Protein Structure Analysis

Thomas Funkhouser
Princeton University
CS597A, Fall 2007

Outline

Protein structure databases
- Primary
- Secondary
- Tertiary
- Quaternary

Protein structure visualization
- Demo

Protein structure analysis
- Geometry
- Relationship to sequence
- Classifications

Outline

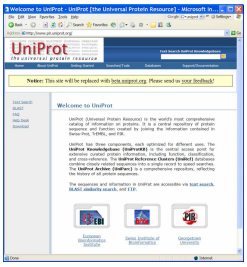
Protein structure databases
- ØPrimary
- Secondary
- Tertiary
- Quaternary

Protein structure visualization
- Demo

Protein structure analysis
- Geometry
- Relationship to sequence
- Classifications

Primary Structure Statistics

Main databases:
- UniProtKB/Swiss-Prot ← curated
- UniProtKB/TrEMBL



http://www.pir.uniprot.org/

Primary Structure Statistics

Information provided:
- Sequence of amino acid types

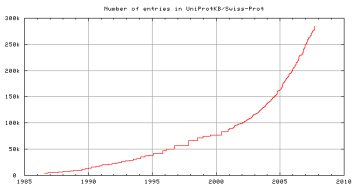
Chain 1GSA:_
Compound Glutathione Synthetase
Type Protein
Molecular Weight 35547
Number of Residues 316

```
  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]

Primary Structure Statistics

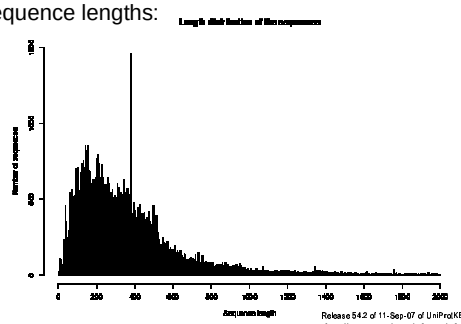
Sequence counts:
- 283,454 sequence entries in UniProtKB/Swiss-Prot
- 4,754,787 sequence entries in UniProtKB/TrEMBL



Release 54.2 of 11-Sep-07 of UniProtKB/Swiss-Prot
http://expasy.org/sprot/relnotes/relstat.html

Primary Structure Statistics

Sequence lengths:



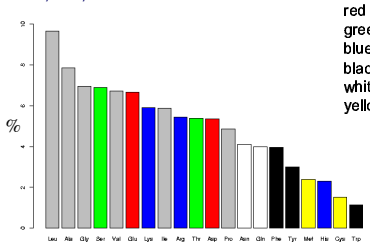
Release 54.2 of 11-Sep-07 of UniProtKB/Swiss-
http://expasy.org/sprot/relnotes/relstat.html

Primary Structure Statistics

Amino acid counts:

- 104,030,551 total residues

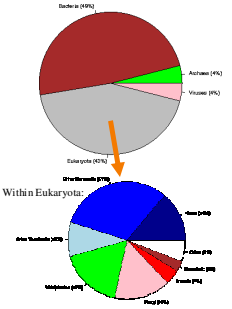
gray = aliphatic
red = acidic
green = small hydroxy
blue = basic
black = aromatic
white = amide
yellow = sulfur



Release 54.2 of 11-Sep-07 of UniProtKB/Swiss-
http://expasy.org/sprot/relnotes/relstat.html

Primary Structure Statistics

Taxonomic distribution:



Most Represented Species:

Count	Species
17169	Homo sapiens (Human)
13826	Mus musculus (Mouse)
6493	Saccharomyces cerevisiae (Baker's yeast)
6312	Rattus norvegicus (Rat)
6065	Arabidopsis thaliana (Mouse-ear cress)
4402	Escherichia coli (strain K12)
4272	Bos taurus (Bovine)
3467	Schizosaccharomyces pombe (Fission yea
3072	Caenorhabditis elegans
2860	Bacillus subtilis

Release 54.2 of 11-Sep-07 of UniProtKB/Swiss-
http://expasy.org/sprot/relnotes/relstat.html

Outline

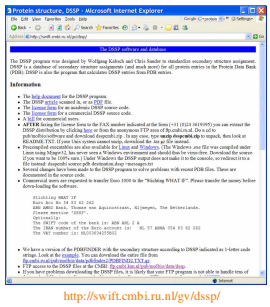
Protein structure databases
- Primary
- Secondary
- Tertiary
- Quaternary

Protein structure visualization
- Demo

Protein structure analysis
- Geometry
- Relationship to sequence
- Classifications

Secondary Structure Statistics

Main database:
- DSSP



http://swift.cmbi.ru.nl/gv/dssp/

Secondary Structure Statistics

Information provided:
- Predicted secondary structure element for every residue

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
      HHHHHHHHTT  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
```

