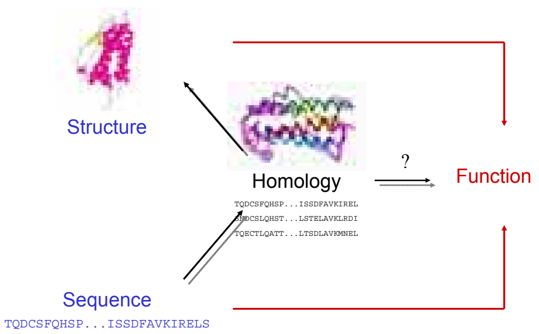
Ancient Protein Evolutionary Relationships Inferred from Structure

Emma E. Hill &
Steven E. Brenner

University of California, Berkeley

IPAM - Structural Proteomics
10 May 2004

Sequence/Structure/Function



Structure

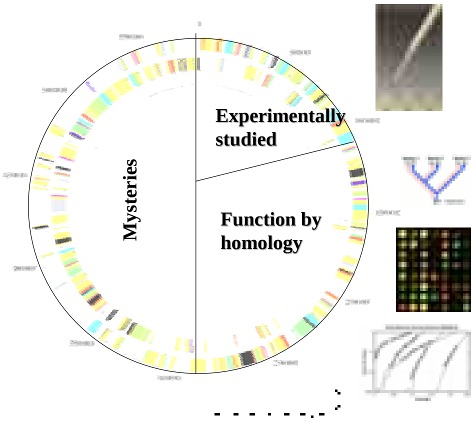
Homology

TQDCSFQHSP...ISSDFAVKIREL
SNDCSLQHST...LSTELAVKLRDI
TQECTLQATT...LTSDLAVKMNEL

?

Function

Sequence
TQDCSFQHSP...ISSDFAVKIRELS



Experimentally studied

Function by homology

Mysteries

Genome Annotation Quality

- What is the quality of genome annotation?
- Quality of sequence well known
- Quality of gene prediction at least roughly understood
- Functional accuracy of 99.5% claimed…
  … but not tested experimentally
- *We rely upon functional assignments for biological interpretation*

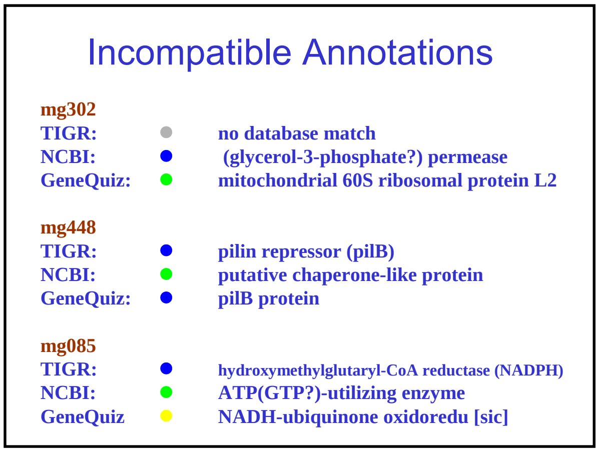
The Annotation of *M. genitalium*

1. TIGR sequences genome and makes initial annotation
2. GeneQuiz consortium automatically annotates
3. Eugene Koonin et al (NCBI) manually make annotations
4. GeneQuiz consortium automatically re-annotates
5. Updates
   - Several groups make automated structural annotations
   - TIGR makes updates to annotation, including new genefinding

Different groups use similar methods and operated sequentially, reviewing each others' results

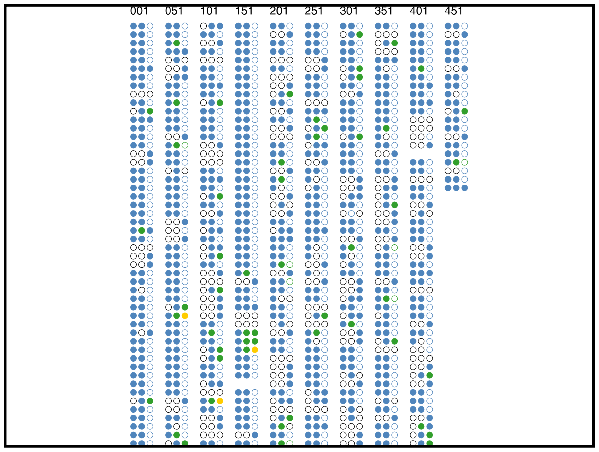
Compatible Annotations

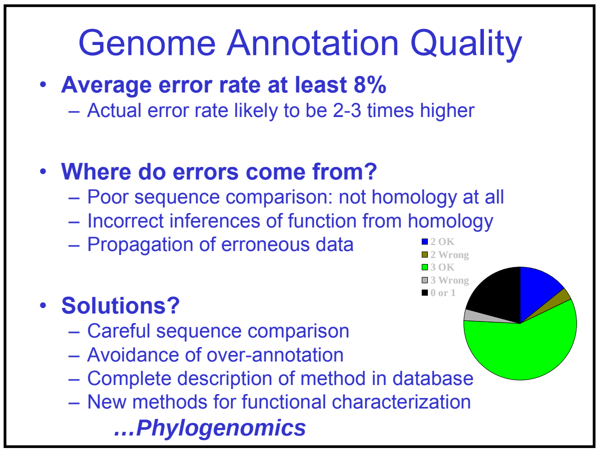
mg463		
TIGR:	●	**high level kasgamycin resistance (ksgA)**
NCBI:	●	**rRNA (adenosine-N6, N6-)-dimethyltransferase (ksgA)**
GeneQuiz:	●	**Dimethyladenosine transfe [sic]**
mg010		
TIGR:	●	**DNA primase (dnaE)**
NCBI:	●	**DNA primase (truncated version) (DnaGp)**
GeneQuiz:	●	**DNA primase (EC 2.7.7.-)**
mg225		
TIGR:	●	**hypothetical protein**
NCBI:	●	**amino acid permease**
GeneQuiz:	●	**histidine permease**



Incompatible Annotations

mg302		
TIGR:	●	**no database match**
NCBI:	●	**(glycerol-3-phosphate?) permease**
GeneQuiz:	●	**mitochondrial 60S ribosomal protein L2**
mg448		
TIGR:	●	**pilin repressor (pilB)**
NCBI:	●	**putative chaperone-like protein**
GeneQuiz:	●	**pilB protein**
mg085		
TIGR:	●	**hydroxymethylglutaryl-CoA reductase (NADPH)**
NCBI:	●	**ATP(GTP?)-utilizing enzyme**
GeneQuiz	●	**NADH-ubiquinone oxidoredu [sic]**

