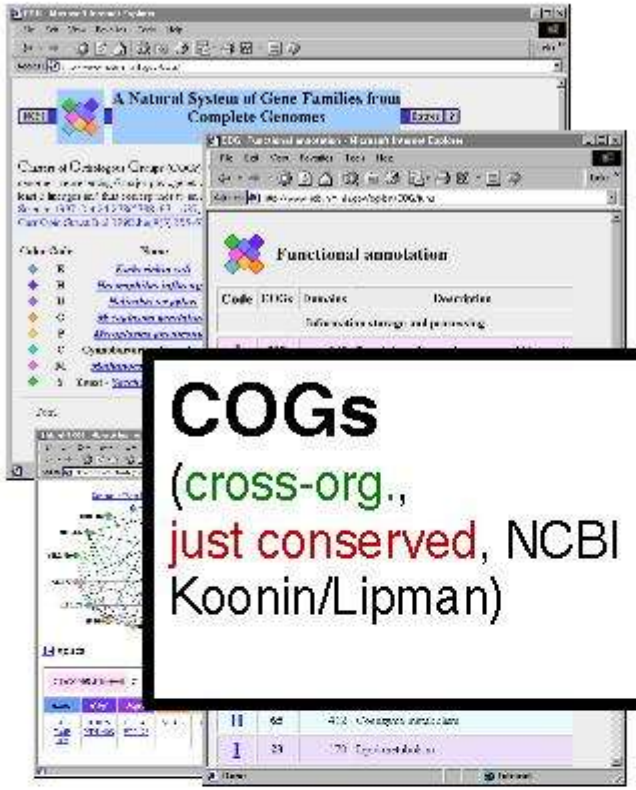
What is Function

- Biochemical function
  - Kinase
  - DNA-binding
- Biological function
  - Cell cycle
- Medical function
  - Cancer related
- Location
  - Mitochondrial

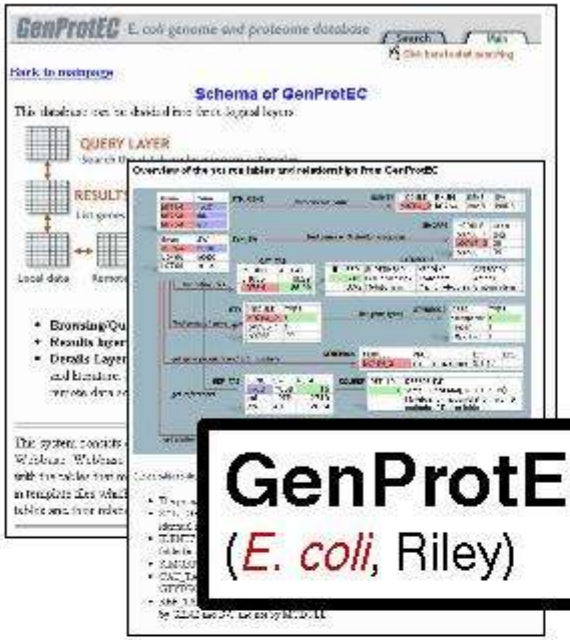
Annotations

- Keywords
- Ontologies
  - Ontologies are 'specifications of a relational vocabulary'. In other words they are sets of defined terms like the sort that you would find in a dictionary, but the terms are networked. The terms in a given vocabulary are likely to be restricted to those used in a particular field, and in the case of GO, the terms are all biological.

Functional Classification



COGs
(cross-org.,
just conserved, NCBI
Koonin/Lipman)

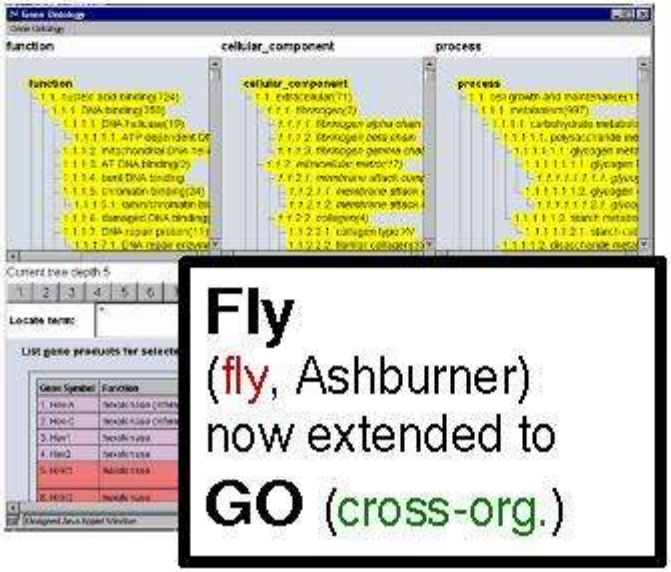


GenProtEC
(*E. coli*, Riley)

ENZYME
(SwissProt
Bairoch/
Apweiler,
just enzymes,
cross-org.)



MIPS/PEDANT
(yeast, Mewes)



Fly
(fly, Ashburner)
now extended to
GO (cross-org.)

Also:

Other
SwissProt
Annotation

WIT, KEGG
(just pathways)

TIGR EGAD
(human ESTs)

SGD (yeast)

(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

EC classifications

First figure	Second figure	Third figure
A. *OXIDOREDUCTASES* Substrate is oxidised-regarded as the hydrogen or electron donor	Describes substrate acted on by enzyme	Type of acceptor
B. *TRANSFERASES* Transfer of a group from one substrate to another	Describes group transferred	Further information on the group transferred
C. *HYDROLASES* Hydrolytic cleavage of bond	Describes type of bond	Nature of substrate
D. *LYASES* Cleavage of bonds by elimination	Type of bond	Further information on the group eliminated
E. *ISOMERASES*	Type of reorganisation	Type of substrate
F. *LIGASES* Enzyme catalysing the joining of two molecules in concert with hydrolysis of ATP	Describes type of bond formed	Describes type of compound formed

GeneOntology
http://www.geneontology.org/

- The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases.
- GO consortium (Examples):
  - FlyBase
  - TIGR
  - annotation of UniProt Knowledgebase
  - Saccharomyces Genome Database (SGD)
  - Etc.....

What GO is not

- GO is not a nomenclature for genes or gene products. The vocabularies describe molecular phenomena (e.g. programmed cell death), not biological objects (e.g. proteins or genes).

GO vocabularies

- Molecular Function (7447 terms)
- Biological Process (9170 terms)
- Cellular Component (1501 terms)

How do I find GO annotations for 'my' genes?

- Several browsers have been created for browsing the GO and finding GO associations for genes and gene products. These can be accessed at the GO Web site. The AmiGO browser, for example, allows searches both by GO term (or a portion thereof) and by gene products. The results include the GO hierarchy for the term, definition and synonyms for the term, external links, and the complete set of gene product associations for the term and any of its children.

Databases with GO annotations

Database	Index File	Source	Date of last update
UniProt Knowledgebase	spkw2go	Evelyn Camon *(Note: spkw2go used to be called swp2go, all files remain the same.)*	Monthly
COG Functional Categories	cog2go	Michael Ashburner and Jane Lomax	June 2004
Enzyme Commission	ec2go	Michael Ashburner	Monthly
EGAD	egad2go	Michael Ashburner	October 2000
GenProtEC	genprotec2go	Heather Butler and Michael Ashburner	December 2000
TIGR Role	tigr2go	Michael Ashburner	January 2004
TIGR Families	tigrfams2go	TIGR Staff	September 2004
InterPro	interpro2go	Nicola Mulder	Monthly
MIPS Funcat	mips2go	Michael Ashburner and Midori Harris	August 2002
MetaCyc Pathways	metacyc2go	Michael Ashburner and Midori Harris	December 2003

GO annotations

- Both manual and automated annotations are made according to two principles: first, every annotation must be attributed to a source, which may be a literature reference, another database or a computational analysis; second, the annotation must indicate what kind of evidence is found in the cited source to support the association between the gene product and the GO term

GO Annotations

- IMP inferred from mutant phenotype
- IGI inferred from genetic interaction [with <database:gene_symbol[allele_symbol]>]
- IPI inferred from physical interaction [with <database:protein_name>]
- ISS inferred from sequence similarity [with <database:sequence_id>]
- IDA inferred from direct assay
- IEP inferred from expression pattern
- IEA inferred from electronic annotation [to <database:id>]
- TAS traceable author statement
- NAS non-traceable author statement
- ND no biological data available
- RCA inferred from reviewed computational analysis
- IC inferred by curator

GO Evidence codes

- IC inferred by curator
- IDA inferred from direct assay
- IEA inferred from electronic annotation
- IEP inferred from expression pattern
- IGI inferred from genetic interaction
- IMP inferred from mutant phenotype
- IPI inferred from physical interaction
- ISS inferred from sequence or structural similarity
- NAS non-traceable author statement
- ND no biological data available
- RCA inferred from reviewed computational analysis
- TAS traceable author statement

GO mappings

- The files contain concepts from systems external to GO e.g. Enzyme Comission numbers, SWISS-PROT keywords and TIGR roles, indexed to equivalent GO terms. The mappings are typically made manually, details can be found in the file header. The files are of the format:
  - external system identifier: external system term name/id > GO: GO term name ; GO id.

