Structure-Based Prediction of Protein Function

Thomas Funkhouser
Princeton University
CS597A, Fall 2005

Outline

Protein structure databases
- Repositories
- Classifications

Protein function databases
- Gene Ontology (GO)
- Enzyme Commission (EC)

Sequence → Structure → Function
- Sequence alignment
- Structure alignment
- Sequence motifs
- Structure motifs

Outline

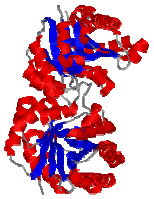
Protein structure databases
- ØRepositories
- Classifications

Protein function databases
- Gene Ontology (GO)
- Enzyme Commission (EC)

Sequence → Structure → Function
- Sequence alignment
- Structure alignment
- Sequence motifs
- Structure motifs

Protein Structure Databases

Repositories:
- Primary ← amino acid sequence
- Secondary ← local fold pattern of small subsequence
- Tertiary ← fold of entire protein chain
- Quaternary ← complex of multiple chains



1tim
[Jena]

Protein Structure Databases

Repositories:
- Primary ← UniProt
- Secondary ← DSSP
- Tertiary ← PDB
- Quaternary ← PQS

Protein Structure Databases

Repositories:
- ØPrimary ← UniProt
- Secondary ← DSSP
- Tertiary ← PDB
- Quaternary ← PQS

Chain 1GSA:_
Compound Glutathione Synthetase
Type Protein
Molecular Weight 35547
Number of Residues 316

```
  1 MIKLGIVMDP IANINIKKDS SFAMLLEAQR RGYELHYMEM GDLYLINGEA
 51 RAHTRTLNVK QNYEEWFSFV GEQDLPLADL DVILMRKDPP FDTEFIYATY
101 ILERAEEKGT LIVNKPQSLR DCNEKLFTAW FSDLTPETLV TRNKAQLKAF
151 WEKHSDIILK PLDGMGGASI FRVKEGDPNL GVIAETLTEH GTRYCMAQNY
201 LPAIKDGDKR VLVVDGEPVP YCLARIPQGG ETRGNLAAGG RGEPRPLTES
251 DWKIARQIGP TLKEKGLIFV GLDIIGDRLT EINVTSPTCI REIEAEFPVS
301 ITGMLMDAIE ARLQQQ
```

http://www.uniprot.org/ [Apweiler04]

Protein Structure Databases

Repositories:

- Primary ← UniProt
- ØSecondary ← DSSP
- Tertiary ← PDB
- Quaternary ← PQS

Chain 1GSA:_
Compound Glutathione Synthetase
Type Protein
Molecular Weight 35547
Number of Residues 316
Number of Alpha 9 Content of Alpha 27.22
Number of Beta 19 Content of Beta 28.16

```
  1 MIKLGIVMDP IANINIKKDS SFAMLLEAQR RGYELHYMEM GDLYLINGEA
      EEEEE S  GGGTTTTTTH HHHHHHHHHH HT EEEEE G GGEEEETTEE
 51 RAHTRTLNVK QNYEEWFSFV GEQDLPLADL DVILMRKDPP FDTEFIYATY
    EEEEEEEEE  S SS  EEE   EEEEEGGGS SEEEE        HHHHHHHH
101 ILERAEEKGT LIVNKPQSLR DCNEKLFTAW FSDLTPETLV TRNKAQLKAF
    HHHHHHHTT  EEES HHHHH HTTTTGGGGG GTTTB  EEE ES HHHHHHH
151 WEKHSDIILK PLDGMGGASI FRVKEGDPNL GVIAETLTEH GTRYCMAQNY
    HHHHSSEEEE  SS  TTTT  EEE TTTTTH HHHHHHHTTT TTS EEEEE
201 LPAIKDGDKR VLVVDGEPVP YCLARIPQGG ETRGNLAAGG RGEPRPLTES
      GGGGG EEE EEEETTEE S EEEEEE  SS  S   GGGT  EEEEEE  HH
251 DWKIARQIGP TLKEKGLIFV GLDIIGDRLT EINVTSPTCI REIEAEFPVS
    HHHHHHHHHT HHHHTT  EE EEEEETTEE  EEE SS    H HHHHHHSS
301 ITGMLMDAIE ARLQQQ
    HHHHHHHHHH HHT
```

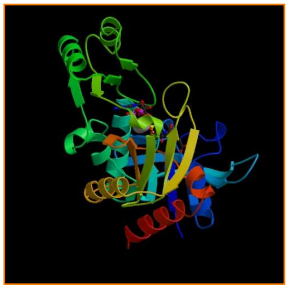
H = helix
B = residue in isolated beta bridge
E = extended beta strand
G = 310 helix
T = hydrogen bonded turn
S = bend

[Kabsch83]

Protein Structure Databases

Repositories:

- Primary ← UniProt
- ØSecondary ← DSSP
- Tertiary ← PDB
- Quaternary ← PQS

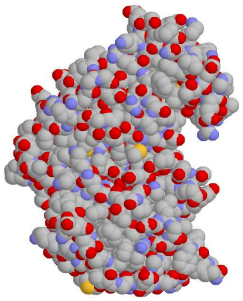


[Kabsch83]