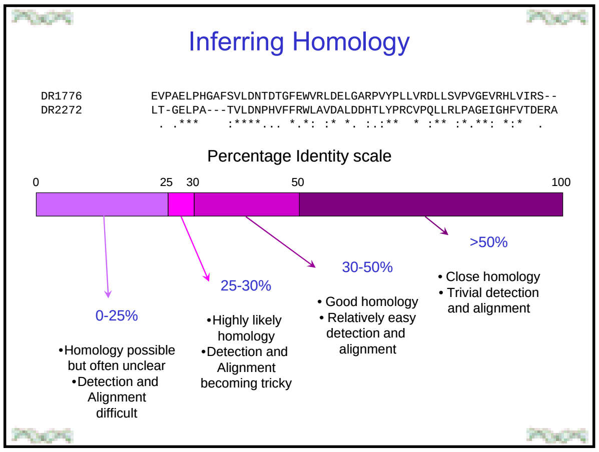




Genome Annotation Quality

- **Average error rate at least 8%**
  - Actual error rate likely to be 2-3 times higher
- **Where do errors come from?**
  - Poor sequence comparison: not homology at all
  - Incorrect inferences of function from homology
  - Propagation of erroneous data
- **Solutions?**
  - Careful sequence comparison
  - Avoidance of over-annotation
  - Complete description of method in database
  - New methods for functional characterization

    …Phylogenomics



Inferring Homology

```
DR1776          EVPAELPHGAFSVLDNTDTGFEWVRLDELGARPVYPLLVRDLLSVPVGEVRHLVIRS--
DR2272          LT-GELPA---TVLDNPHVFFRWLAVDALDDHTLYPRCVPQLLRLPAGEIGHFVTDERA
                 . .***    :****... *.*: :* *. :.:**  * :** :*.**: *:*  .
```

Percentage Identity scale

0-25%

•Homology possible but often unclear
•Detection and Alignment difficult

25-30%

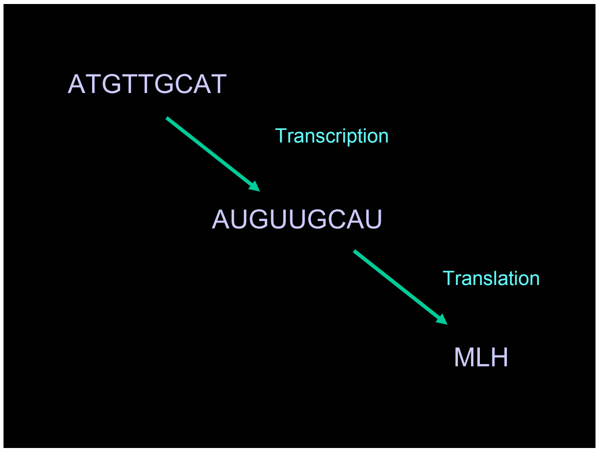
•Highly likely homology
•Detection and Alignment becoming tricky

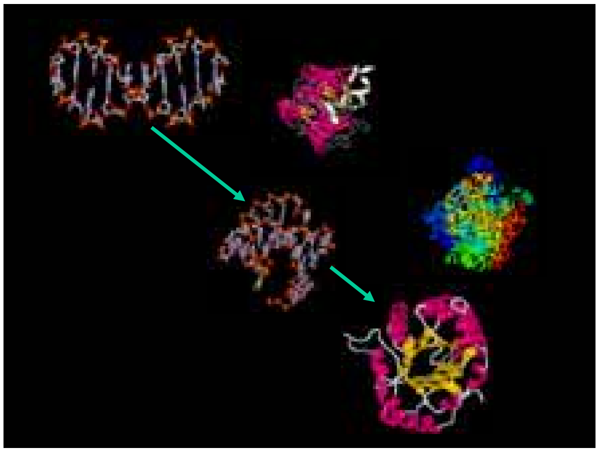
30-50%

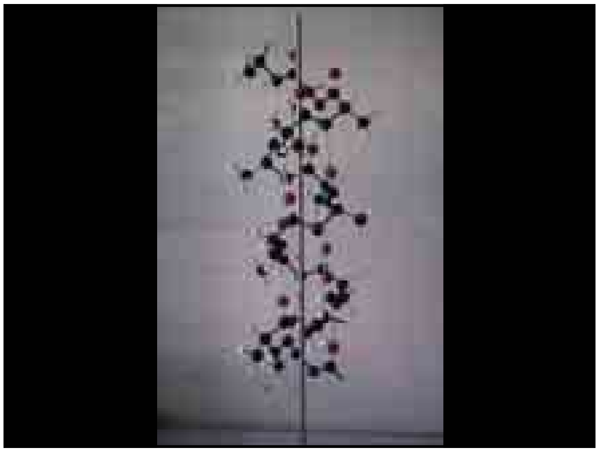
• Good homology
• Relatively easy detection and alignment

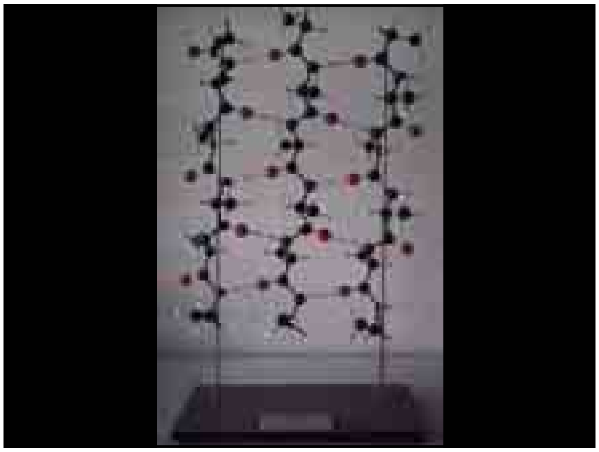
>50%

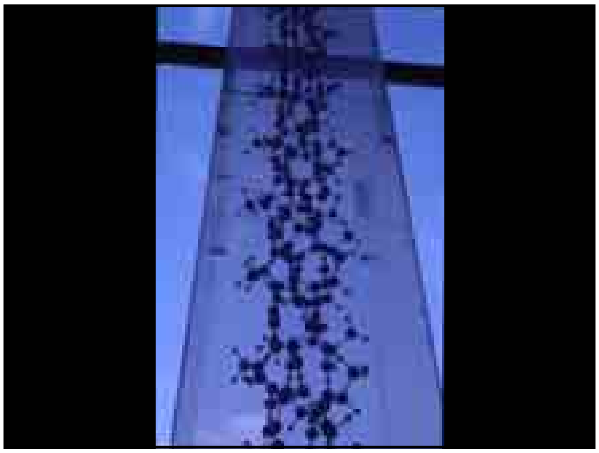
• Close homology
• Trivial detection and alignment