GeneOntology Classifications

- # GO:0008150 : biological_process (109503)
- # GO:0005575 : cellular_component (98453)
- # GO:0003674 : molecular_function (108120)

GO: biological_process

- # GO:0007610 : behavior (2414)
- # GO:0000004 : biological_process unknown (28719)
- # GO:0009987 : cellular process (38756)
- # GO:0007275 : development (16478)
- # GO:0007582 : physiological process (70981)
- # GO:0050789 : regulation of biological process (14629)
- # GO:0016032 : viral life cycle (225)

GO: cellular_component

- # GO:0005623 : cell (71940)
- # GO:0008372 : cellular_component unknown (20397)
- # GO:0005576 : extracellular (9217)
- # GO:0031012 : extracellular matrix (960)
- # GO:0043226 : organelle (48954)
- # GO:0043234 : protein complex (9408)
- # GO:0019012 : virion (96)

GO: molecular_function

- * GO:0016209 : antioxidant activity (478)
- * GO:0005488 : binding (31317)
- * GO:0003824 : catalytic activity (35260)
- * GO:0030188 : chaperone regulator activity (14)
- * GO:0030234 : enzyme regulator activity (2087)
- * GO:0005554 : molecular_function unknown (29597)
- * GO:0003774 : motor activity (522)
- * GO:0045735 : nutrient reservoir activity (36)
- * GO:0004871 : signal transducer activity (8356)
- * GO:0005198 : structural molecule activity (3428)
- * GO:0030528 : transcription regulator activity (8552)
- * GO:0045182 : translation regulator activity (687)
- * GO:0005215 : transporter activity (9054)
- * GO:0030533 : triplet codon-amino acid adaptor activity (555)

Open Biology Onthologies

Domain	Prefix	Ontology	Defs file
Arabidopsis gross anatomy	TAIR	arabidopsis anatomy.ontology	arabidopsis anatomy.definitions
Arabidopsis development	TAIR	arabidopsis development.ontology	arabidopsis development.definitions
Cell type	CL	cell.obo	included in cell.obo
Cereal plant gross anatomy	GRO	anatomy gr ont	anatomy gr def
Cereal plant development	GRO	temporal gr ont	temporal gr def
Cereal plant trait ontology	TO	trait ontology	trait definitions
Chemical entities of biological interest	CHEBI	ontology.obo	included in ontology.obo
Protein covalent bond	CV	[none]	[none]
Protein-protein Interaction	MI	psi-mi.dag	psi-mi.def
Maize gross anatomy	ZEA	Zea mays anatomy ontology.txt	Zea mays anatomy ontology definitions.txt
Dictyostelium anatomy	DDANAT	anatomy.ontology	anatomy.definitions
Drosophila gross anatomy	FBbt	fly anatomy.ontology	fly anatomy.definitions
Habronattus courtship		protege source	included in protege source
Loggerhead nesting		protege source	included in protege source
Human anatomy and development	EV	ontologies	[none]
Microarray experimental conditions		MGEDOntology.daml	included in MGEDOntology.daml
Physical-chemical methods and properties	FIX	fix.ontology	[none]
Fungal gross anatomy	FAO	fungal anatomy.ontology	fungal anatomy.definitions
Molecular function	GO	gene_ontology.obo	included in gene_ontology.obo
Biological process	GO	gene_ontology.obo	included in gene_ontology.obo
Cellular component	GO	gene_ontology.obo	included in gene_ontology.obo

How is function and structure related

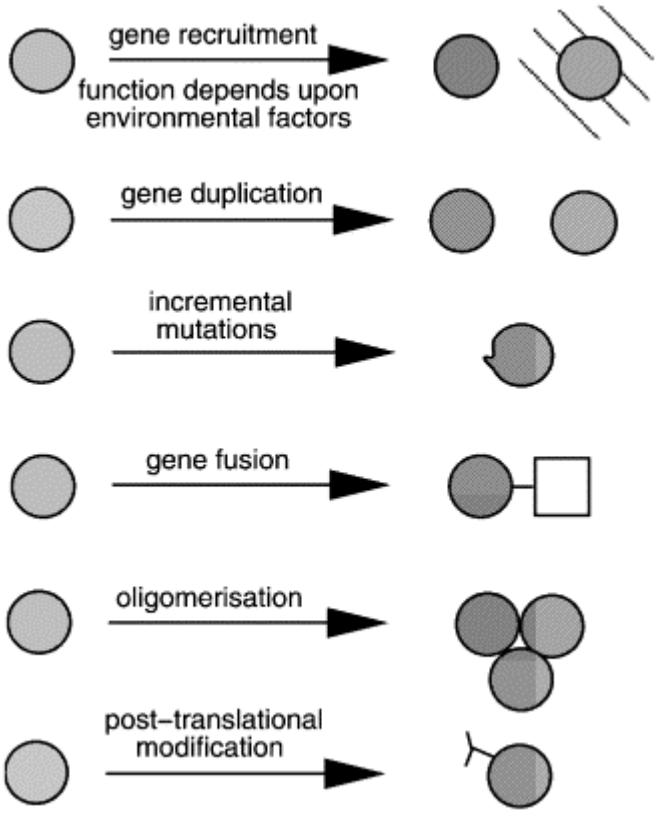
- Molecular Function most structure related
- Function by homology
  - But close homologs might have very different functions
- Function ab-initio from structure
  - Active site residues

Functional Evolution

- Gene Duplications
  - Orthologs are expected to have more similar functions
  - >95% (88%) of all genes origin from duplications in human (yeast)
- Gene Fusion
  - If two protein are fused in one organism the two individual proteins are often “functionally related”

Some mechanisms new functions are created

- Gene recruitment
- Post translational modifications
- Alternative splicing
- Gene duplications
- Incremental mutations
- Gene fusion
- oligomerization



One gene, two or more functions

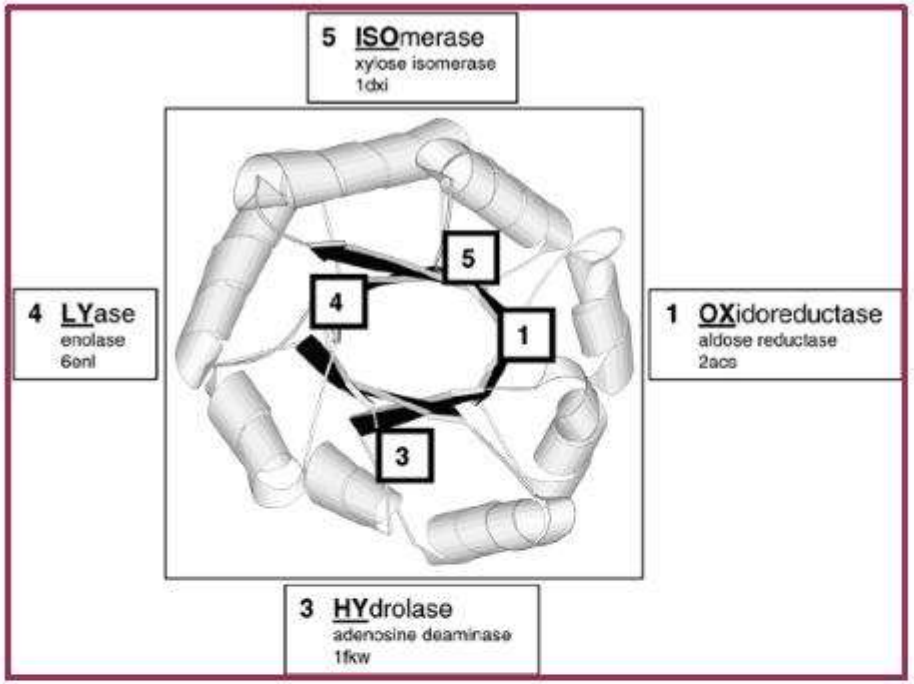
- Recruited for new functions
  - Enzymes to Crystallins
- Post-translational modifications
  - Non -identical proteins
- Alternative splicing
  - Non-identical proteins

From Structure to function

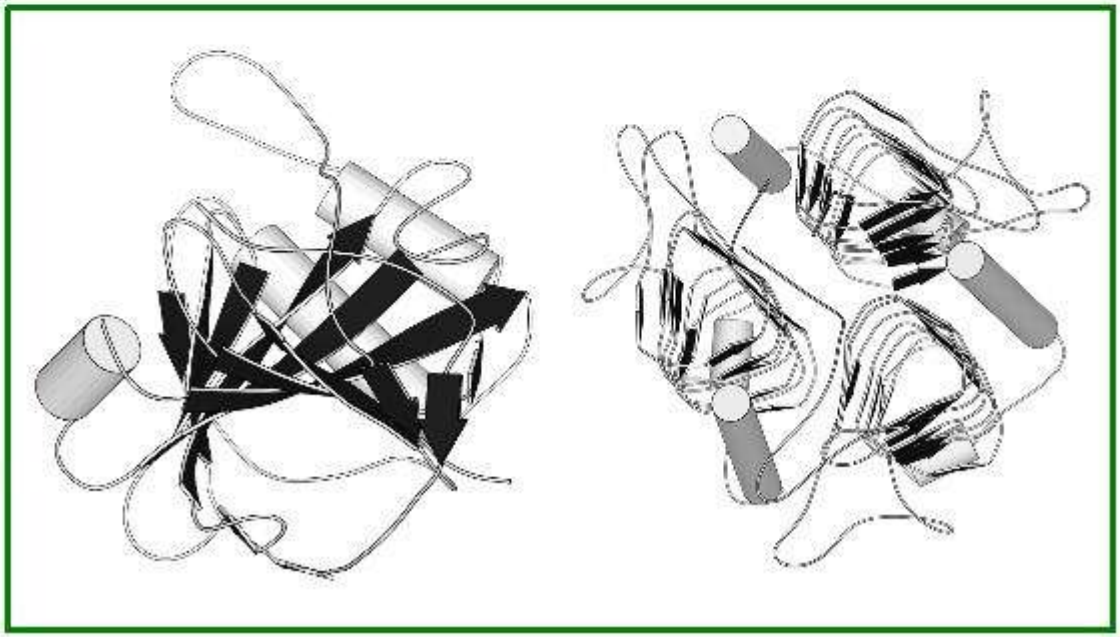
- Ligands bound provide functional clues
- Conserved residues are often functionally important

Fold-Function Combinations

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



Two Different Folds Catalyze the Same Reaction -- e.g. Carbonic Anhydrases (4.2.1.1)



(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

Some examples

- Loss of enzyme activity
  - Duck crystallin and non-enzyme 94% identical
  - Enzyme/non-enzyme
    - Human lysozyme vs human lacalbumin, 40 %ID
  - Identical functions and homologs
    - Hemoglobin from P. Marinus and V. Stercoraria 8% ID
  - Different enzymatic activitied, same superfamily
    - Adenylyl cyclase (EC.4.6.1.1) and DNA pplymerase (EC 2.7.7.7) 12 % ID

More examples

- Similar folds different functions
  - Acylphosphatase, DNA binding domain
- Different folds, identical function
  - B-lactamse class B vs class A, C, D (EC 3.5.2.6)
- Different folds samt function
  - Serine endopeptidases
  - Subtilisin (EC 3.4.21.62) and chymotrypsin (EC 3.4.21.1)

Structural class and function

- Heme – alpha proteins
- DNA-binding alpha or alpha/beta
- Nucleotide binding alpha/beta
- Enzymes non-alpha