Protein Structure Databases

Repositories:

- Primary ← UniProt
- Secondary ← DSSP
- ØTertiary ← PDB
- Quaternary ← PQS

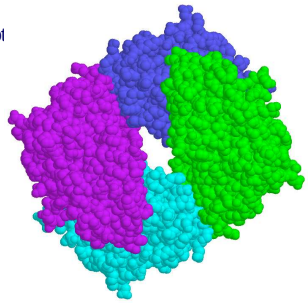


http://www.rcsb.org/pdb/ [Berman00]

Protein Structure Databases

Repositories:

- Primary ← UniProt
- Secondary ← DSSP
- ØTertiary ← PDB
- Quaternary ← PQS

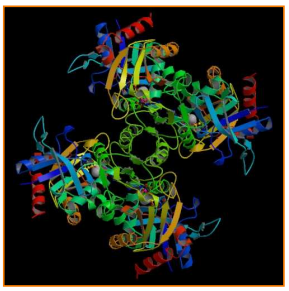


http://pqs.ebi.ac.uk/ [Hendrick98]

Protein Structure Databases

Repositories:

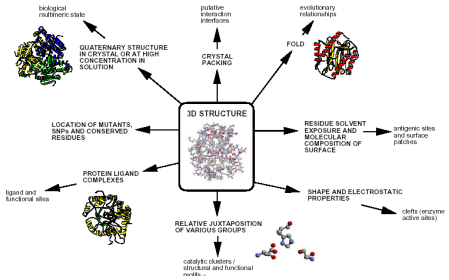
- Primary ← UniProt
- Secondary ← DSSP
- Tertiary ← PDB
- ØQuaternary ← PQS



http://pqs.ebi.ac.uk/ [Hendrick98]

Slide courtesy of Philip Bourne

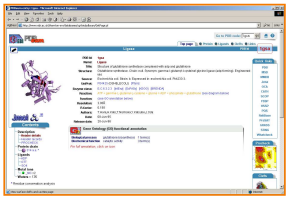
Protein Structure Databases



Protein Structure Databases

Repositories:
- Primary ← UniProt
- Secondary ← DSSP
- Tertiary ← PDB
- Quaternary ← PQS

Summaries:
- ØPDBsum
- Jena
- MSD

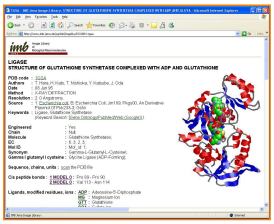


http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/ [Laskowski05]

Protein Structure Databases

Repositories:
- Primary ← UniProt
- Secondary ← DSSP
- Tertiary ← PDB
- Quaternary ← PQS

Summaries:
- PDBsum
- ØJena
- MSD

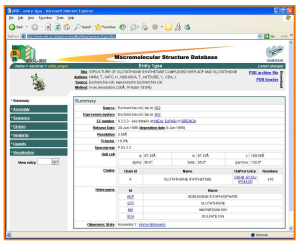


http://www.imb-jena.de/IMAGE.html

Protein Structure Databases

Repositories:
- Primary ← UniProt
- Secondary ← DSSP
- Tertiary ← PDB
- Quaternary ← PQS

Summaries:
- PDBsum
- Jena
- ØMSD



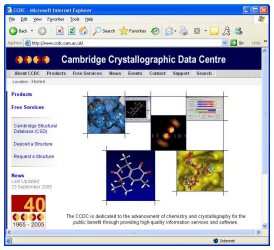
http://www.ebi.ac.uk/msd/ [Velankar05]

Protein Structure Databases

Repositories:
- Primary ← UniProt
- Secondary ← DSSP
- Tertiary ← PDB
- Quaternary ← PQS

Summaries:
- PDBsum
- Jena
- MSD

Small molecules:
- ØCCDC



http://www.ccdc.cam.ac.uk/

Outline

Protein structure databases
- Repositories
- ØClassifications

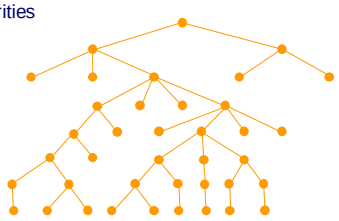
Protein function databases
- Gene Ontology (GO)
- Enzyme Commission (EC)

Sequence → Structure → Function
- Sequence alignment
- Structure alignment
- Sequence motifs
- Structure motifs

Protein Structure Classifications

Clustering
- Fold similarities
- Evolutionary relationships
- Sequence similarities

Examples:
- CATH
- SCOP



Protein Structure Classifications

CATH hierarchy:

- Class
- Architecture
- Topology
- Homology
- S35 (Family)
- S95
- S100

http://cathwww.biochem.ucl.ac.uk/ [Orengo97]

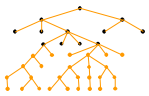
Protein Structure Classifications

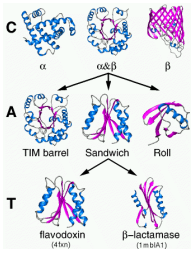
CATH hierarchy:

- Class
- Architecture
- Topology

} Structural Layout

- Homology
- S35 (Family)
- S95
- S100





http://cathwww.biochem.ucl.ac.uk/ [Orengo97]

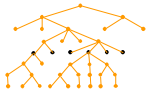
Protein Structure Classifications

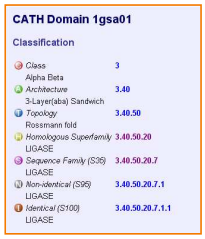
CATH hierarchy:

- Class
- Architecture
- Topology
- Homology

} Evolution

- S35 (Family)
- S95
- S100





http://cathwww.biochem.ucl.ac.uk/ [Orengo97]

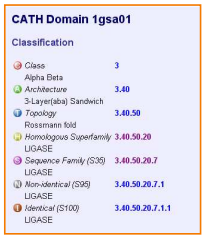
Protein Structure Classifications

CATH hierarchy:

- Class
- Architecture
- Topology
- Homology
- S35 (Family)

} Sequence Identity

- S95
- S100



http://cathwww.biochem.ucl.ac.uk/ [Orengo97]

Protein Structure Classifications

SCOP hierarchy:

- Class
- Fold
- Superfamily
- Family
- Protein Domain
- Species
- PDB

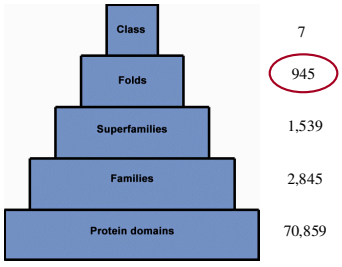
SCOP: 1gsa

1. Root:
   scop
2. Class:
   Alpha and beta proteins (a/b) [51349]
3. Fold:
   PreATP-grasp domain [52439]
4. Superfamily:
   PreATP-grasp domain [52440]
5. Family:
   Prokaryotic glutathione synthetase, N-terminal domain [52457]
6. Protein:
   Prokaryotic glutathione synthetase, N-terminal domain [52458]
7. Species:
   Escherichia coli [52459]

http://scop.mrc-lmb.cam.ac.uk/scop/ [Murzin95]

Protein Structure Classifications

SCOP hierarchy:

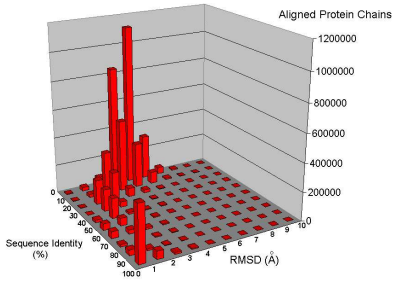


SCOP: Structural Classification of Proteins (1.69 release)

Slide courtesy of Philip Bourne

Protein Structure Classifications

Protein folds are highly redundant

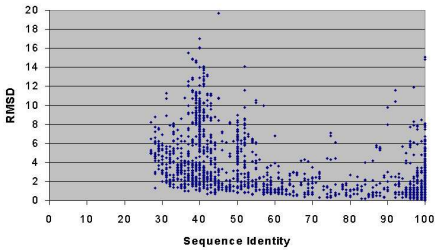


Structure Alignments using CE with z>4.0

Slide courtesy of Philip Bourne

Sequence → Structure → Function

Sequence determine structure, but …



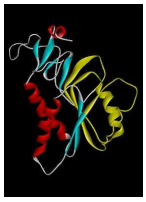
Structure Comparison of 30% of PDBSelect Set

Slide courtesy of Philip Bourne

Sequence → Structure → Function

Similar sequence, different structure & function

1PIV:1 (Viral Capsid Protein)



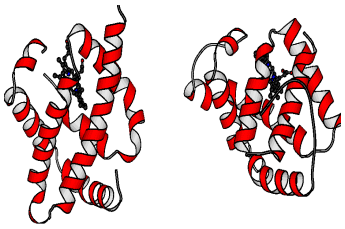
1HMP:A (Glycosyltransferase)

80 Residue Stretch (Yellow) with Over 40% Sequence Identity

Slide courtesy of Philip Bourne

Sequence → Structure → Function

Different sequence, similar structure & function



The globin fold is resilient to amino acid changes. *V. stercoraria* (bacterial) hemoglobin (left) and *P. marinus* (eukaryotic) hemoglobin (right) share just 8% sequence identity, but their overall fold and function is identical.

Slide courtesy of Philip Bourne

Sequence → Structure → Function

Some folds have many functions

NAD binding domain

This is a double beta-alpha-beta-alpha-beta motif, and is a common structural motif of enzymes binding NAD, NADP and other related cofactors, for example, NAD is found in dehydrogenases as the hydrogen acceptor. The domain is found as a common core unit in many structures, with other structural units at the periphery.

Number of EC numbers associated with this fold (to the third level): 5

P-loop NTP hydrolase fold

This fold consists of alpha/beta/alpha, parallel or mixed beta sheets of variable size. The fold binds the phosphate of ATP or GTP and is found in ATP and GTP binding proteins such as adenylate kinase. The P-loop is a phosphate binding loop which binds the phosphate groups of ATP and GTP,and is a glycine-rich sequence with the consensus sequence (A,G)xxxxGK(T,S). The P-loop residues are shown in detail (left) in guanylate kinase.