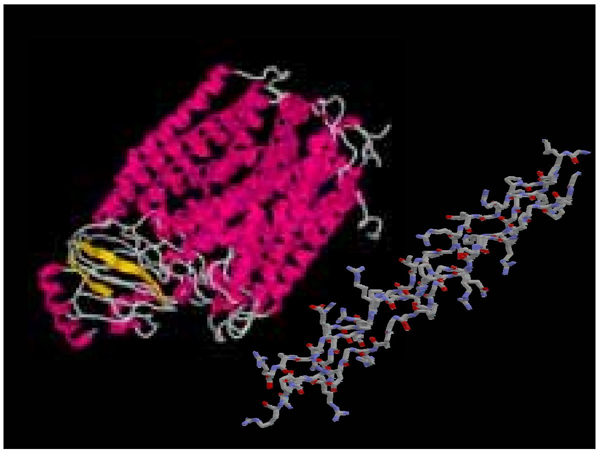


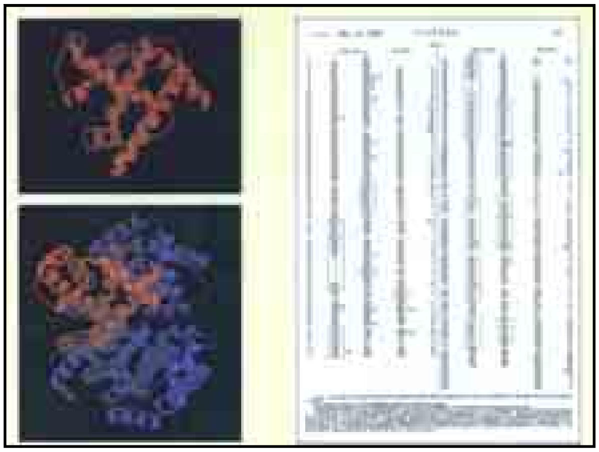




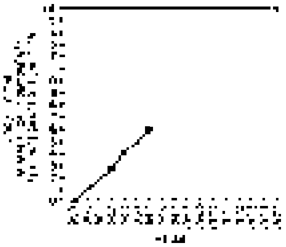








Moore’s Law

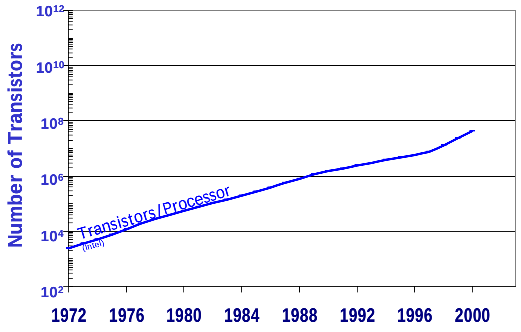


Abstract:
“With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip”
Actually: 5,000 in 1974
Intel 8080

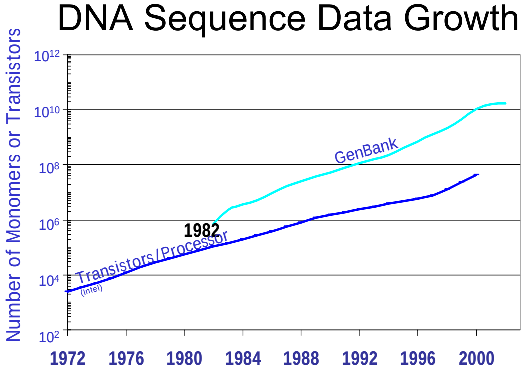
Further:
by 2003, there would be 2^{32} (~17 billion; ~10^{10}) components on each integrated circuit
Actually: 4 x 10^8 ($2^{28.6}$)
Intel Madison

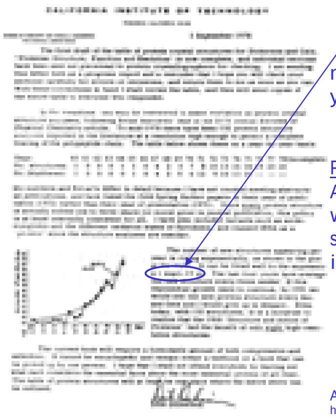
MOORE, G. E. *Cramming More Components Onto Integrated Circuits.* **Electronics** 38, 8 (April 19 **1965**), 114--117.

Moore’s Law



DNA Sequence Data Growth





CALIFORNIA INSTITUTE OF TECHNOLOGY

n = exp(0.19 y)

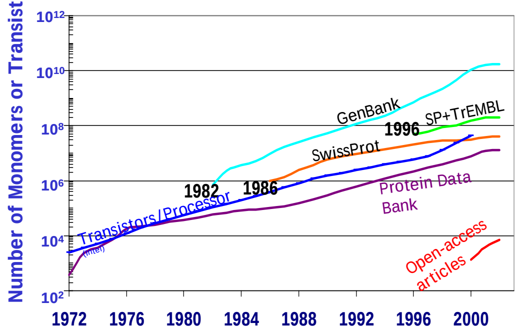
n = new structures/year
y = years since 1960

Prediction:
At the end of 2001, there would be **13,941** crystal structure entries available in the PDB

14,000 crystal structures were actually available!

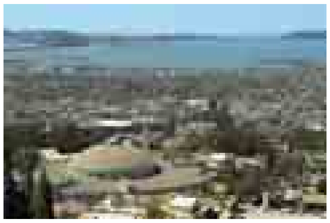
Arthur Arnone
http://www.rcsb.org/pdb/pdb_news2002.html

Database growth

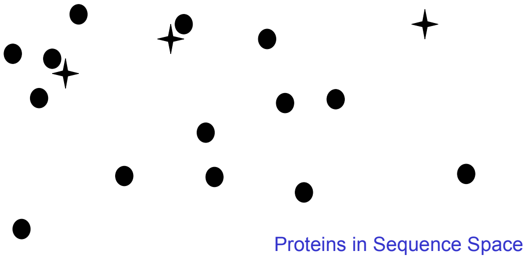


Structural Genomics

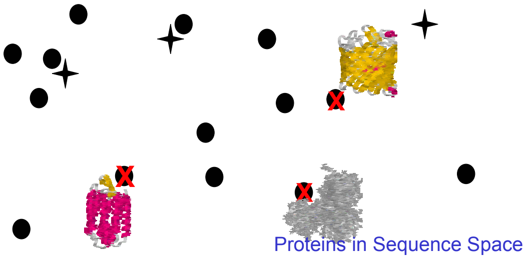
- Provide a 3D structure or quality model for every tractable biomacromolecule
- LBNL Berkeley Structural Genomics Center is one of 9 NIH-funded pilot centers
- Comparable worldwide efforts, esp Japan
- Experimental & computational effort