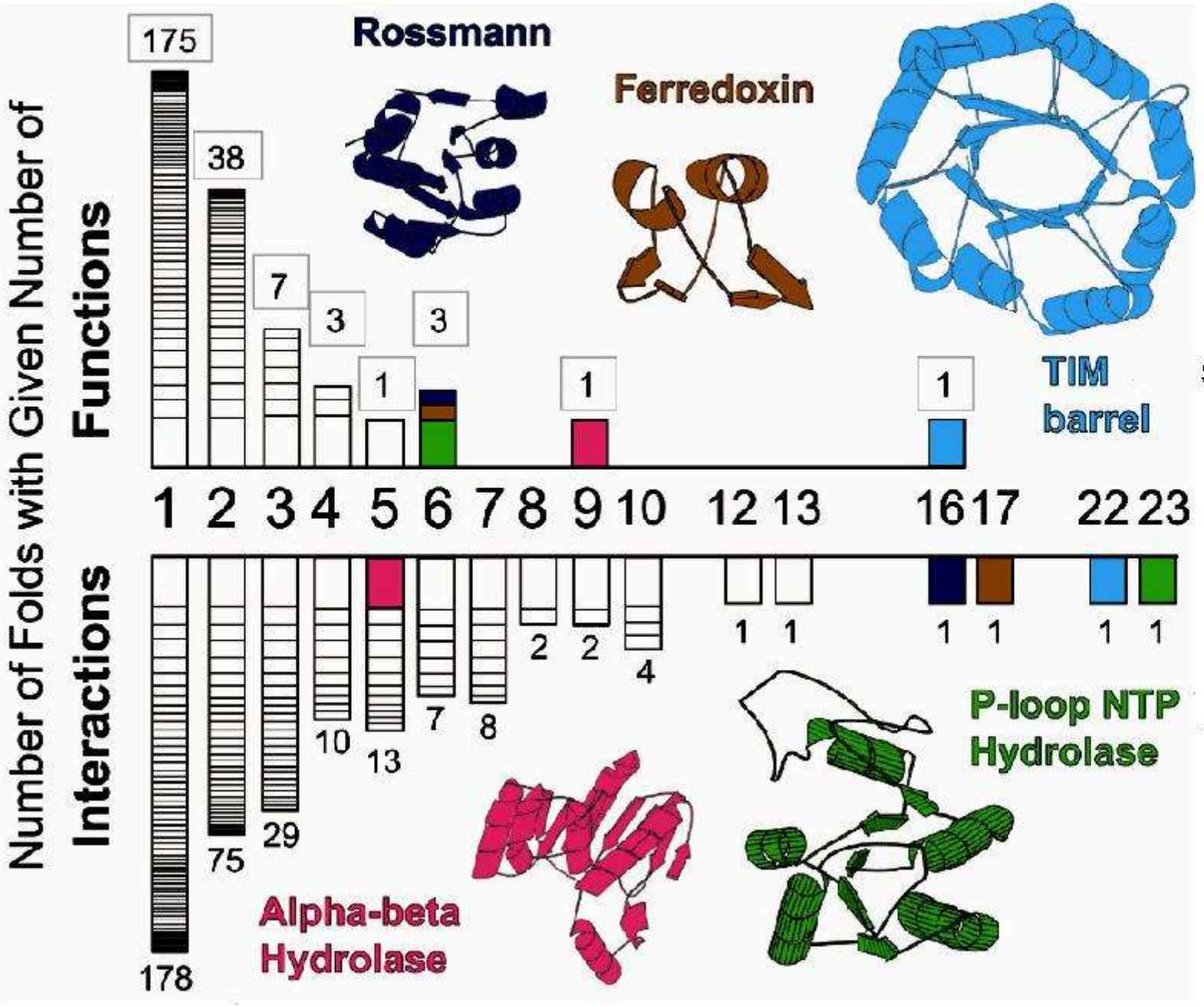
Homology and function classifications

- Orthologs are thought to be more conserved functionally
  - No good test done to my knowledge
- Conservation of functional sites
  - If active site is conserved, functions are often conserved
  - Active blast

Fold similarity and structural analogs

- Families with many functions
  - Superfolds or frequently occuring domains
    - Tim barrels
    - Rossman folds

Most Versatile Folds ? Relation to Interactions



Similar results Martin et al. (1998)

The number of interactions for each fold = the number of other folds it is found to contact in the PDB

(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

Functional predictions without homology

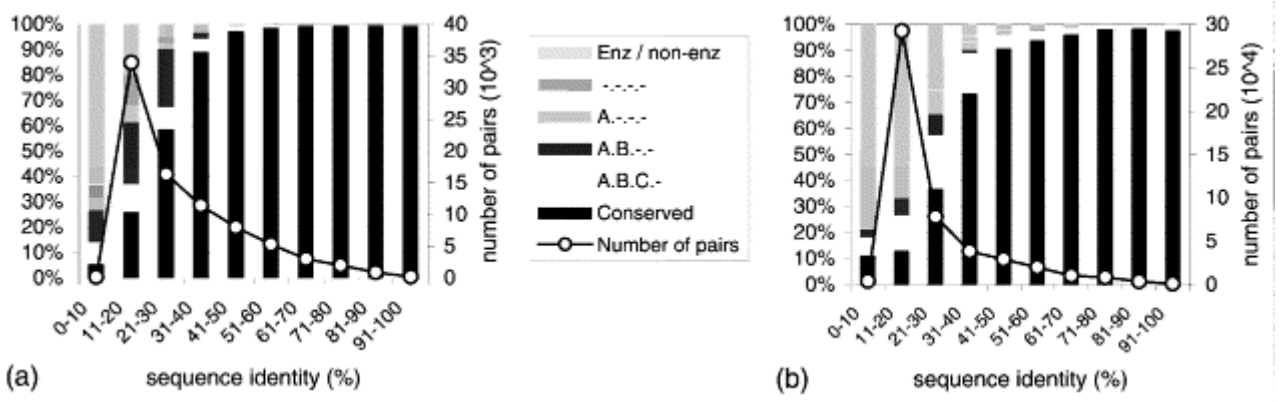
- Identification of active residues
  - TESS, PROCAT
    - Search for active sites
  - SPASM, RIGOR
  - FFF
  - Protein Sidechain patterns

Structural analysis

- Identificaton of active site
  - On the surface but in a cleft
  - Conserved
- Interaction sites
  - Highly exposed
  - Hydrophobic

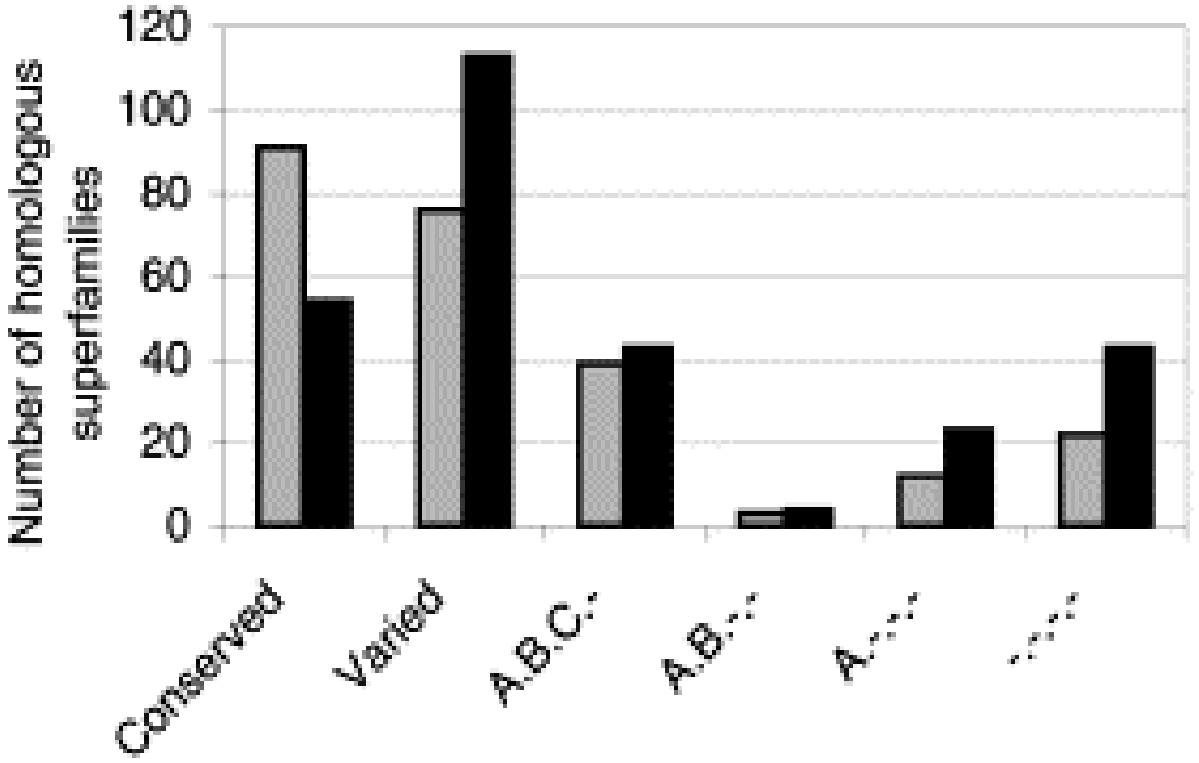
Evolution of a protein function from a structural perspective

- Study of 31 functionally diverse enzyme superfamilies, by Todd et al 2001



Substrate specicity

- 19/31 completely diverse specificity

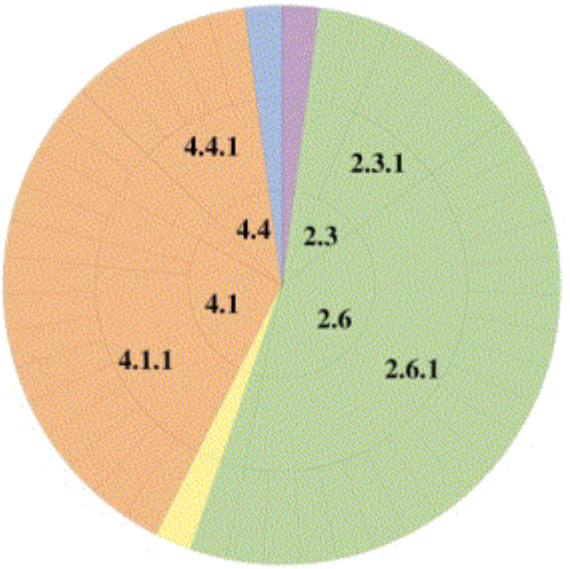


Reaction Chemistry

- Conserved chemistry
  - 2/28 the reaction chemistry is conserved
- Semi-conserved chemistry
  - 21/28 families
- Poorly conserved chemistry
  - 3/28 families
- Variation in chemistry
  - 2/28 families

Catalytic residues

- Same active site framework may be used to catalyze a host of diverse activities
  - / hydrolase superfamily
- Different catalytic apparatus may exist in related protein with very similar function
  - (SER-HIS-ASP triad)