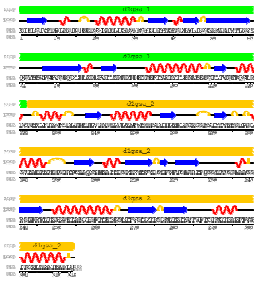
H = helix
B = residue in isolated beta bridge
E = extended beta strand
G = 310 helix
T = hydrogen bonded turn
S = bend

[Kabsch83]

Secondary Structure Statistics

Information provided:

- Predicted secondary structure element for every residue



[Kabsch83]

Outline

Protein structure databases

- Primary
- Secondary

ØTertiary

- Quaternary

Protein structure visualization

- Demo

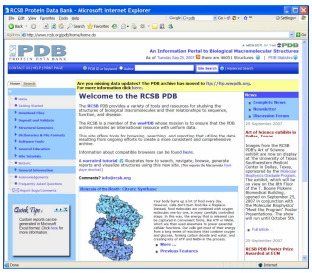
Protein structure analysis

- Geometry
- Relationship to sequence
- Classifications

Tertiary Structure Statistics

Main database:

- PDB



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

Tertiary Structure Statistics

Information provided:

- Atomic coordinates for every atom
- Remarks and info about experiment

```
HEADER    LIGASE                                  08-JUN-95
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: GLUTATHIONE SYNTHETASE;
COMPND   3 CHAIN: NULL;
COMPND   4 SYNONYM: GAMMA-L-GLUTAMYL-L-CYSTEINE\:GLYCINE LIGASE
COMPND   5 (ADP-FORMING);
COMPND   6 EC: 6.3.2.3;
COMPND   7 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI;
SOURCE   3 STRAIN: B;
SOURCE   4 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   5 EXPRESSION_SYSTEM_STRAIN: JM109;
SOURCE   6 EXPRESSION_SYSTEM_PLASMID: PGS500, AN DERIVATIVE PLASMID OF
SOURCE   7 PKK233-3;
SOURCE   8 EXPRESSION_SYSTEM_GENE: GSHI
KEYWDS    LIGASE, GLUTATHIONE SYNTHETASE
EXPDTA    X-RAY DIFFRACTION - SINGLE CRYSTAL
AUTHOR    T.HARA,H.KATO,T.NISHIOKA,Y.KATSUBE,J.ODA
```