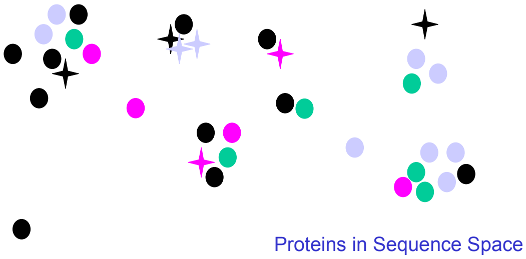
Practical Target Selection



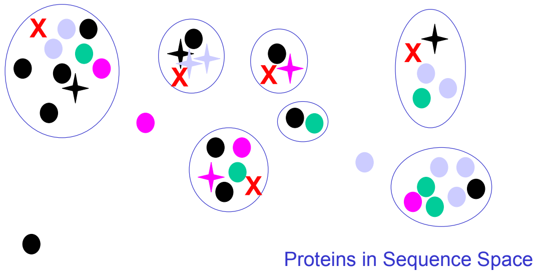
Practical Target Selection



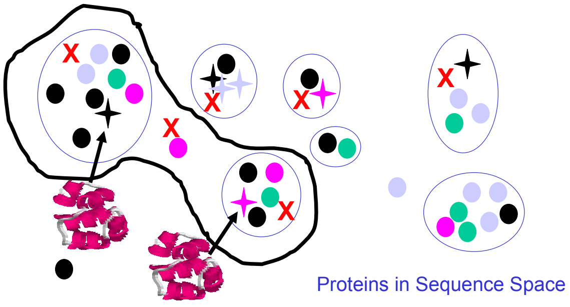
Practical Target Selection



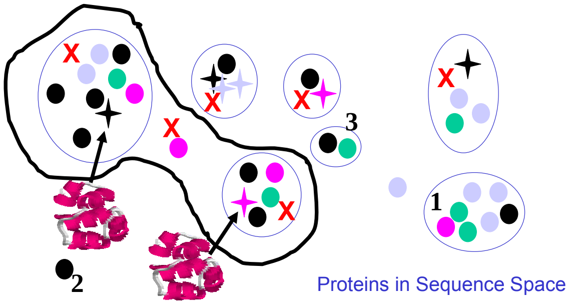
Practical Target Selection



Practical Target Selection

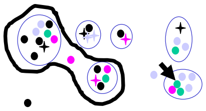


Practical Target Selection

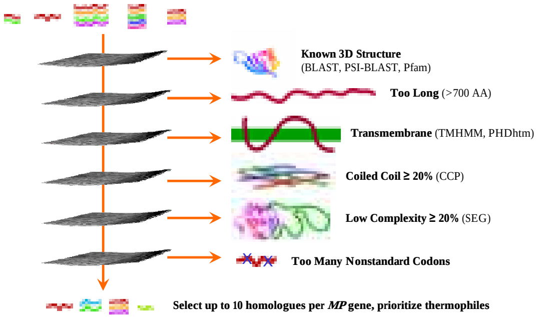


Protein Selection

- All proteins in *Mycoplasma* are potential targets
- Potentially easy to characterize members
  - Cloning (accessible organisms)
  - Expression (UGA/Trp)
  - Purification (thermo) & Crystallization



BSGC Target Selection (Current)



Known 3D Structure (BLAST, PSI-BLAST, Pfam)

Too Long (>700 AA)

Transmembrane (TMHMM, PHDhtm)

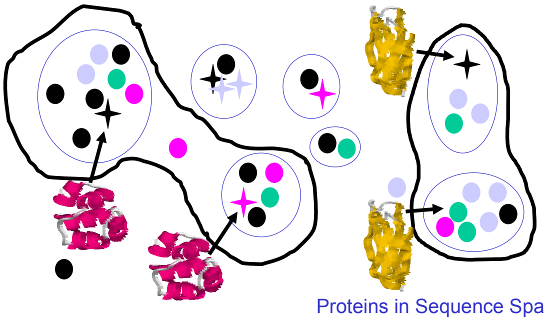
Coiled Coil ≥ 20% (CCP)

Low Complexity ≥ 20% (SEG)

Too Many Nonstandard Codons

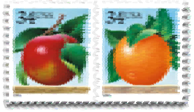
Select up to 10 homologues per *MP* gene, prioritize thermophiles

Practical Structural Genomics



Why Classify

"Physics or Stamp Collecting"
--Ernest Rutherford

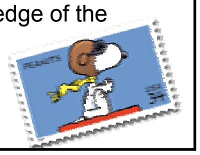


General trends provide insight into underlying principles

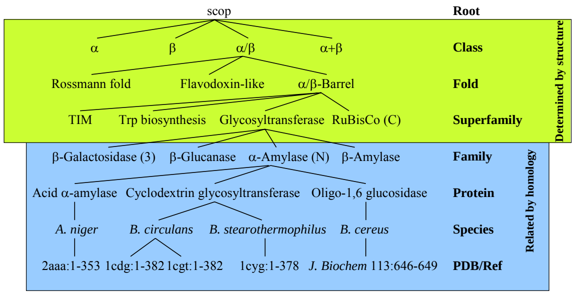
The Jenny

Unusual features only become apparent with knowledge of the principles

Assist predictions



SCOP Sample Hierarchy



scop — Root

α β α/β α+β — Class

Rossmann fold Flavodoxin-like α/β-Barrel — Fold

TIM Trp biosynthesis Glycosyltransferase RuBisCo (C) — Superfamily

β-Galactosidase (3) β-Glucanase α-Amylase (N) β-Amylase — Family

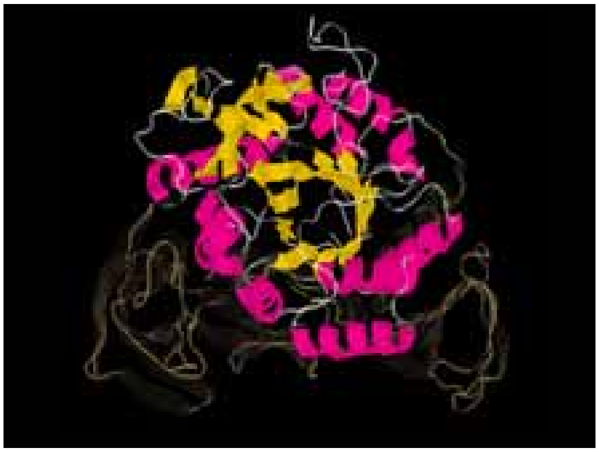
Acid α-amylase Cyclodextrin glycosyltransferase Oligo-1,6 glucosidase — Protein