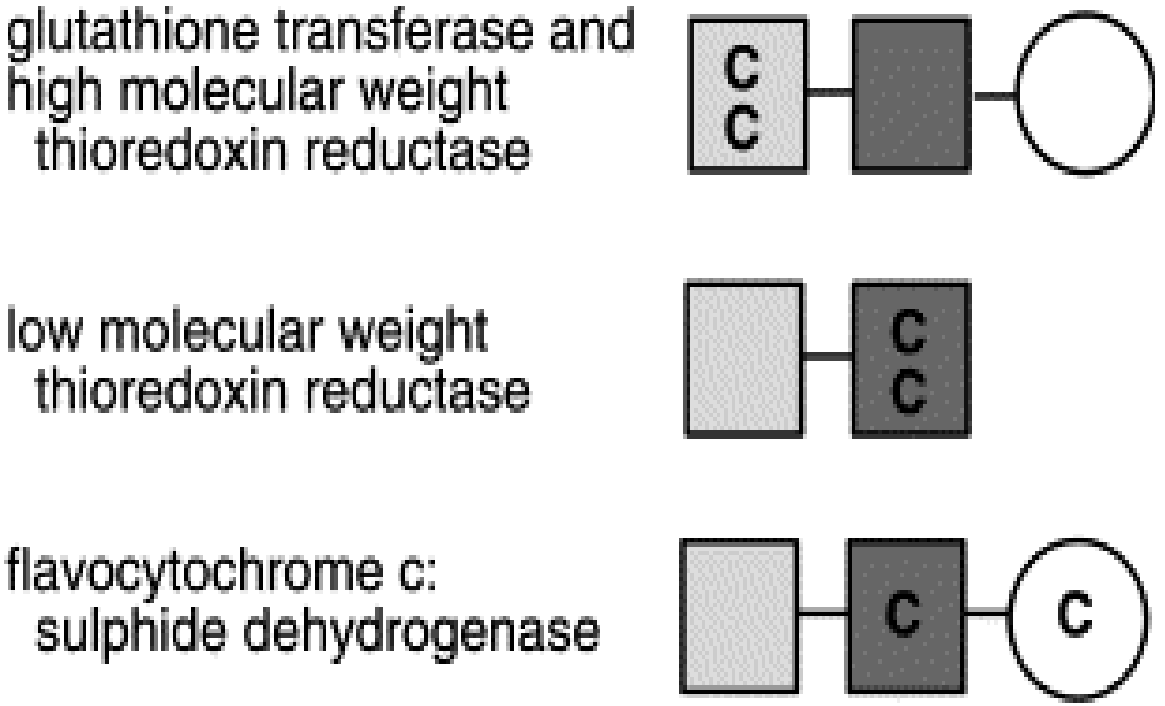
Diversity of enzyme functions catalysed by members of the PLP-dependent type I aspartate aminotransferase superfamily.

Domain enlargement

- Functional core of a domain
  - Superfamilies varies in size
    - 11/31 more than 50% in size
  - Helixes/sheets are added/deleted
    - Addition more common than deletetions
  - Oligomeric state

Domain organization

- Gene fusiona and gene rearrangements



Structural Genomics

- High througput structural determination
  - http://targetdb.pdb.org/
- For some proteins only sequence and structure will be known, how will function be known ?
  - Bound ligands etc.
  - Similarity
  - Guessing

Status report by center

SGTDB release 20-SEP-04

69658 unique target proteins

20centers worldwide

1409solved structures in PDB

progress in the past 2 weeks

398 new structures deposited in PDB this past week

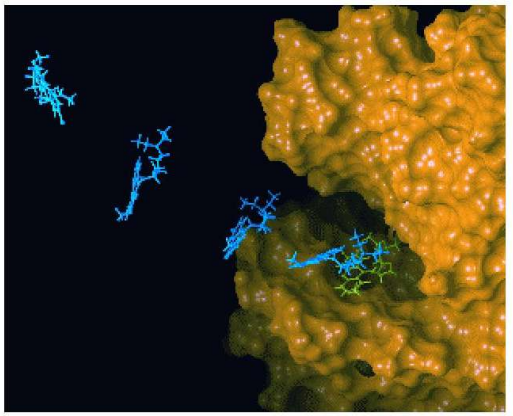
Structural Based genome assignments

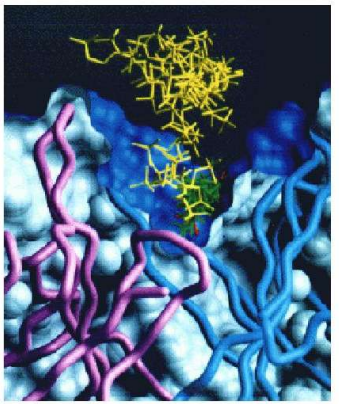
- Assignments of structures
- Analysis
- Evolutionary consequences

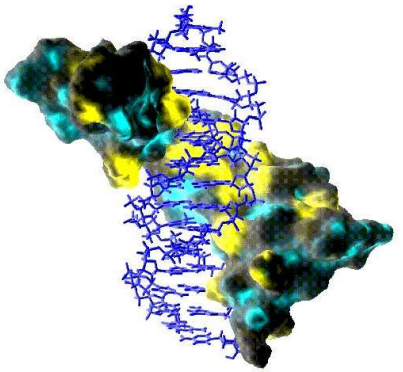
Major Application I: Designing Drugs

- Understanding How Structures Bind Other Molecules (Function)
- Designing Inhibitors
- Docking, Structure Modeling

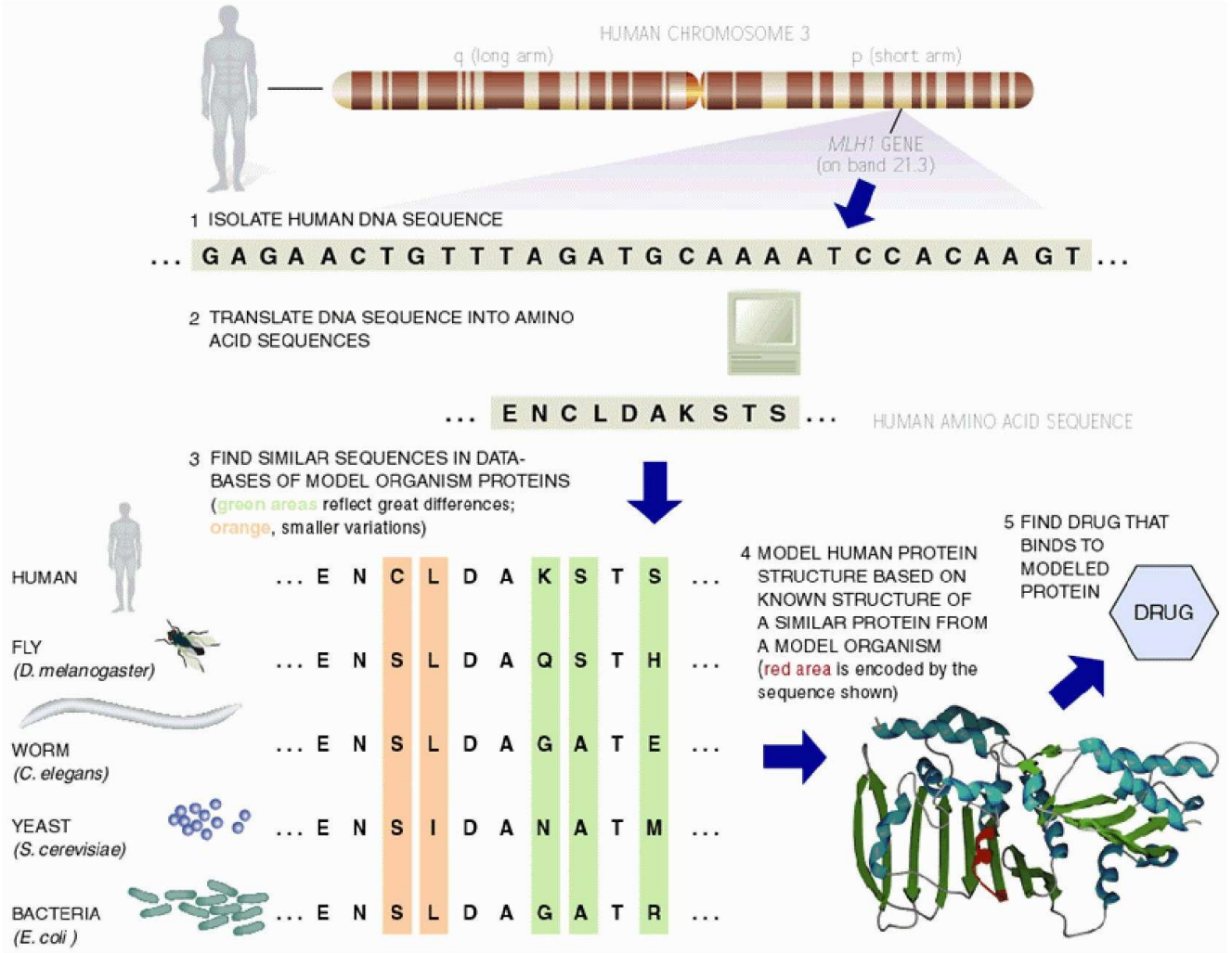
(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).







Major Application II: Finding Homologs

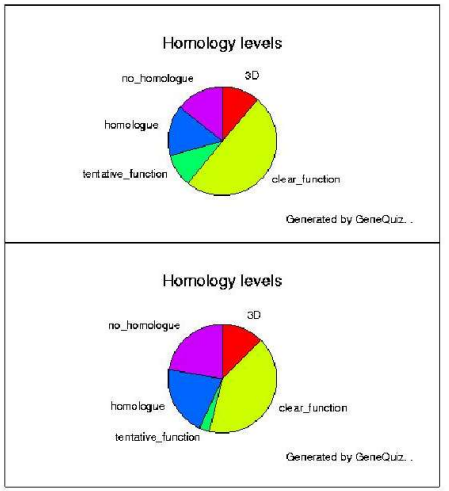


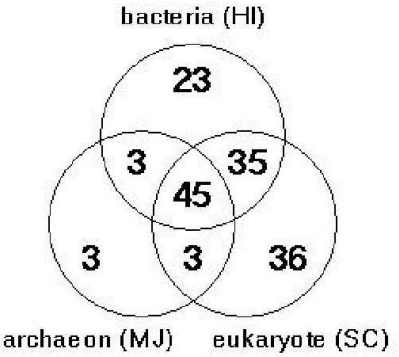
(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

Major Application I|I: Overall Genome Characterization

- Overall Occurrence of a Certain Feature in the Genome
  - ◊ e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - ◊ Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics

(Clock figures, yeast v. Synechocystis,
adapted from GeneQuiz Web Page, Sander Group, EBI)