• • •

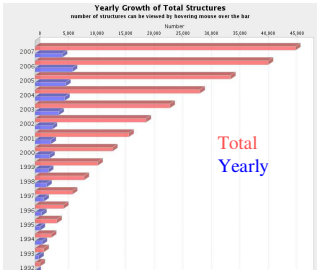
```
ATOM      1  N   MET A   1      31.467  -6.342  40.148  1.00 34.76
ATOM      2  CA  MET A   1      30.541  -6.042  49.013  1.00 32.18
ATOM      3  C   MET A   1      30.378  -5.075  49.799  1.00 29.56
ATOM      4  O   MET A   1      29.614  -4.985  48.741  1.00 30.36
ATOM      5  CB  MET A   1      32.985  -5.385  49.850  1.00 34.51
ATOM      6  CG  MET A   1      33.289  -4.072  49.133  1.00 38.14
ATOM      7  SD  MET A   1      34.638  -4.118  47.476  1.00 42.62
ATOM      8  CE  MET A   1      32.774  -3.418  46.457  1.00 40.00
ATOM      9  N   ILE A   2      30.166  -4.481  50.861  1.00 26.67
ATOM     10  CA  ILE A   2      29.216  -3.412  51.049  1.00 21.85
ATOM     11  C   ILE A   2      29.937  -2.163  50.584  1.00 19.42
ATOM     12  O   ILE A   2      31.147  -2.020  50.777  1.00 18.91
ATOM     13  CB  ILE A   2      28.849  -3.371  52.527  1.00 21.13
ATOM     14  CG1 ILE A   2      28.165  -4.679  52.917  1.00 22.01
ATOM     15  CG2 ILE A   2      27.928  -2.210  52.892  1.00 17.73
ATOM     16  CD1 ILE A   2      27.877  -4.876  54.380  1.00 22.74
ATOM     17  N   LYS A   3      29.238  -1.277  49.965  1.00 18.72
ATOM     18  CA  LYS A   3      29.765  0.060  49.567  1.00 17.79
ATOM     19  C   LYS A   3      29.887  0.969  50.359  1.00 18.09
ATOM     20  O   LYS A   3      27.769  1.016  50.359  1.00 18.09
```

• • •

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

Tertiary Structure Statistics

Structure count:



~45,000 Protein Structures

http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html

Tertiary Structure Statistics

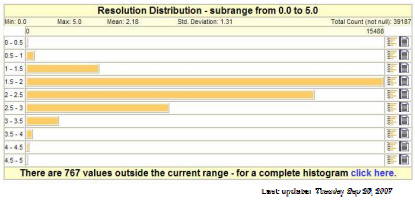
Experimental method:

		Molecule Type				
		Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
Exp. Method	X-ray	36466	989	1705	24	39184
	NMR	5696	784	134	7	6621
	Electron Microscopy	106	10	38	0	154
	Other	82	4	4	2	92
	Total	42350	1787	1881	33	46051

http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html

Tertiary Structure Statistics

Resolution:



http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html

Outline

Protein structure databases
- Primary
- Secondary
- Tertiary

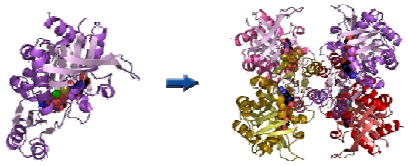
ØQuaternary

Protein structure visualization
- Demo

Protein structure analysis
- Geometry
- Relationship to sequence
- Classifications

Quaternary Structure Statistics

Main databases:
- PQS
- PISA



Asymmetric unit

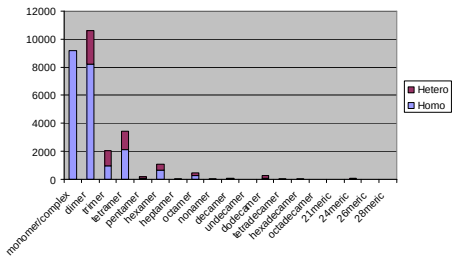
1gsa

Biological unit, tetramer
- as defined in PDB file (see also PQS)

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

Quaternary Structure Statistics

Counts of monomeric / oligomeric proteins in PQS:

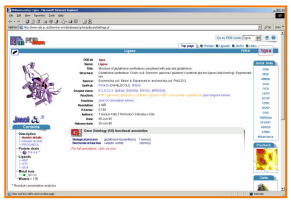


PQS

Protein Structure Databases

Useful resources:

ØPDBsum
- Jena
- MSD

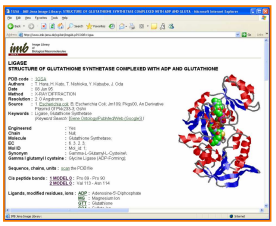


http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/ [Laskowski05]

Protein Structure Databases

Useful resources:
- PDBsum

ØJena
- MSD



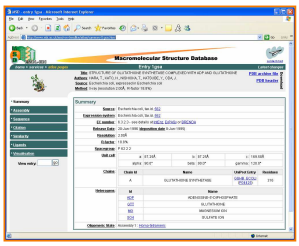
http://www.imb-jena.de/IMAGE.html

Protein Structure Databases

Useful resources:

- PDBsum
- Jena

ØMSD



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

Outline

Protein structure databases
- Primary
- Secondary
- Tertiary
- Quaternary

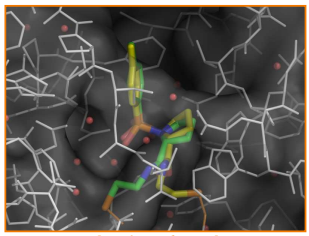
Protein structure visualization
- Demo

Protein structure analysis
- Geometry
- Relationship to sequence
- Classifications

Protein Structure Visualization

Some tools:

- PyMOL
- OpenRasMOL
- Protein Explorer
- Grasp
- etc.



[pymol.sourceforge.net]

Protein Structure Visualization

Demo!

by Meghan Bellows

Outline

Protein structure databases
- Primary
- Secondary
- Tertiary
- Quaternary

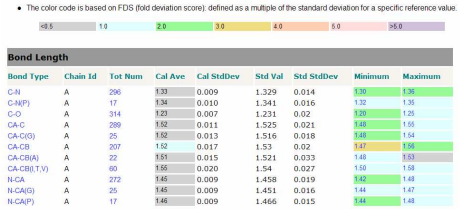
Protein structure visualization
- Demo

Protein structure analysis

ØGeometry
- Relationship to sequence
- Classifications

Protein Structure Analysis

Bond types (for 1gsa):



- The color code is based on FDS (fold deviation score): defined as a multiple of the standard deviation for a specific reference value.

<0.5 | 1.0 | 2.0 | 3.0 | 4.0 | 5.0 | >5.0

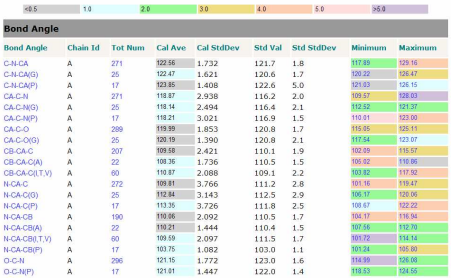
Bond Length

Bond Type	Chain Id	Tot Num	Cal Ave	Cal StdDev	Std Val	Std StdDev	Minimum	Maximum
C-N	A	296	1.33	0.009	1.329	0.014	1.30	1.36
C-N(P)	A	17	1.34	0.010	1.341	0.016	1.32	1.35
C-O	A	314	1.23	0.007	1.231	0.02	1.20	1.25
CA-C	A	289	1.52	0.011	1.525	0.021	1.48	1.55
CA-C(G)	A	25	1.52	0.013	1.516	0.018	1.48	1.54
CA-CB	A	207	1.52	0.017	1.53	0.02	1.47	1.56
CA-CB(A)	A	22	1.51	0.015	1.521	0.033	1.48	1.53
CA-CB(I,T,V)	A	60	1.55	0.020	1.54	0.027	1.50	1.58
N-CA	A	272	1.45	0.009	1.458	0.019	1.42	1.48
N-CA(G)	A	25	1.45	0.009	1.451	0.016	1.44	1.47
N-CA(P)	A	17	1.46	0.009	1.466	0.015	1.44	1.48

http://www.rcsb.org/pdb/explore/geometryDisplay.do?structureId=1GSA

Protein Structure Analysis

Bond angles (for 1gsa):