A. niger *B. circulans* *B. stearothermophilus* *B. cereus* — Species

2aaa:1-353 1cdg:1-382 1cgt:1-382 1cyg:1-378 *J. Biochem* 113:646-649 — PDB/Ref

Determined by structure

Related by homology



SCOP Sample Hierarchy

scop — Root

α β α/β α+β — Class

Rossmann fold Flavodoxin-like α/β-Barrel — Fold

TIM Trp biosynthesis Glycosyltransferase RuBisCo (C) — Superfamily

β-Galactosidase (3) β-Glucanase α-Amylase (N) β-Amylase — Family

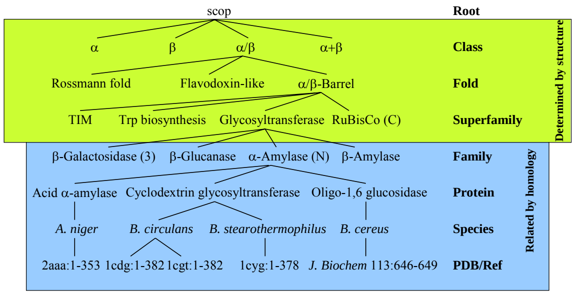
Acid α-amylase Cyclodextrin glycosyltransferase Oligo-1,6 glucosidase — Protein

A. niger *B. circulans* *B. stearothermophilus* *B. cereus* — Species

2aaa:1-353 1cdg:1-382 1cgt:1-382 1cyg:1-378 *J. Biochem* 113:646-649 — PDB/Ref

Determined by structure

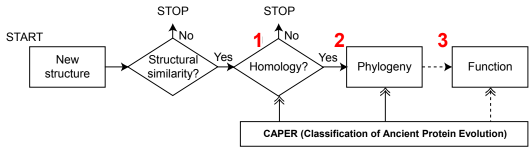
Related by homology



Making Structure Classification Consistent and Automated

1. Automatically determine homology from these features for proteins of known structure
2. Calculate phylogenies for protein superfamilies
3. Apply phylogenomic techniques to predict protein function

Identifying & learning on protein structure features



START

New structure

Structural similarity?

STOP

No

Yes

Homology?

STOP

No

Yes

1

2

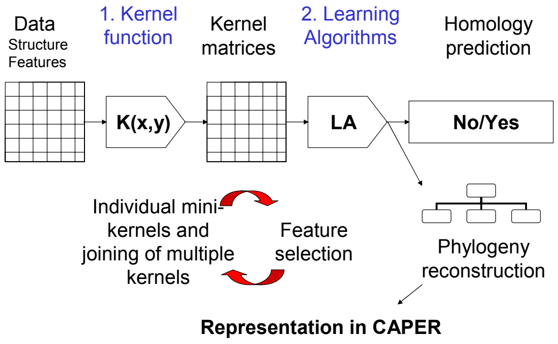
3

Phylogeny

Function

CAPER (Classification of Ancient Protein Evolution)

The Plan



Data Structure Features

1. Kernel function

Kernel matrices

2. Learning Algorithms

Homology prediction

K(x,y)

LA

No/Yes

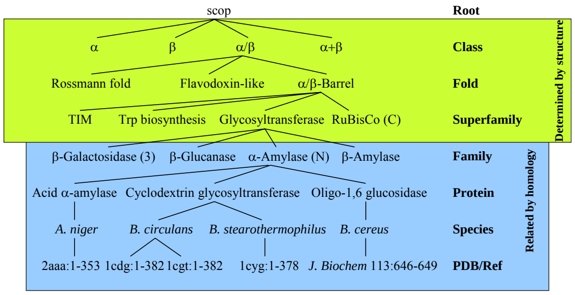
Individual mini-kernels and joining of multiple kernels

Feature selection

Phylogeny reconstruction

Representation in CAPER

Validation on the SCOP database



scop — Root

α β α/β α+β — Class

Rossmann fold Flavodoxin-like α/β-Barrel — Fold

TIM Trp biosynthesis Glycosyltransferase RuBisCo (C) — Superfamily

β-Galactosidase (3) β-Glucanase α-Amylase (N) β-Amylase — Family

Acid α-amylase Cyclodextrin glycosyltransferase Oligo-1,6 glucosidase — Protein

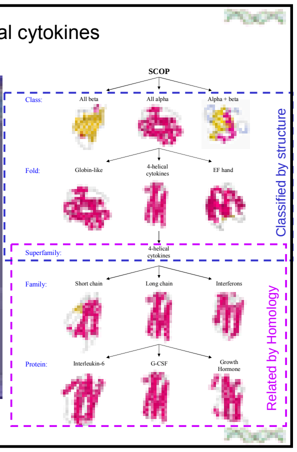
A. niger *B. circulans* *B. stearothermophilus* *B. cereus* — Species

2aaa:1-353 1cdg:1-382 1cgt:1-382 1cyg:1-378 *J. Biochem* 113:646-649 — PDB/Ref

Determined by structure

Related by homology

Four-helical cytokines



http://scop.berkeley.edu

Low Sequence Identity but Definite Homology

Long Chain

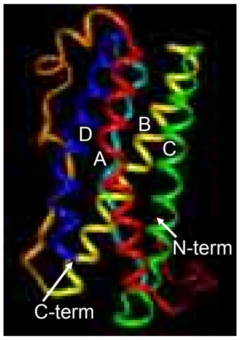
seed	LIF d1lki	Somatotropin d1hgu	CNTF d1cnt1	Leptin d1ax8
d1rh				
d1al				
d1lk				
d1hg				
d1cr				
d1ax				

Short Chain

	Erythropoietin d1eera_ 2.1	GM-CSF d2gmfa_ 2.2	IL-4 d1iara_ 2.3	IL-5 d1hula_ 2.4	M-CSF d1hmca_ 2.5	Flt3 d1etea_ 2.6	IL-2 d3inkc_ 2.7	IL-3 d1jli__ 2.8
d1eera_ 2.1	2.80E-71 166	**None**	**None**	**None**	**None**	**None**	**None**	**None**
d2gmfa_ 2.2	**None**	7.90E-65 121	**None**	**None**	**None**	**None**	**None**	**None**
d1iara_ 2.3	**None**	**None**	2.20E-54 129	**None**	**None**	**None**	**None**	**None**
d1hula_ 2.4	**None**	**None**	**None**	1.00E-51 108	**None**	**None**	**None**	**None**
d1hmca_ 2.5	**None**	**None**	**None**	**None**	7.40E-82 145	**None**	**None**	**None**
d1etea_ 2.6	**None**	**None**	**None**	**None**	**None**	1.40E-80 134	**None**	**None**
d3inkc_ 2.7	**None**	**None**	**None**	**None**	**None**	**None**	7.60E-64 128	E-value **None** Length of match
d1jli__ 2.8	**None**	**None**	**None**	**None**	**None**	**None**	**None**	2.50E-43 111

- Long Chain Family - 3 inter-protein hits.
- Short Chain Family - 0 inter-protein matches.