Number of EC numbers associated with this fold (to the third level): 5

Ferredoxin-like fold

This fold consists of an alpha/beta sandwich with an antiparallel beta sheet. The ferredoxin-like fold is associated with predominantly with non-enzymatic ferredoxins, like the example shown (Ferredoxin II from *D. gigas*, left). Ferredoxins are iron-sulphur clusters involved in electron transport, and often form part of multi-subunit assemblies. An example of an enzyme with this fold is muconolactone isomerase (EC 5.3.3.4).

Number of EC numbers associated with this fold (to the third level): 5

Outline

Protein structure databases
- Repositories
- Classifications

Protein function databases ←
- Gene Ontology (GO)
- Enzyme Commission (EC)

Sequence → Structure → Function
- Sequence alignment
- Structure alignment
- Sequence motifs
- Structure motifs

Protein Function

Possible levels of functional characterization

Molecular Function = elemental activity/task

- the tasks performed by individual gene products; examples are *carbohydrate binding* and *ATPase activity*

Biological Process = biological goal or objective

- broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by ordered assemblies of molecular functions

Cellular Component = location or complex

- subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *RNA polymerase II holoenzyme*

[Karen Christie et al., GO Annotation Camp 2005]

Protein Function

Possible levels of functional characterization

Molecular Function = elemental activity/task

- the tasks performed by individual gene products; examples are *carbohydrate binding* and *ATPase activity*

Biological Process = biologi

- broad biological goals, such as *mitosis* or ordered assemblies of molecular function

plished by

Structure is mostly useful for predicting molecular function

Cellular Component = location or complex

- subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *RNA polymerase II holoenzyme*

[Karen Christie et al., GO Annotation Camp 2005]

Protein Function Databases

Gene Ontology

- Directed acyclic graph of gene product attributes – provides functional terms at multiple levels for genes

Enzyme Commission (EC)

- Hierarchical classification of enzymes based on the chemical reactions they catalyze

Protein Function Databases

Gene Ontology

- Directed acyclic graph of gene product attributes – provides functional terms at multiple levels for genes

Enzyme Commission (EC)

- Hierarchical classification of enzymes based on the chemical reactions they catalyze

Protein Function Databases

Gene Ontology (top-level terms)

GO:0008150 : biological_process (137592)
- GO:0007610 : behavior (4736)
- GO:0000004 : biological process unknown (33210)
- GO:0009987 : cellular process (87883)
- GO:0007275 : development (18172)
- GO:0040007 : growth (3591)
- GO:0044419 : interaction between organisms (871)
- GO:0007582 : physiological process (92179)
- GO:0043473 : pigmentation (174)
- GO:0050789 : regulation of biological process (18806)
- GO:0000003 : reproduction (5312)
- GO:0016032 : viral life cycle (305)

GO:0005575 : cellular_component (126814)
- GO:0005623 : cell (94448)
- GO:0008372 : cellular component unknown (25532)
- GO:0031012 : extracellular matrix (1045)
- GO:0005576 : extracellular region (9981)
- GO:0043226 : organelle (65495)
- GO:0043234 : protein complex (13646)
- GO:0019012 : virion (133)

GO:0003674 : molecular_function (138386)
- GO:0016209 : antioxidant activity (542)
- GO:0005488 : binding (46487)
- GO:0003824 : catalytic activity (44201)
- GO:0030188 : chaperone regulator activity (50)
- GO:0030234 : enzyme regulator activity (2515)
- GO:0005554 : molecular function unknown (34234)
- GO:0003774 : motor activity (594)
- GO:0045735 : nutrient reservoir activity (43)
- GO:0031386 : protein tag (18)
- GO:0004871 : signal transducer activity (10883)
- GO:0005198 : structural molecule activity (3986)
- GO:0030528 : transcription regulator activity (9526)
- GO:0045182 : translation regulator activity (836)
- GO:0005215 : transporter activity (11587)
- GO:0030533 : triplet codon-amino acid adaptor activity (1217)

http://www.godatabase.org/

Protein Function Databases

Gene Ontology (molecular function terms)

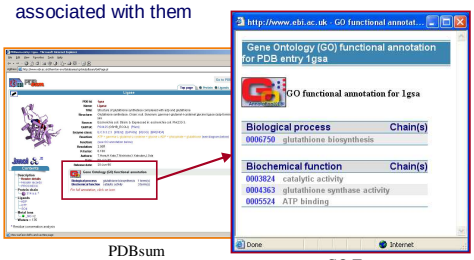
Term name	Total Gene Products	Percent
binding	46487	33.5
catalytic activity	44201	31.9
molecular function unknown	34234	24.7
transporter activity	11587	8.37
signal transducer activity	10883	7.86
transcription regulator activity	9526	6.88
structural molecule activity	3986	2.88
enzyme regulator activity	2515	1.81
triplet codon-amino acid adaptor activity	1217	0.87
translation regulator activity	836	0.60
motor activity	594	0.42
antioxidant activity	542	0.39
chaperone regulator activity	50	0.03
nutrient reservoir activity	43	0.03
protein tag	18	0.01
molecular_function	1	0.00

http://www.godatabase.org/

Protein Function Databases

Gene Ontology (molecular function terms)