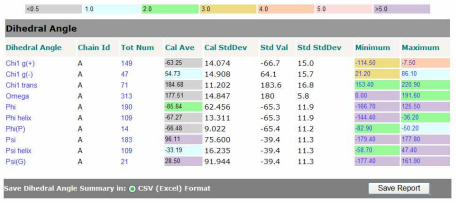
Bond Angle

Bond Angle	Chain Id	Tot Num	Cal Ave	Cal StdDev	Std Val	Std StdDev	Minimum	Maximum
C-N-CA	A	271	122.56	1.732	121.7	1.8	117.89	129.16
C-N-CA(G)	A	25	122.47	1.621	120.6	1.7	120.22	124.47
C-N-CA(P)	A	17	123.85	1.408	122.6	5.0	121.03	126.15
CA-C-N	A	271	118.87	2.938	116.2	2.0	109.57	128.03
CA-C-N(G)	A	25	118.14	2.494	116.4	2.1	112.52	121.37
CA-C-N(P)	A	17	118.21	3.021	116.9	1.5	110.61	123.00
CA-C-O	A	289	119.39	1.853	120.8	1.7	115.05	125.11
CA-C-O(G)	A	25	120.15	1.390	120.8	2.1	117.54	123.07
CB-CA-C	A	207	109.58	2.421	110.1	1.9	102.59	116.67
CB-CA-C(A)	A	22	108.36	1.736	110.5	1.5	105.92	110.86
CB-CA-C(LT,V)	A	60	110.87	2.088	109.1	2.2	103.92	117.92
N-CA-C	A	272	109.81	3.766	111.2	2.8	101.16	119.47
N-CA-C(G)	A	25	112.84	3.143	112.5	2.9	106.37	120.06
N-CA-C(P)	A	17	113.35	3.726	111.8	2.5	108.67	122.22
N-CA-CB	A	190	110.06	2.092	110.5	1.7	104.17	116.94
N-CA-CB(A)	A	22	110.21	1.444	110.4	1.5	107.56	112.70
N-CA-CB(LT,V)	A	60	109.59	2.097	111.5	1.7	103.72	114.14
N-CA-CB(P)	A	17	103.75	1.082	103.0	1.1	101.24	105.60
O-C-N	A	296	121.15	1.772	123.0	1.6	114.99	126.08
O-C-N(P)	A	17	121.01	1.447	122.0	1.4	118.53	124.55

http://www.rcsb.org/pdb/explore/geometryDisplay.do?structureId=1GSA

Protein Structure Analysis

Dihedral angles (for 1gsa):



Dihedral Angle

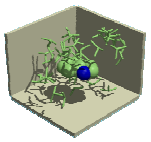
Dihedral Angle	Chain Id	Tot Num	Cal Ave	Cal StdDev	Std Val	Std StdDev	Minimum	Maximum
Chi1 g(+)	A	149	-63.25	14.074	-66.7	15.0	-114.50	-7.50
Chi1 g(-)	A	47	54.73	14.908	64.1	15.7	21.20	86.10
Chi1 trans	A	71	184.68	11.202	183.6	16.8	153.40	220.90
Omega	A	313	177.51	14.847	180	5.8	0.00	191.60
Phi	A	190	-65.54	62.456	-65.3	11.9	-166.70	126.50
Phi helix	A	109	-67.27	13.311	-65.3	11.9	-144.60	-36.26
Phi(P)	A	14	-66.48	9.022	-65.4	11.2	-82.90	-50.20
Psi	A	183	96.11	75.600	-39.4	11.3	-179.60	177.80
Psi helix	A	109	-33.19	16.235	-39.4	11.3	-68.70	47.46
Psi(G)	A	21	28.59	91.944	-39.4	11.3	-177.40	161.90