(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

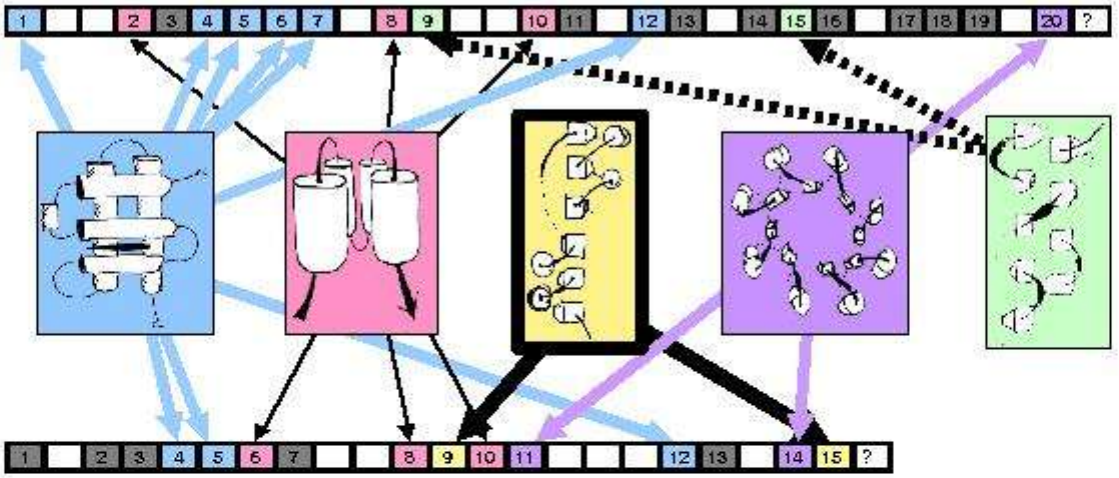
Global Surveys of a Finite Set of Parts from Many Perspectives



(human)

(*T. pallidum*)



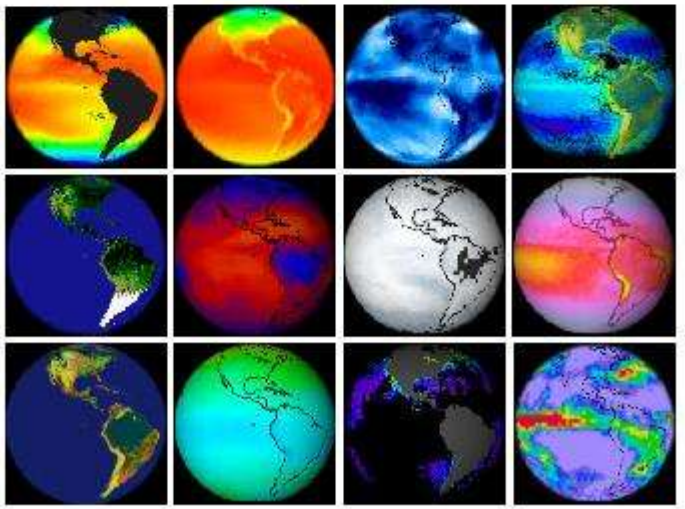


~30000 genes

~1000 folds

~1000 genes

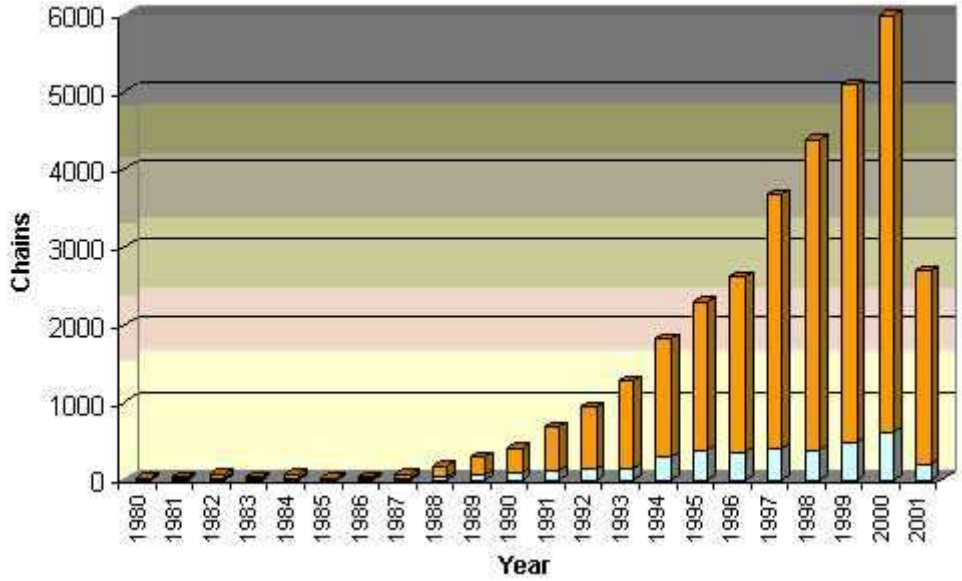
World of structures is even more finite than that of sequences, providing simplification?

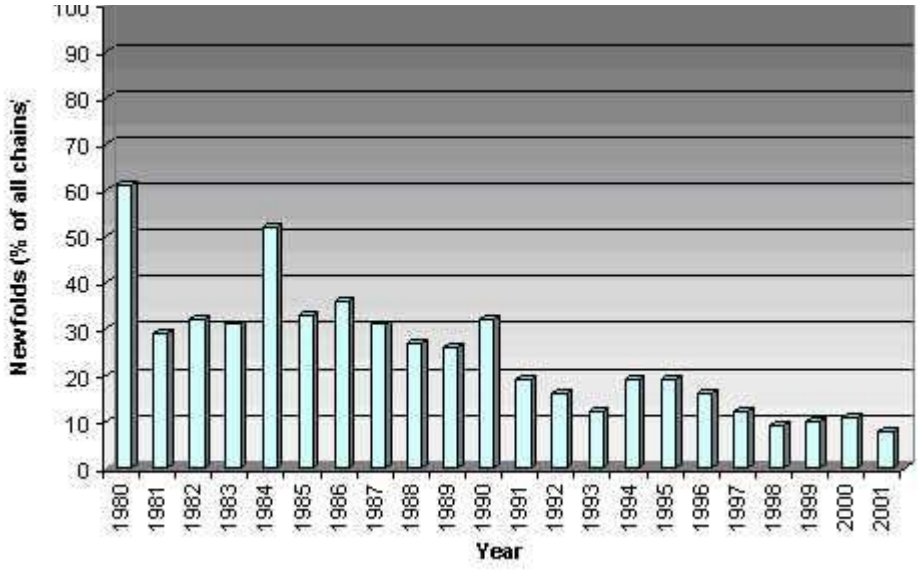


(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

Methods to assign structures

- Fold Recognition
  - 3D-PSSM
- HMMs
  - Superfamily
- SCOP/CATH
- PFAM ?





Folds as the best parts, the final annotation for the human genome. Why?

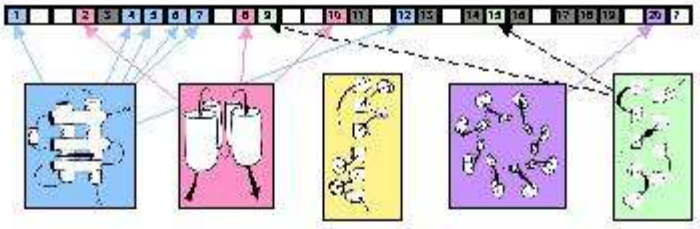
1 Most Highly Conserved

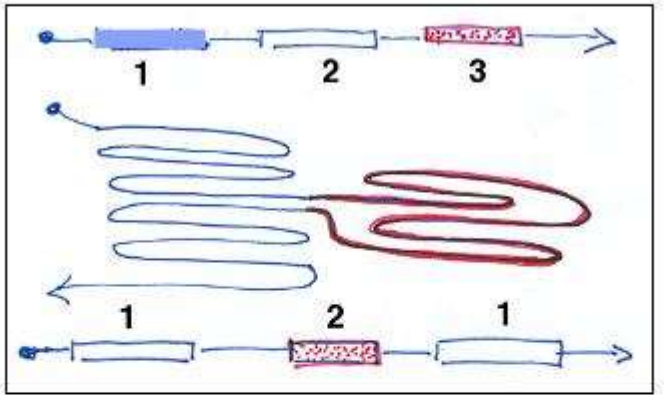
2 Precisely Defined Modules

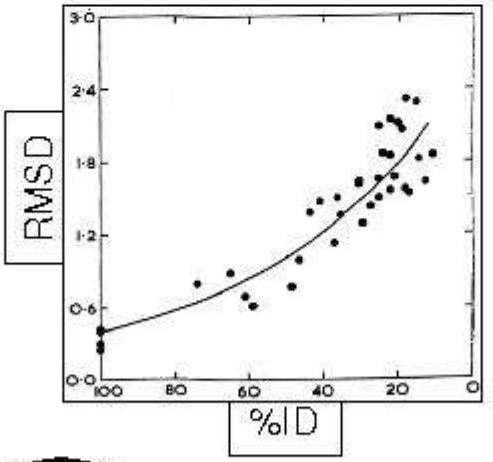
3 Seq. ⇔ Struc. Clearer than Seq. ⇔Func.

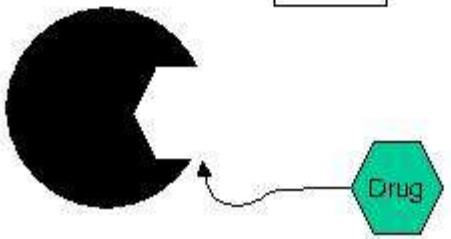
4 Link to Chemistry, biochemical function, drugs

Here mostly
scop+astral+auto-alignments
(Chothia, Levitt, Brenner, Murzin)