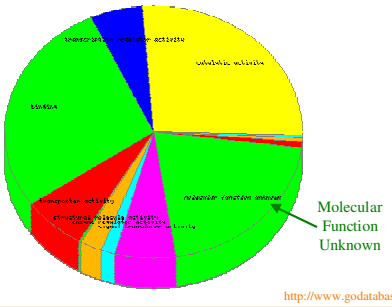
- Most proteins have several GO terms associated with them



PDBsum GO Terms

Protein Function Databases

Gene Ontology (molecular function terms)



http://www.godatabase.org/

Protein Function Databases

Gene Ontology

- Directed acyclic graph of gene product attributes – provides functional terms at multiple levels for genes

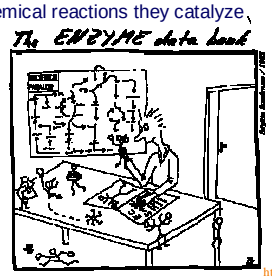
Enzyme Commission (EC) ←

- Hierarchical classification of enzymes based on the chemical reactions they catalyze

Protein Function Databases

Enzyme Commission (EC) numbers

- Hierarchical classification of enzymes based on the chemical reactions they catalyze



http://www.expasy.org/enzyme/

Protein Function Databases

Enzyme Commission (EC) numbers

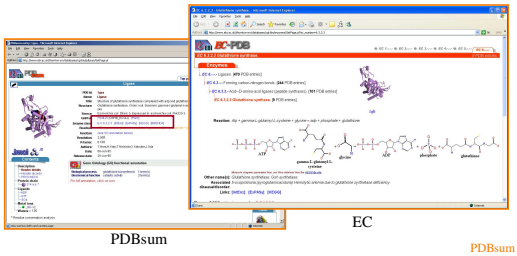
```
EC 1 Oxidoreductases
     EC 1.1 Acting on the CH-OH group of donors
          EC 1.1.1 With NAD+ or NADP+ as acceptor
          EC 1.1.2 With a cytochrome as acceptor
          EC 1.1.3 With oxygen as acceptor
          EC 1.1.4 With a disulfide as acceptor
          EC 1.1.5 With a quinone or similar compound as acceptor
          EC 1.1.99 With other acceptors
     EC 1.2 Acting on the aldehyde or oxo group of donors
          EC 1.2.1 With NAD+ or NADP+ as acceptor
          EC 1.2.2 With a cytochrome as acceptor
          EC 1.2.3 With oxygen as acceptor
          EC 1.2.4 With a disulfide as acceptor
          EC 1.2.7 With an iron-sulfur protein acceptor
          EC 1.2.99 With other acceptors
     EC 1.3 Acting on the CH-CH group of donors
          EC 1.3.1 With NAD+ or NADP+ as acceptor
          EC 1.3.2 With a cytochrome as acceptor
          EC 1.3.3 With oxygen as acceptor
          EC 1.3.5 With a quinone or related compound as acceptor
          EC 1.3.7 With an iron-sulfur protein as acceptor
          EC 1.3.99 With other acceptors
     EC 1.4 Acting on the CH-NH2 group of donors
          EC 1.4.1 With NAD+ or NADP+ as acceptor
          EC 1.4.2 With a cytochrome as acceptor
          EC 1.4.3 With oxygen as acceptor
          EC 1.4.4 With a disulfide as acceptor
          EC 1.4.7 With an iron-sulfur protein as acceptor
          EC 1.4.99 With other acceptors
etc.
```

http://www.expasy.org/enzyme/

Protein Function Databases

Enzyme Commission (EC) numbers

- Specify exact reaction catalyzed by enzyme



PDBsum EC

PDBsum

Outline

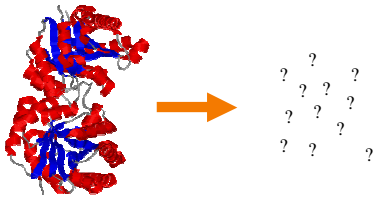
Protein structure databases
- Repositories
- Classifications

Protein function databases
- Gene Ontology (GO)
- Enzyme Commission (EC)

Sequence → Structure → Function ←
- Sequence alignment
- Structure alignment
- Sequence motifs
- Structure motifs

Sequence → Structure → Function

Goal:
- Given a protein sequence/structure, predict its function

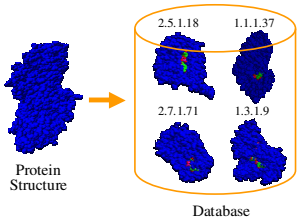


Sequence → Structure → Function

General strategy:
1. Given a protein with unknown function
2. Match it to proteins/templates with known functions
3. Transfer function from statistically significant matches

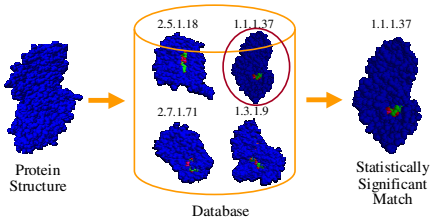
Sequence → Structure → Function

General strategy:
1. Given a protein with unknown function
2. Match it to proteins/templates with known functions
3. Transfer function from statistically significant matches