Why are we certain they are homologous?



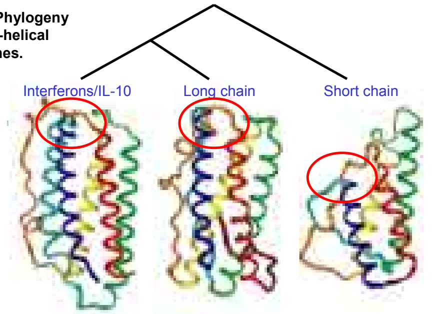
Unique topology (up-up-down-down)

Restricted range of functions

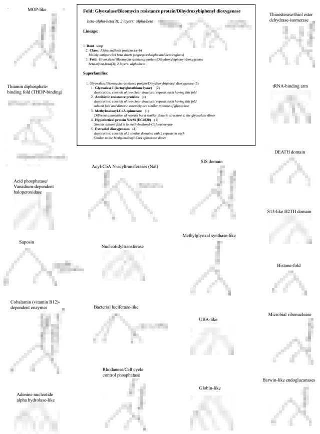
Homologous receptors

Conserved exon/intron boundaries

Family-level Phylogeny of the four-helical cytokines.



Family:	Interferons/IL-10	Long chain	Short chain
Protein (PDBID):	Ifn-β (1au1)	LIF (1lki)	IL-4 (1rcb)
Reference:	Karpusas *et al*., 1997	Robinson *et al*., 1994	Wlodawer *et al*., 1992
Av. Chain Length:	180	180	140
Av. Helix Length:	20-30	20-30	10-20
AB Helix Linker: **(w.r.t Helix D)**	Outside	Outside	Inside
Other:	CD linker forms 5th helix	Small Helix in AB linker	Small 2 stranded β-sheet


MOP-like
Thioesterase/thiol ester dehydrase-isomerase
Thiamin diphosphate-binding fold (THDP-binding)
tRNA-binding arm
DEATH domain
Acyl-CoA N-acyltransferases (Nat)
SIS domain
Acid phosphatase/ Vanadium-dependent haloperoxidase
S13-like H2TH domain
Saposin
Nucleotidyltransferase
Methylglyoxal synthase-like
Histone-fold
Cobalamin (vitamin B12)-dependent enzymes
Bacterial luciferase-like
UBA-like
Microbial ribonuclease
Rhodanese/Cell cycle control phosphatase
Barwin-like endoglucanases
Globin-like
Adenine nucleotide alpha hydrolase-like

Fold: Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase

beta-alpha-beta(3); 2 layers: alpha/beta

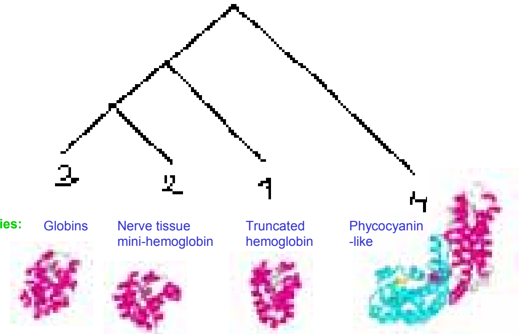
Lineage:

1. **Root** : scop
   2. **Class:** Alpha and beta proteins (a+b)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
   3. **Fold:** Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase
   beta-alpha-beta(3); 2 layers: alpha/beta

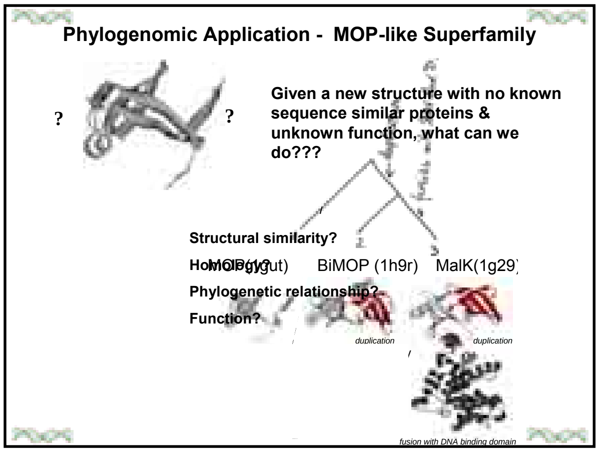
Superfamilies:

1. Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase (5)
   1. **Glyoxalase I (lactoylglutathione lyase)** (2)
   duplication: consists of two clear structural repeats each having this fold
   2. **Antibiotic resistance proteins** (4)
   duplication: consists of two clear structural repeats each having this fold
   subunit fold and dimeric assembly are similar to those of glyoxalase
   3. **Methylmalonyl-CoA epimerase** (1)
   Different assosiation of repeats but a similar dimeric structure to the glyoxalase dimer
   4. **Hypothetical protein YecM (EC4020)** (1)
   Similar subunit fold is to methylmalonyl-CoA epimerase
   5. **Extradiol dioxygenases** (4)
   duplication: consists of 2 similar domains with 2 repeats in each
   Similar to the Methylmalonyl-CoA epimerase dimer

The Globin-like superfamily



Families:	Globins	Nerve tissue mini-hemoglobin	Truncated hemoglobin	Phycocyanin -like
	prototype	lacks first helix (but more similar to globins than truncated hemoglobins.)	lacks first helix	Oligomers of two different homologous subunits each subunit has two additional N-terminus helices binds a chromophore



Phylogenomic Application - MOP-like Superfamily

?

?

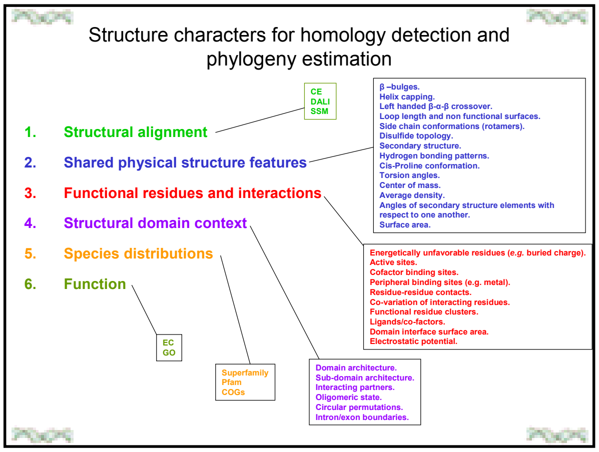
Given a new structure with no known sequence similar proteins & unknown function, what can we do???