Save Dihedral Angle Summary in: CSV (Excel) Format Save Report

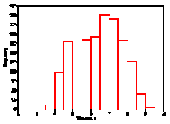
http://www.rcsb.org/pdb/explore/geometryDisplay.do?structureId=1GSA

Protein Structure Analysis

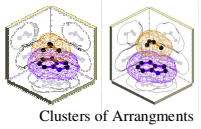
Side-chain arrangements: TRP↔LEU



Sample of 3D interactions



Distributions of distances



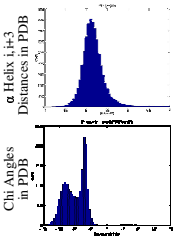
Clusters of Arrangments

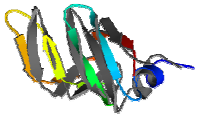
http://www.biochem.ucl.ac.uk/bsm/sidechains/Trp/Leu/ip.html

Protein Structure Analysis

Example application:

- Scott McAllister: "Generating Likely Distance Bounds for Efficient Protein Structure Prediction" (COS 597A)





	Original		Distance Bound		Both Bounds	
Rank	Energy	RMSD	Energy	RMSD	Energy	RMSD
1	-412.7	10.11	-407.6	7.55	-365.4	11.21
2	-397.0	10.68	-378.3	3.88	-329.5	10.18
3	-388.5	7.37	-375.3	10.35	-310.2	10.24
4	-369.9	8.40	-363.8	7.41	-305.9	10.98
Best	-282.4	5.50	-378.3	3.88	-209.3	4.93

Outline

Protein structure databases
- Primary
- Secondary
- Tertiary
- Quaternary

Protein structure visualization
- Demo

Protein structure analysis
- Geometry

ØRelationship to sequence

- Classifications

Sequence → Structure?

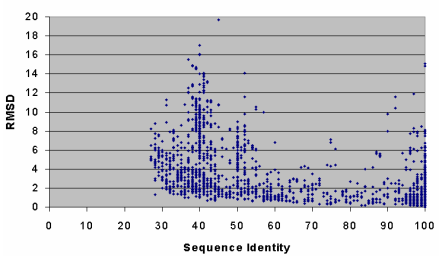
If proteins have similar sequences
they probably have similar structures

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

Slide courtesy of Philip Bourne

Sequence → Structure?

Relationships between sequence and structure

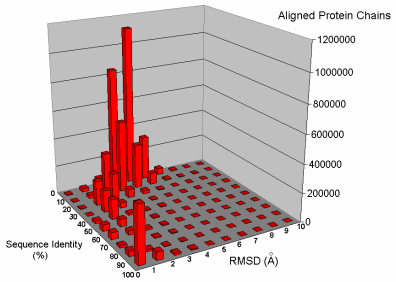


Structure Comparison of 30% of PDBSelect Set

Slide courtesy of Philip Bourne

Sequence → Structure?

Relationships between sequence and structure

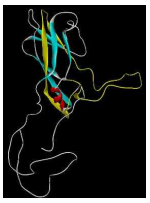


Structure Alignments using CE with z>4.0

Slide courtesy of Philip Bourne

Sequence → Structure?

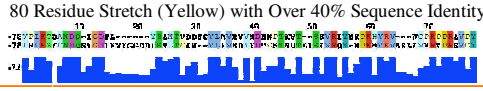
Similar sequence, different structure & function



1PIV:1 (Viral Capsid Protein)

1HMP:A (Glycosyltransferase)

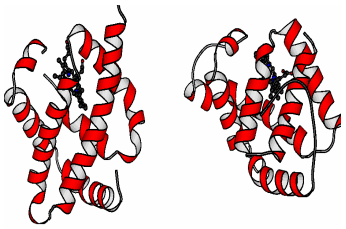
80 Residue Stretch (Yellow) with Over 40% Sequence Identity



Slide courtesy of Philip Bourne

Sequence → Structure?