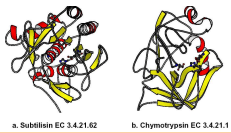
Sequence → Structure → Function

General strategy:
1. Given a protein with unknown function
2. Match it to proteins/templates with known functions
3. Transfer function from statistically significant matches



Sequence → Structure → Function

Evolution:
- Divergent evolution
  - § Homology: proteins share a common ancestor
    - – Orthology: separated by a speciation event
    - – Paralogy: separated by a gene duplication event
- Convergent evolution
  - § Analogy: similar structure evolves independently in two species due to similar selective pressures



Sequence → Structure → Function

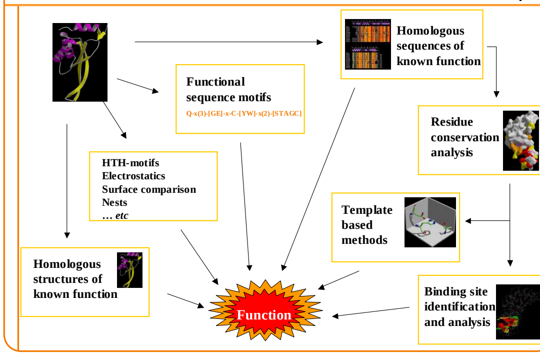
If proteins have similar sequences and structures they probably have similar functions

- >30% sequence identity
  - § Usually same structure & function
- 20-30% sequence identity
  - § Maybe related structure & function
  - § "Twighlight zone"
- <20% sequence identity
  - § Unlikely to be related
  - § "Midnight zone"

Structure is better preserved than sequence through evolution

Function Prediction (ProFunc)

Slide courtesy of James Watson



Outline

Protein structure databases
- Repositories
- Classifications

Protein function databases
- Gene Ontology (GO)
- Enzyme Commission (EC)

Sequence → Structure → Function
- Ø Sequence alignment
- Structure alignment
- Sequence motifs
- Structure motifs

Sequence Alignment

Find optimal alignment of protein sequences, and compute score for overlaps, substitutions, & gaps

```
> fasta34 d1gv0a d1guya

               10        20        30        40        50        60
d1gv0a AGVLDSARFRSFIAMELGVSMQDVTACVLGGHGDAMVPVVKYTTVAGIPVADLISAERIA
       :::::.::.:.::::: :::..:: : ..::::: :::. ...:..::::...:. .:.:
d1guya AGVLDAARYRTFIAMEAGVSVEDVQAMLMGGHGDEMVPLPRFSTISGIPVSEFIAPDRLA
               10        20        30        40        50        60

               70        80        90       100       110       120
d1gv0a ELVERTRTGGAEIVNHLKQGSAFYSPATSVVEMVESIVLDRKRVLTCAVSLDGQYGIDGT
       ..::::: ::.:::: :: :::.:.::.....:::... :.:::.   :. : ::::..
d1guya QIVERTRKGGGEIVNLLKTGSAYYAPAAATAQMVEAVLKDKKRVMPVAAYLTGQYGLNDI
               70        80        90       100       110       120

              130       140       150       160
d1gv0a FVGVPVKLGKNGVEHIYEIKLDQSDLDLLQKSAKIVDENCKML
       . :::: :: .:::.: :. :.. .. ::. ::: :
d1guya YFGVPVILGAGGVEKILELPLNEEEMALLNASAKAVRATLDTL
              130       140       150       160

54.487% identity
156 out of 163 amino acids overlap
Smith-Waterman score: 588
```

[Pearson88]

Sequence Alignment

Find optimal alignment of protein sequences, and compute score for overlaps, substitutions, & gaps


```
> fasta34 d1gv0a d1guya

               10        20        30        40        50        60
d1gv0a AGVLDSARFRSFIAMELGVSMQDVTACVLGGHGDAMVPVVKYTTVAGIPVADLISAERIA
d1guya AGVLDAARYRTFIAMEAGVSVEDVQAMLMGGHGDEMVPLPRFSTISGIPVSEFIAPDRLA

              130       140       150       160
d1gv0a FVGVPVKLGKNGVEHIYEIKLDQSDLDLLQKSAKIVDENCKML
d1guya YFGVPVILGAGGVEKILELPLNEEEMALLNASAKAVRATLDTL
              130       140       150       160
```

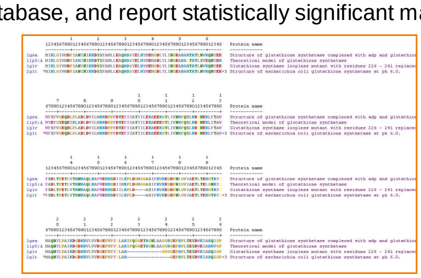

54.487% identity
156 out of 163 amino acids overlap
Smith-Waterman score: 588


```
54.487% identity
156 out of 163 amino acids overlap
Smith-Waterman score: 588
```


[Pearson88]

Sequence Database Search

Compute sequence alignment for query with all in database, and report statistically significant matches