Structural similarity?

Homology?

Phylogenetic relationship?

Function?



Structure characters for homology detection and phylogeny estimation

1. Structural alignment
2. Shared physical structure features
3. Functional residues and interactions
4. Structural domain context
5. Species distributions
6. Function

CE
DALI
SSM

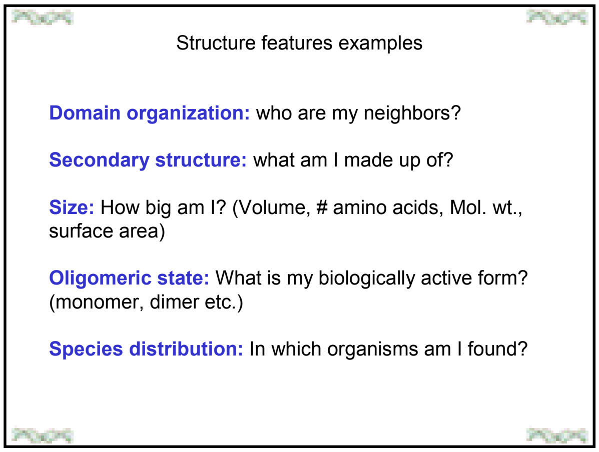
β –bulges.
Helix capping.
Left handed β-α-β crossover.
Loop length and non functional surfaces.
Side chain conformations (rotamers).
Disulfide topology.
Secondary structure.
Hydrogen bonding patterns.
Cis-Proline conformation.
Torsion angles.
Center of mass.
Average density.
Angles of secondary structure elements with respect to one another.
Surface area.

Energetically unfavorable residues (*e.g.* buried charge).
Active sites.
Cofactor binding sites.
Peripheral binding sites (e.g. metal).
Residue-residue contacts.
Co-variation of interacting residues.
Functional residue clusters.
Ligands/co-factors.
Domain interface surface area.
Electrostatic potential.

Domain architecture.
Sub-domain architecture.
Interacting partners.
Oligomeric state.
Circular permutations.
Intron/exon boundaries.

Superfamily
Pfam
COGs

EC
GO



Structure features examples

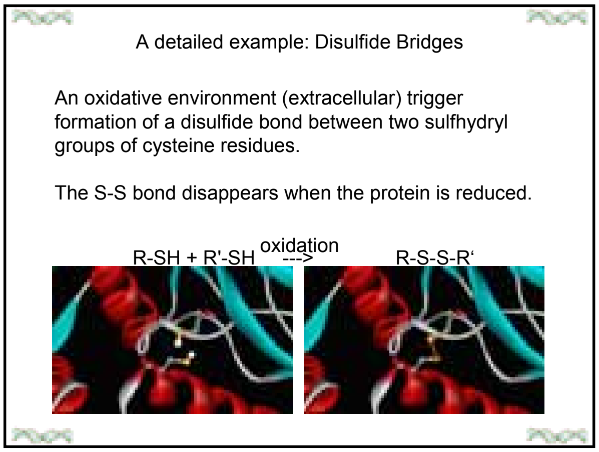
Domain organization: who are my neighbors?

Secondary structure: what am I made up of?

Size: How big am I? (Volume, # amino acids, Mol. wt., surface area)

Oligomeric state: What is my biologically active form? (monomer, dimer etc.)

Species distribution: In which organisms am I found?

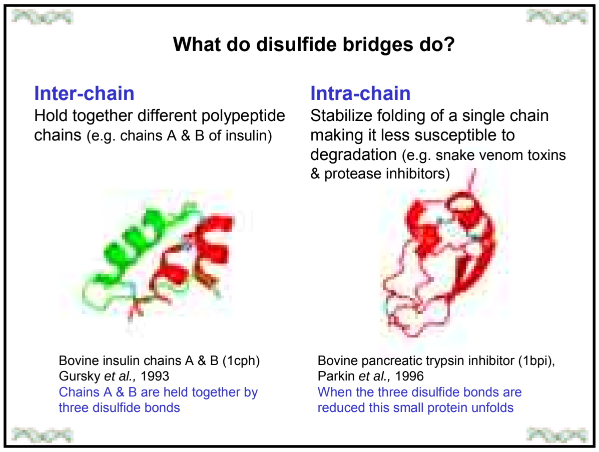


A detailed example: Disulfide Bridges

An oxidative environment (extracellular) trigger formation of a disulfide bond between two sulfhydryl groups of cysteine residues.

The S-S bond disappears when the protein is reduced.

$$\text{R-SH + R'-SH} \xrightarrow{\text{oxidation}} \text{R-S-S-R'}$$



What do disulfide bridges do?

Inter-chain

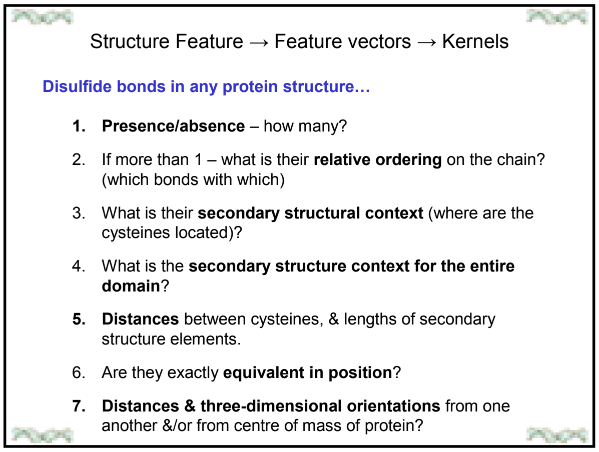
Hold together different polypeptide chains (e.g. chains A & B of insulin)

Intra-chain

Stabilize folding of a single chain making it less susceptible to degradation (e.g. snake venom toxins & protease inhibitors)

Bovine insulin chains A & B (1cph)
Gursky *et al.*, 1993
Chains A & B are held together by three disulfide bonds

Bovine pancreatic trypsin inhibitor (1bpi),
Parkin *et al.*, 1996
When the three disulfide bonds are reduced this small protein unfolds



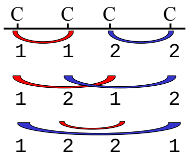
Structure Feature → Feature vectors → Kernels

Disulfide bonds in any protein structure…

1. **Presence/absence** – how many?
2. If more than 1 – what is their **relative ordering** on the chain? (which bonds with which)
3. What is their **secondary structural context** (where are the cysteines located)?
4. What is the **secondary structure context for the entire domain**?
5. **Distances** between cysteines, & lengths of secondary structure elements.
6. Are they exactly **equivalent in position**?
7. **Distances & three-dimensional orientations** from one another &/or from centre of mass of protein?