Different sequence, similar structure & function

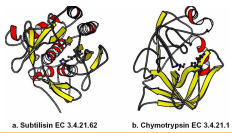


The globin fold is resilient to amino acid changes. *V. stercoraria* (bacterial) hemoglobin (left) and *P. marinus* (eukaryotic) hemoglobin (right) share just 8% sequence identity, but their overall fold and function is identical.

Sequence → Structure?

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



Outline

Protein structure databases

- Primary
- Secondary
- Tertiary
- Quaternary

Protein structure visualization

- Demo

Protein structure analysis

- Geometry
- Relationship to sequence

ØClassifications

Protein Structure Classifications

Main databases:

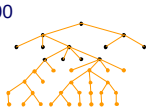
- CATH
- SCOP

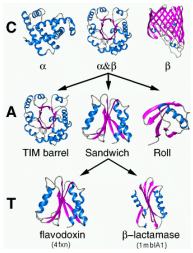
Protein Structure Classifications

CATH hierarchy:

- Class
- Architecture
- Topology
- Homology
- S35 (Family)
- S60
- S95
- S100

} Structural Layout





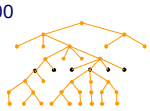
http://cathwww.biochem.ucl.ac.uk/ [Orengo97]

Protein Structure Classifications

CATH hierarchy:

- Class
- Architecture
- Topology
- Homology
- S35 (Family)
- S60
- S95
- S100

} Evolution



CATH Domain 1gsa01

Classification

Class Alpha Beta	3
Architecture 3-Layer(aba) Sandwich	3.40
Topology Rossmann fold	3.40.50
Homologous Superfamily LIGASE	3.40.50.20
Sequence Family (S35) LIGASE	3.40.50.20.7
Non-identical (S95) LIGASE	3.40.50.20.7.1
Identical (S100) LIGASE	3.40.50.20.7.1.1

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

Protein Structure Classifications

CATH hierarchy:

- Class
- Architecture
- Topology
- Homology
- S35 (Family)
- S60
- S95
- S100

} Sequence Identity

CATH Domain 1gsa01

Classification

Class Alpha Beta	3
Architecture 3-Layer(aba) Sandwich	3.40
Topology Rossmann fold	3.40.50
Homologous Superfamily LIGASE	3.40.50.20
Sequence Family (S35) LIGASE	3.40.50.20.7
Non-identical (S95) LIGASE	3.40.50.20.7.1
Identical (S100) LIGASE	3.40.50.20.7.1.1

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

Protein Structure Classifications

CATH hierarchy:

- Class (4)
- Architecture (40)
- Topology (1084)
- Homology (2091)
- S35 (7794)
- S60 (10363)
- S95 (13781)
- S100 (25491)

CATH Domain 1gsa01

Classification

Class Alpha Beta	3
Architecture 3-Layer(aba) Sandwich	3.40
Topology Rossmann fold	3.40.50
Homologous Superfamily LIGASE	3.40.50.20
Sequence Family (S35) LIGASE	3.40.50.20.7
Non-identical (S95) LIGASE	3.40.50.20.7.1
Identical (S100) LIGASE	3.40.50.20.7.1.1

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

Protein Structure Classifications

SCOP hierarchy:

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

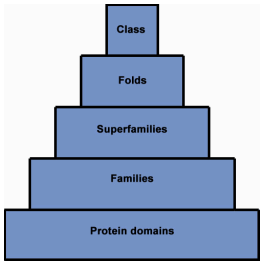
SCOP: 1gsa

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:



Class	7
Folds	971
Superfamilies	1,589
Families	3,004
Protein domains	75,930

SCOP: Structural Classification of Proteins (1.71 release)

Summary

Protein structure databases

- Lots of structural data is available

Protein structure analysis

- Structure provides information about function
- Need good tools for analysis!