FASTA

[Pearson88]

Sequence Database Search

Compute sequence alignment for query with all in database, and report statistically significant matches

Smith-Waterman score	%-tage identity	a.a. overlap	Length	z-score	E value		PDB code	Protein name
-	-	-	314	-	-		1gsa	*Structure of glutathione synthetase complexed with*
1914	90.1%	313	315	2255.5	3.4e-11	☑	1lp5:A	*Theoretical model of glutathione synthetase*
1906	94.3%	314	299	1729.1	7.1e-9	☑	1glv	*Glutathione synthase loopless mutant with residues*
1887	93.6%	314	296	1717.5	3.2e-8	☑	2glt	* *Structure of escherichia coli glutathione synthet*
"	"	"	"	"	"		= 1gsh	*Structure of escherichia coli glutathione synth*
"	"	"	"	"	"		= 1glt	*Glutathione synthase*

FASTA

[Pearson88]

Outline

Protein structure databases
- Repositories
- Classifications

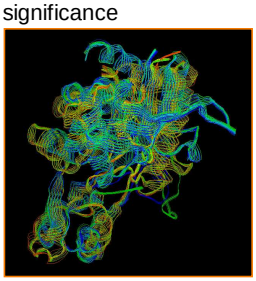
Protein function databases
- Gene Ontology (GO)
- Enzyme Commission (EC)

Sequence → Structure → Function
- Sequence alignment

ØStructure alignment
- Sequence motifs
- Structure motifs

Structure Alignment

Align protein structures
Report score and significance

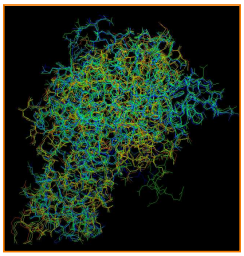


SSM

[Krissnel04]

Structure Alignment

Align protein structures
Report score and significance



SSM

[Krissnel04]

Structure Alignment

Align protein structures
Report score and significance

Hit	Select	Z-score	No. SSE	RMSD (Å)	Seq. id	PDB entry	Name
1	☑	22.2	28	0.00	100.0%	1gsa	Structure of glutathione synthetase complexed with adp and glutath
2	☐	19.9	27	0.39	99.3%	1gsh	Structure of escherichia coli glutathione synthetase at ph 7.5
3	☐	19.4	27	0.42	99.3%	2glt	Structure of escherichia coli glutathione synthetase at ph 6.0.
4	☐	19.3	27	0.46	99.0%	1glv	Glutathione synthase loopless mutant with residues 226 - 241 replac
5	☐	19.0	26	0.55	89.8%	1lp5A	Theoretical model of glutathione synthetase
6	☐	7.9	18	2.49	26.2%	1uc9A	Crystal structure of a lysine biosynthesis enzyme, lysx, from thermu
7	☐	7.6	17	2.43	26.4%	1uc8A	Crystal structure of a lysine biosynthesis enzyme, lysx, from thermu
8	☐	6.7	18	2.67	18.7%	1pk8C	Crystal structure of rat synapsin i c domain complexed to ca.Atp
9	☐	6.4	18	2.71	21.0%	1f7G	Crystal structure analysis of the complex of thE C domain of synaps
10	☐	6.4	18	2.71	18.2%	1auxB	Structure of thE C domain of synapsin Ia from bovine brain with calc

View Key to table entries

Alignment

SSM

[Krissnel04]

Outline

Protein structure databases
- Repositories
- Classifications

Protein function databases
- Gene Ontology (GO)
- Enzyme Commission (EC)

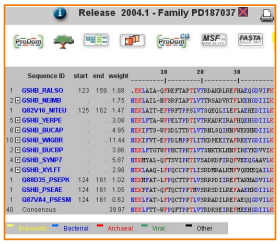
Sequence → Structure → Function
- Sequence alignment
- Structure alignment

ØSequence motifs
- Structure motifs

Sequence Motifs

Recognize local patterns (motifs) in protein sequences associated with specific functions

Example: Prokaryotic glutathione synthetase ATP-binding domain



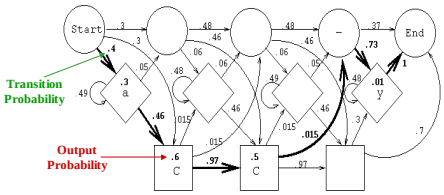
ProDom

http://prodes.toulouse.inra.fr/prodom/

Sequence Motifs

Recognize local patterns (motifs) in protein sequences associated with specific functions

- Represent as Hidden Markov Model (HMM)

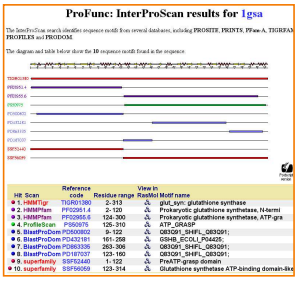


Sequence Motifs

Match query sequence against all motifs

Examples:

- InterPro
- PROSITE
- PRINTS
- PFam-A
- TIGRFAM
- PROFILES
- PRODOM



InterPro [Zdobnov01]