Relative orientations (# and their ordering)

For a protein with two S-S bonds there are
3 possible orientations



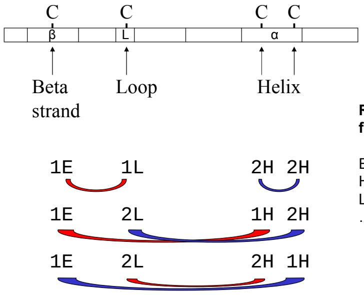
# DS	# C	# poss orders
1	2	1
2	4	3
3	6	15
n	**2n**	**$X_n = (2n-1)X_{n-1}$**

Consideration of non-bonded Cysteines

n	**2n + c**	**$X_n = [(2n-1)X_{n-1}](2n+1)^c$**

(0) 1(0)2(0)2 (0)1(0)

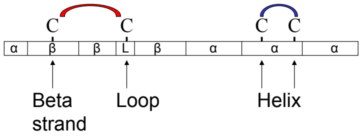
Ordering & Secondary structure context



Representative alphabet for secondary structure

E – extended (beta-strand)
H – Helix
L – Loop
…

Ordering & More secondary structure context



H_1,$1E_1$,E_2, $1L_1$,E_2, H_2, $22H_3$, H_4

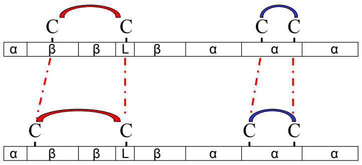
Next step is to add…

- # residues in each secondary structure element
- actual position of C within the secondary structure element

Structural equivalence???

If two proteins have the same signature:
H_1,$1E_1$,E_2, $1L_1$,E_2, H_2, $22H_3$, H_4

Are their disulfide bridges at equivalent positions???
Do they have to be the same to be homologous?



Mini-kernels of increasing complexity

Feature representation	Protein A SCOP classification	Protein B SCOP classification	…
# DS bonds	n	n	
# & orientation	11022	1212	
#, orientation & SSE	1H1E0L2H2H	1H2H1E2L	
#, orientation & all SSE	$H_1 1E_1 E_2 1L_1 E_2 H_2 22H_3$	$H_1 1E_1 E_2 1L_1 E_2 H_2 22H_3$	
Above + specific lengths	$(8)H_1(6)1E_1(9)E_2(3)1L_1$	$(8)H_1(6)1E_1(9)E_2(3)1L_1$	
Alignment	Dynamic programming	Dynamic programming	

Will we need alignment information?

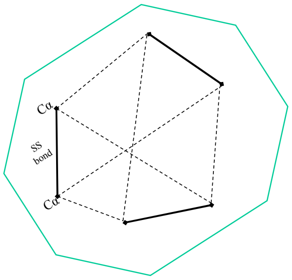
A more three-dimensional approach… (this could remove the need to make alignments?)

Fix first Cα of first SS bond (relative to sequence) as planar, (x=0 for both Cα atoms) with x,y,z of first Cys as (0,0,0)

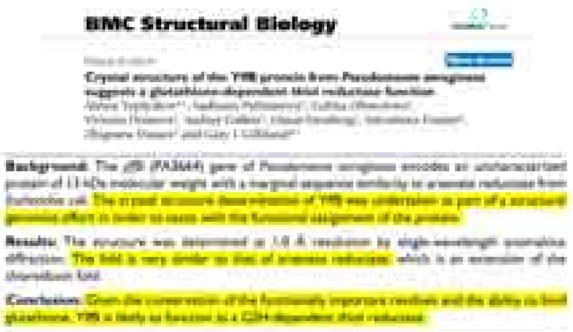
Give relative positions of other Cα atoms from other disulfide Cysteines

Calculate euclidean distances

Calculate RMSDs based on fitting the disulfide bonds



Structural similarity → Function prediction



BMC Structural Biology

Crystal structure of the YffB protein from *Pseudomonas aeruginosa* suggests a glutathione-dependent thiol reductase function

[illegible]

Background: The *yffB* (PA3664) gene of *Pseudomonas aeruginosa* encodes an uncharacterized protein of 13 kDa molecular weight with a marginal sequence similarity to arsenate reductase from *Escherichia coli*. The crystal structure determination of YffB was undertaken as part of a structural genomics effort in order to assist with the functional assignment of the protein.

Results: The structure was determined at 1.0 Å resolution by single-wavelength anomalous diffraction. The fold is very similar to that of arsenate reductase, which is an extension of the thioredoxin fold.

Conclusion: Given the conservation of the functionally important residues and the ability to bind glutathione, YffB is likely to function as a GSH-dependent thiol reductase.

Phylogeny → Understanding protein structure evolution

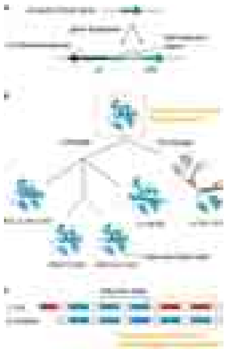
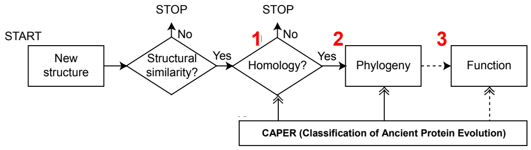


Fig. 1 from Newlove *et al.*, Structure 2004

Overview of our research



1. **A prediction facility to calculate homology from structure**
2. **Phylogenetic reconstructions for homologous proteins of known structure**
3. **A method to predict function based on phylogenetic location**

Applications

This has applications for…

Function prediction

Reconstructing phylogenetic relationships

Relating phylogenetic lineage to protein structure evolution
→ Understanding how changes have occurred in protein structures
→ Resolving questions about the ancestral form of proteins

Elucidating **which structure features are important** in which superfamilies
→ Relating these features back to the proteins in question

Acknowledgements

Steven E. Brenner (UCB)
John-Marc Chandonia
Marcin Joachimiak
Donna Hendrix
Gavin Crooks
Ed Green

Michael Jordan (UCB)
Barbara Engelhardt
Gert Lanckriet
Neil Lawrence
Matthias Seeger
Guillaume Obozinski

Alexey G. Murzin (CPE Cambridge, UK)
Cyrus Chothia (MRC-LMB, UK)
Michael Levitt (Stanford)
Tim Hubbard (Sanger Centre, UK)
Loredana Lo Conte (MRC-LMB, UK)
Veronica Morea (La Sapienza, Rome)
Dick Karp (UCB)