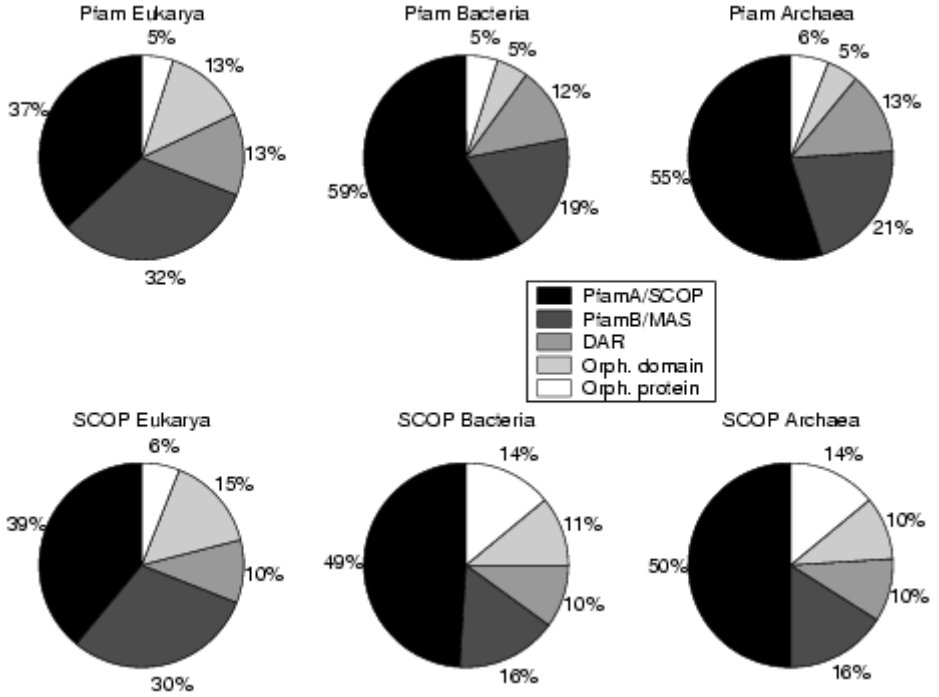




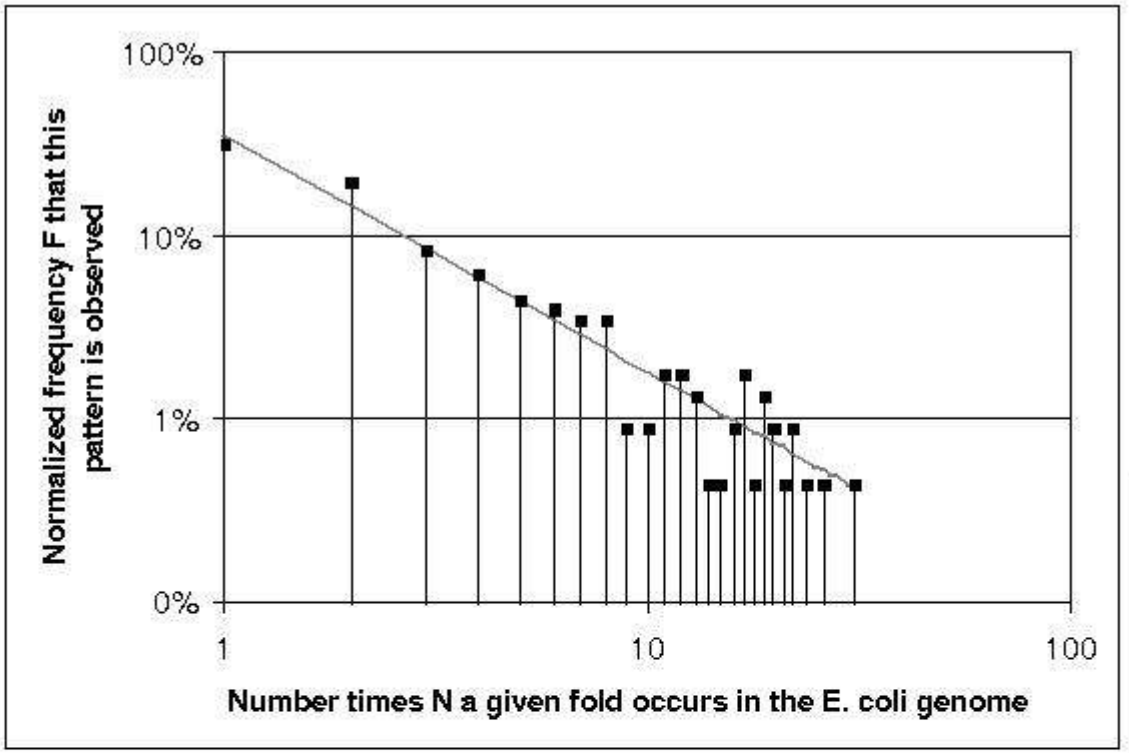


(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

Fraction residues assigned

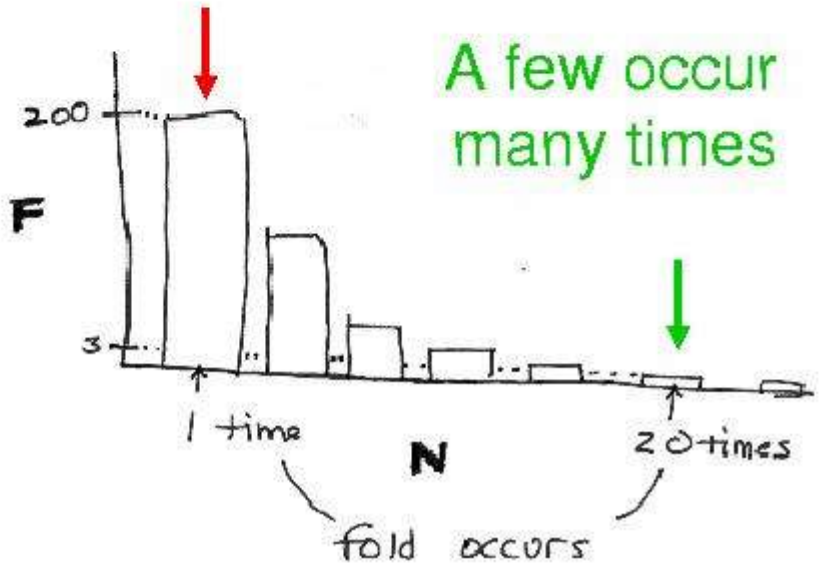


Part Occurrence Frequencies Follow a Power Law



Most occur just once

A few occur many times



Overall falloff follows a Power Law (Zipf law, ?80-20? rule)

$$F = AN^{-B}$$

(A=.35, B=1.2 above)

(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

analysis

- Some domains very common
- Some very rare (only one copy)
- Some duplicated very often

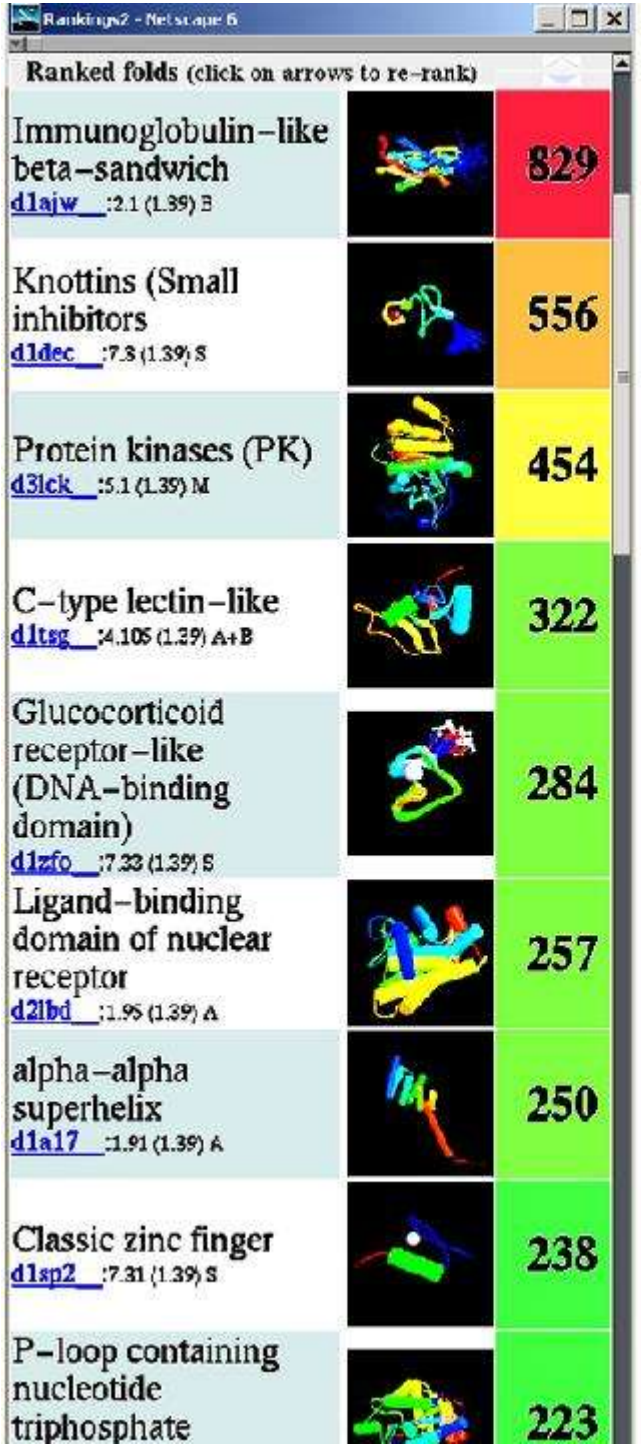
Evolution of power law behavior

- Duplications
  - Larger families are more likely to duplicate
  - Can reproduce most of what is seen
- 50% still remains unclassified
  - Many orphan domains
  - What are the origin of these

Most frequent domains

- E.coli (Eubacteria)
  - P-loop hydrolase
  - Rossman
- Meth. Thermo. (Archeae)
  - P-loop
  - Ferrodoxins
- Sacch. Cerv. (Unicellular eukaryot)
  - P-loop hydrolase
  - Protein Kinase
- C. Elegans