Outline

Protein structure databases

- Repositories
- Classifications

Protein function databases

- Gene Ontology (GO)
- Enzyme Commission (EC)

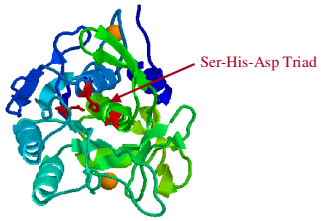
Sequence → Structure → Function

- Sequence alignment
- Structure alignment
- Sequence motifs

ØStructure motifs

Structure Motifs

Recognize local patterns (motifs) in protein structures associated with specific functions

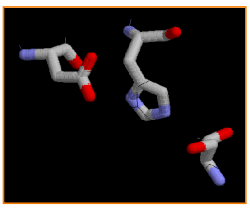


Subtilisin (B. amyloliquefaciens)

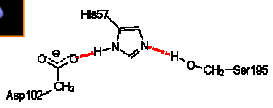
http://chemistry.umeche.maine.edu/

Structure Motifs

Example: serine proteases



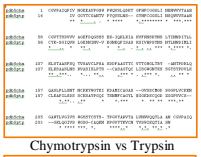
Ser-His-Asp Triad



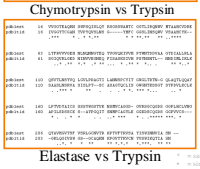
http://chemistry.umeche.maine.edu/

Structure Motifs

Example: serine proteases:



Chymotrypsin vs Trypsin



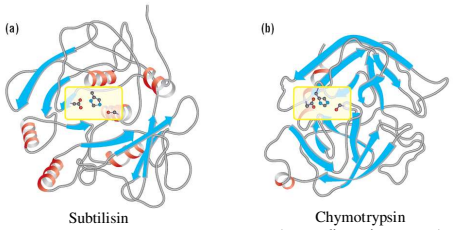
Elastase vs Trypsin

Trypsin (yellow)
Elastase (green)
Chymotrypsin (blue)

http://chemistry.umeche.maine.edu/

Structure Motifs

Example: serine proteases



Subtilisin
(bacterial serine protease)

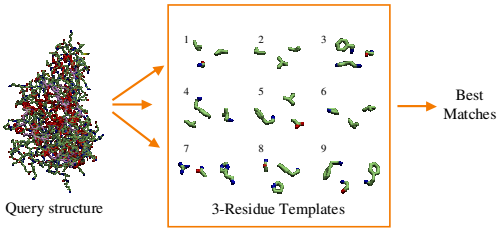
Chymotrypsin
(mammalian serine protease)

http://www.sinauer.com/pdf/06-09-03-1-16.pdf

Structural Motifs

Slide courtesy of James Watson

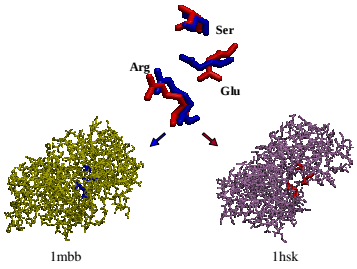
Match protein residues to functional templates, and measure quality/significance of matches



Structural Motifs

Slide courtesy of James Watson

Match protein residues to functional templates, and measure quality/significance of matches



Structural Motifs

Match protein residues to functional templates, and measure quality/significance of matches

b. Probable matches (1.00E-06 < E-value < 0.01)

Hit no.	E-value	Similarity score	Neighbours identical [equiv]	Template id	Matched PDB entry	Longest fitted segment	Seq lengths query/target	Overlap	%-tage seq id	Structural similarity
1	2.01E-04	257.34	10/11 [21]	ATPc0011	1dv2: The structure of biotin carboxylase, mutant e288k, complexed with adp	89/188	314/450	378	16.56%	94.1%
2	6.15E-04	242.38	11/11 [20]	POBc0003	1lov: Complex of d-ala d-ala ligase with adp and a phosphonyl phosphonate	138/219	314/306	360	14.71%	99.2%

c. Possible matches (0.01 < E-value < 0.10)
... no hits in this category

d. Long shots (0.10 < E-value < 10.0)

Hit no.	E-value	Similarity score	Neighbours identical [equiv]	Template id	Matched PDB entry	Longest fitted segment	Seq lengths query/target	Overlap	%-tage seq id	Structural similarity
3	1.514	137.69	8/9 [17]	I4Pc0004	1z2o: inositol 1,3,4-trisphosphate 5/6-kinase in complex with mg2+/adpnhs (1,3,4,6)p4	75/176	314/311	375	16.08%	91.4%
4	12.601	109.28	9/9 [14]	PHYc0003	1e4e: D-alanyl-d-lacate ligase	76/156	314/341	390	19.43%	91.3%

[Laskowski05]

Summary

Protein structure databases
- Repositories
- Classifications

Protein function databases
- Gene Ontology (GO)
- Enzyme Commission (EC)

Sequence → Structure → Function
- Sequence alignment
- Structure alignment
- Sequence motifs
- Structure motifs

Going Forward

Study algorithms
- Next meeting: structural alignment

Think about new methods
- e.g., new structural motifs

Think about how methods interact

Acknowledgements

Slides:
- James Watson
- Philip Bourne