Common Folds



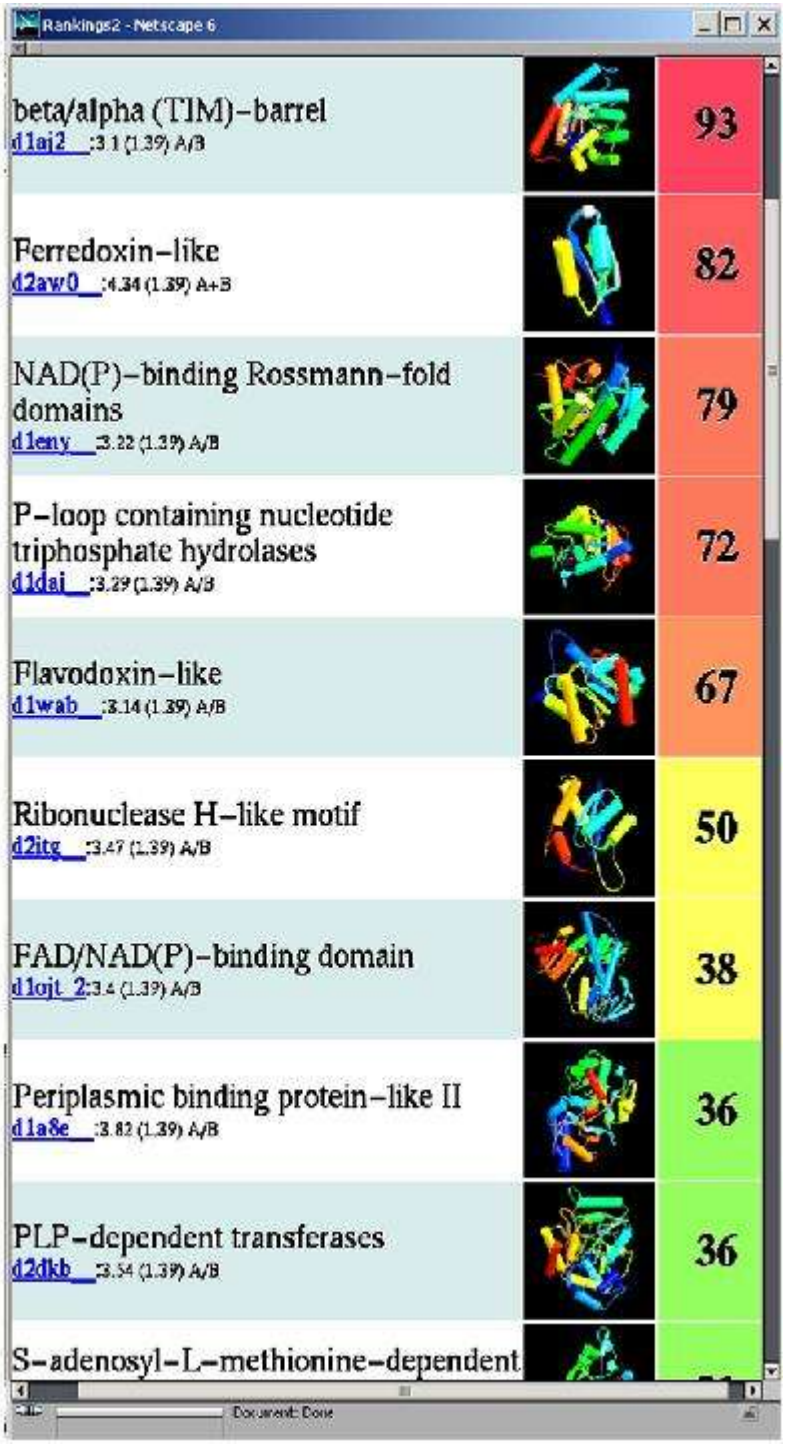
⇦ **Worm**

Extracellular, signaling and trans. factor folds (Ig, kinase, lectin, nuc. receptor, ZnF)

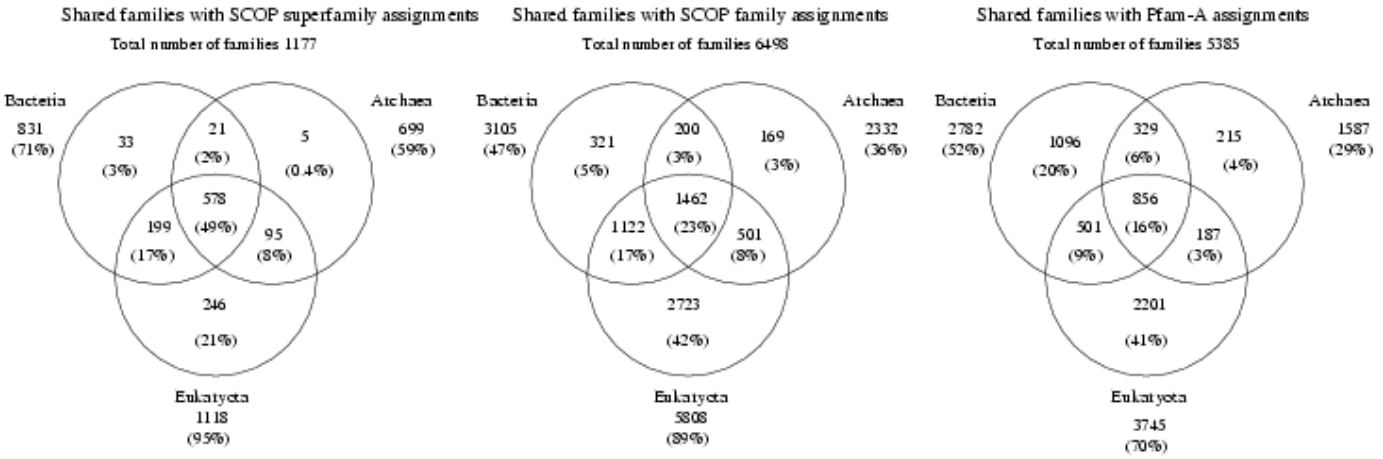
E. coli ⇨

Metabolic Folds (TIM, Ferrodoxin, Rossman, P-loop hydrolase, FAD-binding)

(partslist.org)



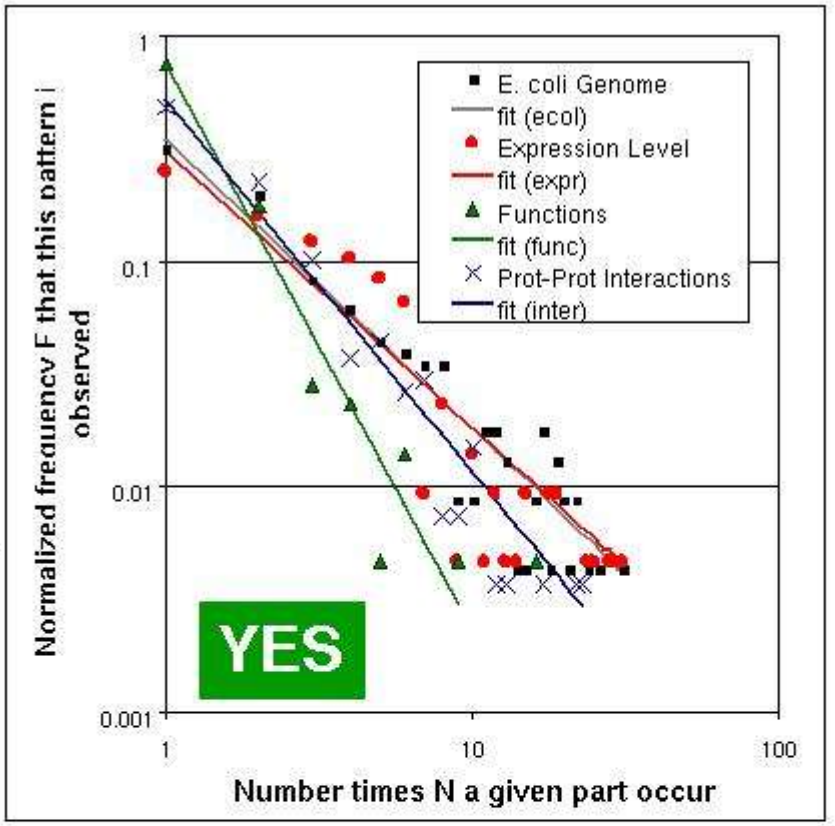
Kingdom specific domains

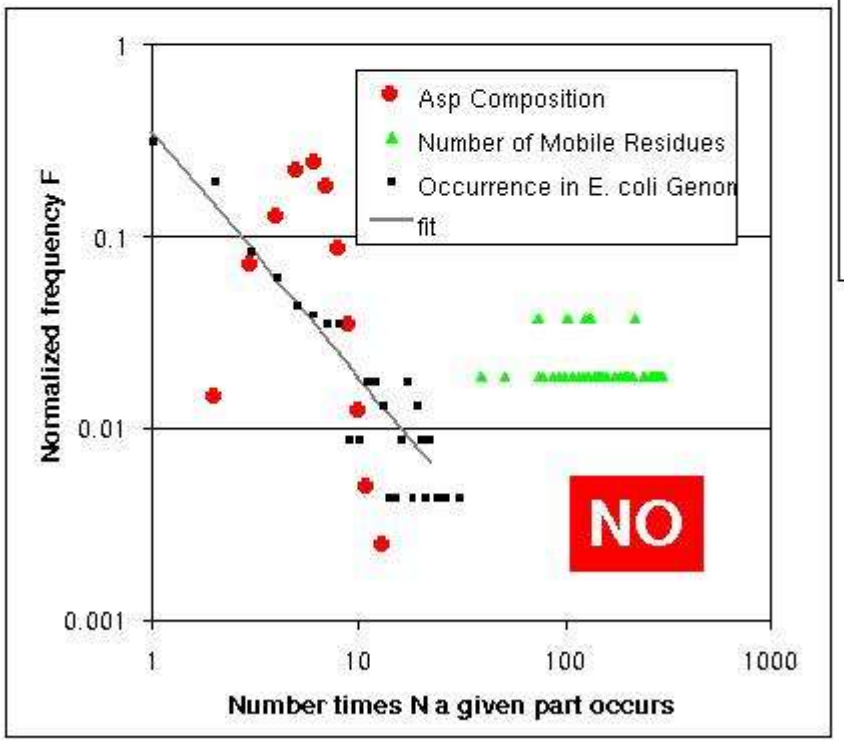


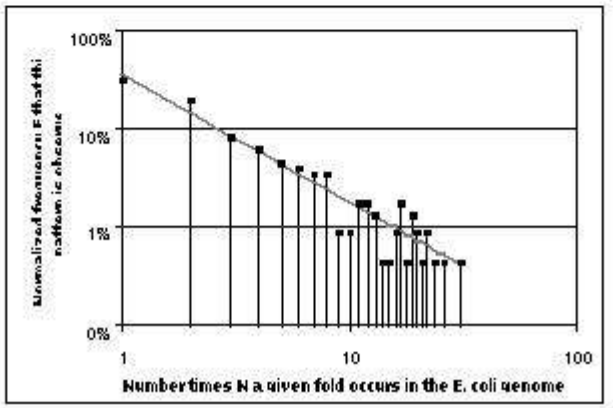
Domain combianation in multidomain families

- Only a small set of all possible combinations seen
- Large families have more combinations
- Conserved N-to-C terminal orientation
- Many combinations are specific to one kingdom
- Multidomain protein involved in cell adhesion

Many (but not all) Part Quantities Follow Power Laws





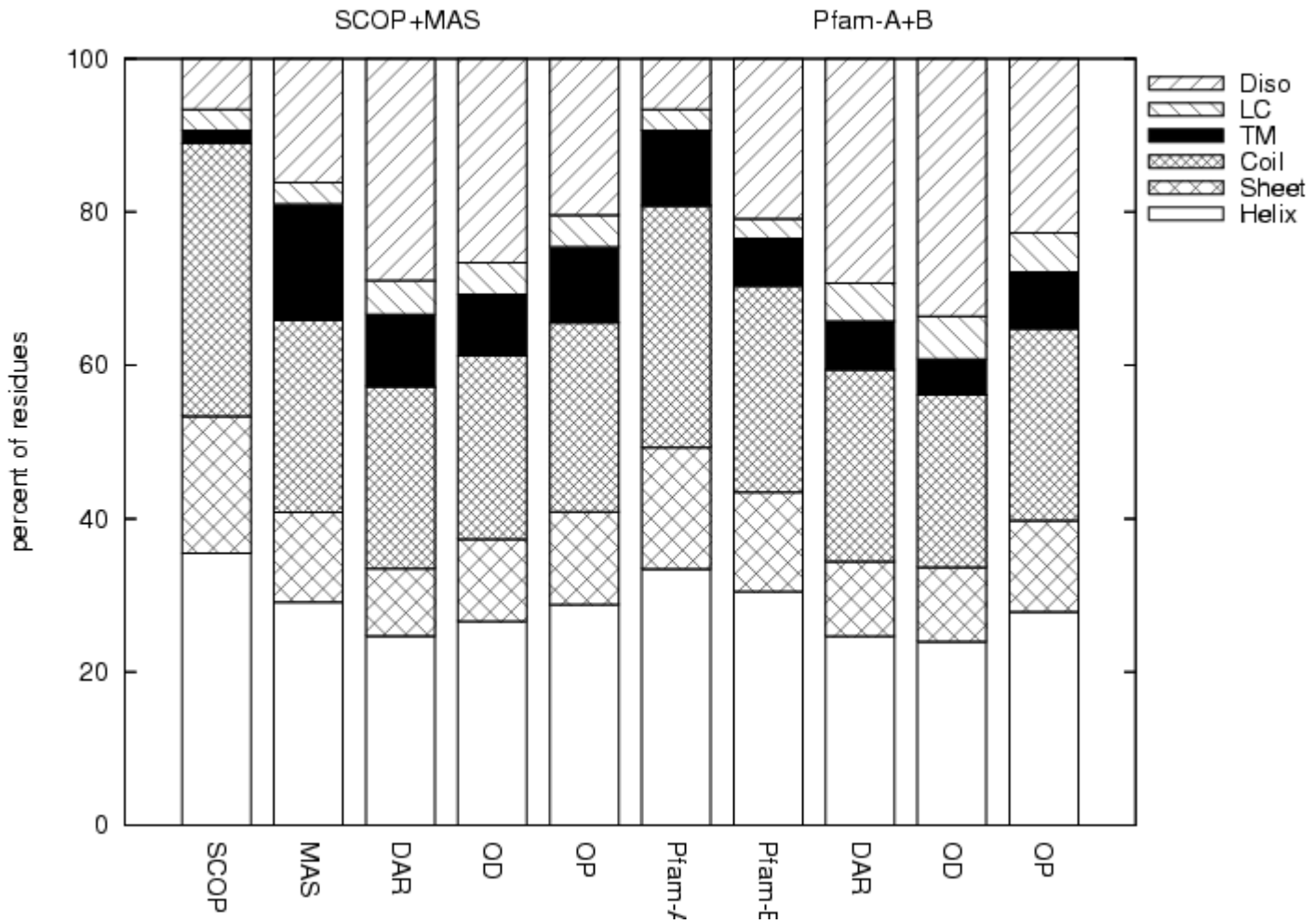


(c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

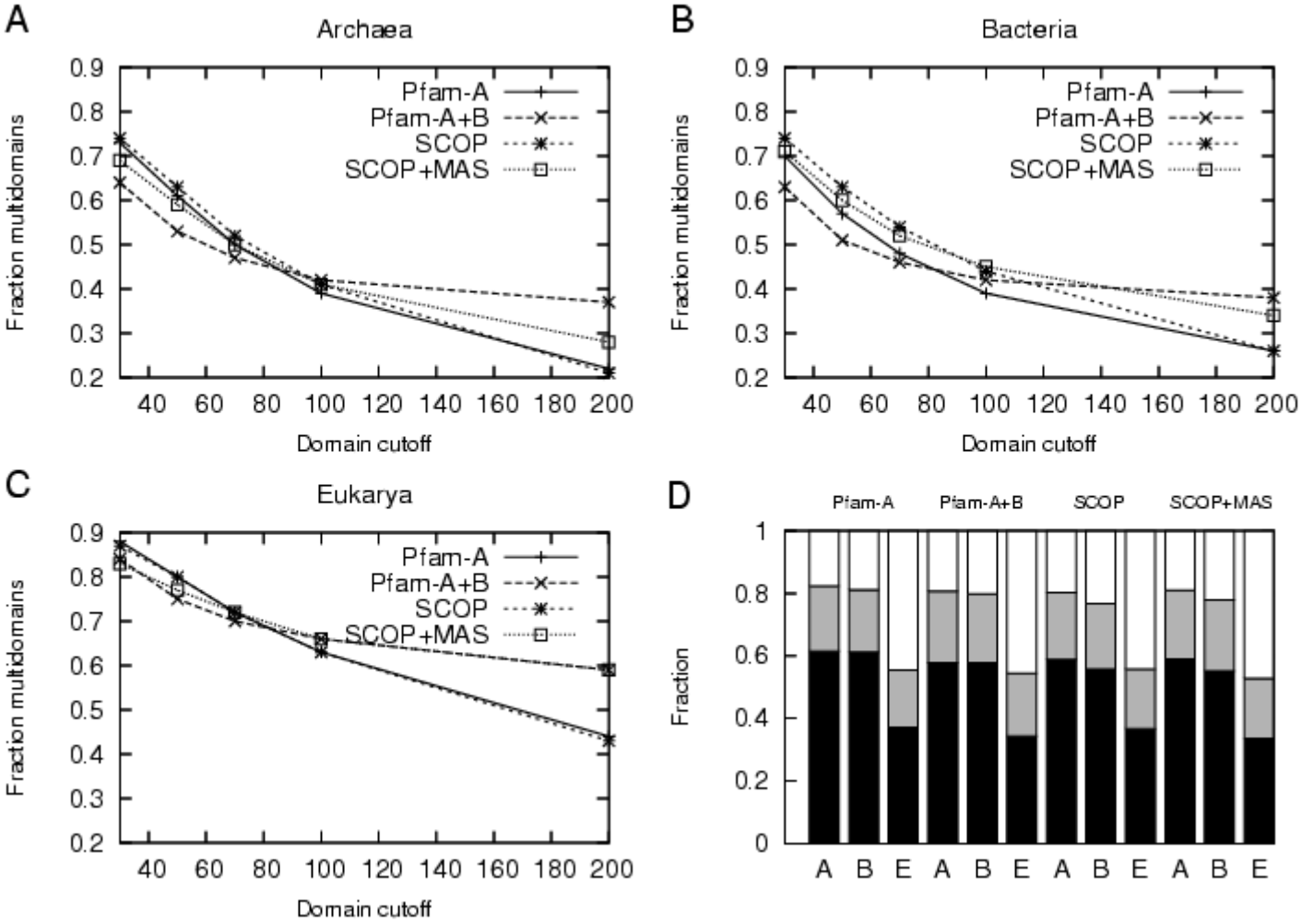
The Unassigned Regions

- 50% of residues can not be assigned
- Some features can be predicted
  - Secondary structure
  - Coiled coils
- Membrane proteins not assigned by SCOP
  - But by Pfam

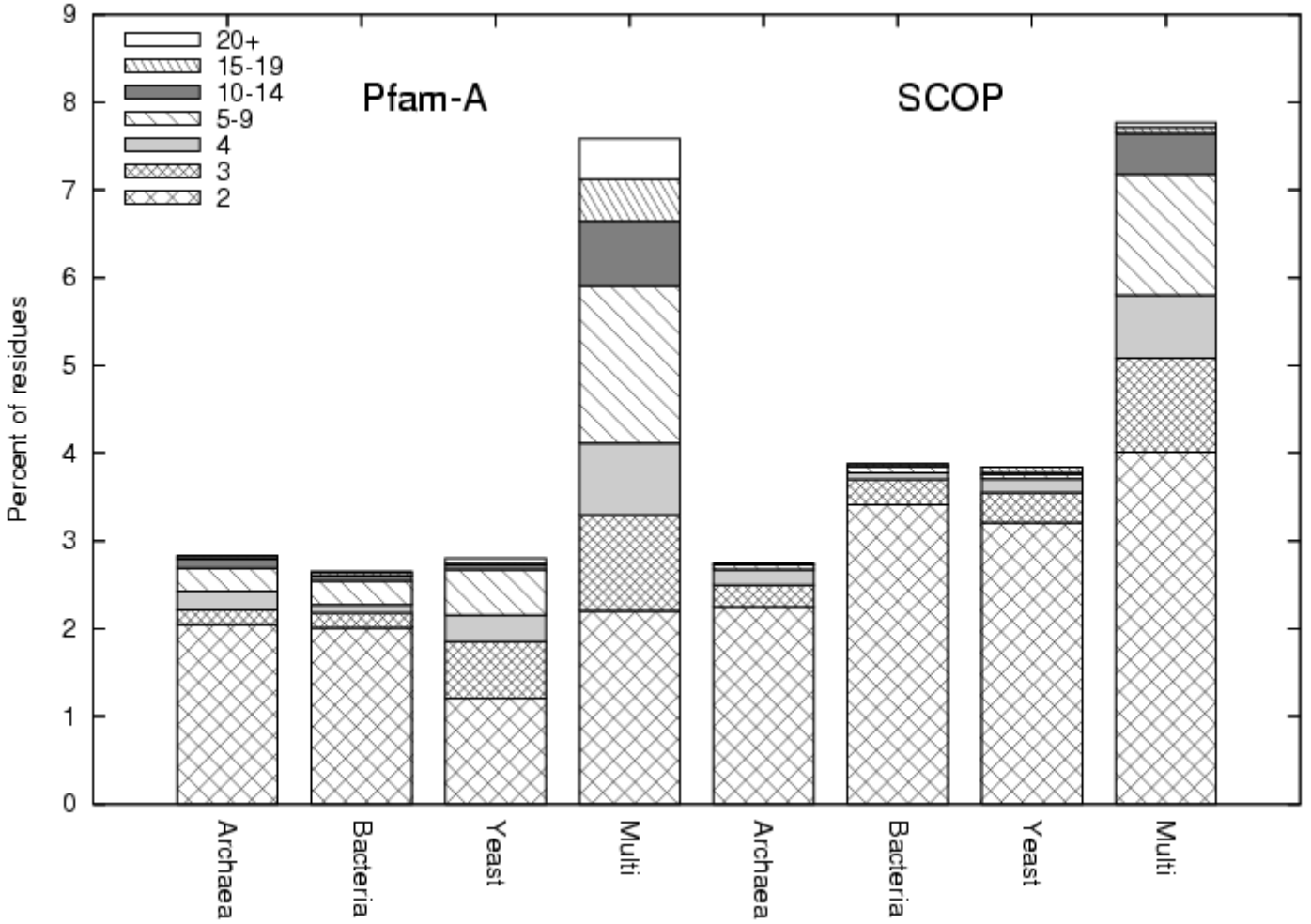
Dissordered regions are common in the unassigned regions



Multidomain proteins



Repetitions



Summary and outlook

- Structurel helps to organize
  - Helps functional assignment and evolutionary analysis
- Practical use for homology modelling
- Many orphan domains remains
  - Are they distant homologs or real orphans ?