Overview of Protein Structure and its classification

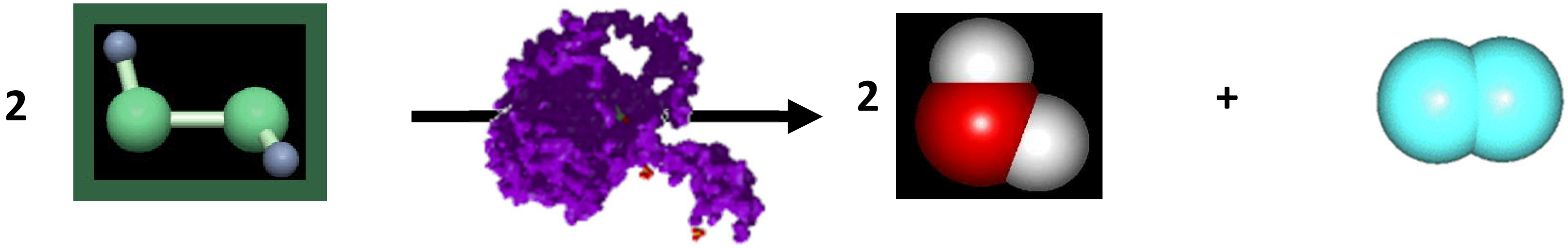
Incorporated with Lawrence Hunter (University of Colorado), Kun Huang (OSU) and Doug Brutlag (Stanford)

Proteins’ roles……

If there is a job to be done in the molecular world of our cells, usually that job is done by a protein.

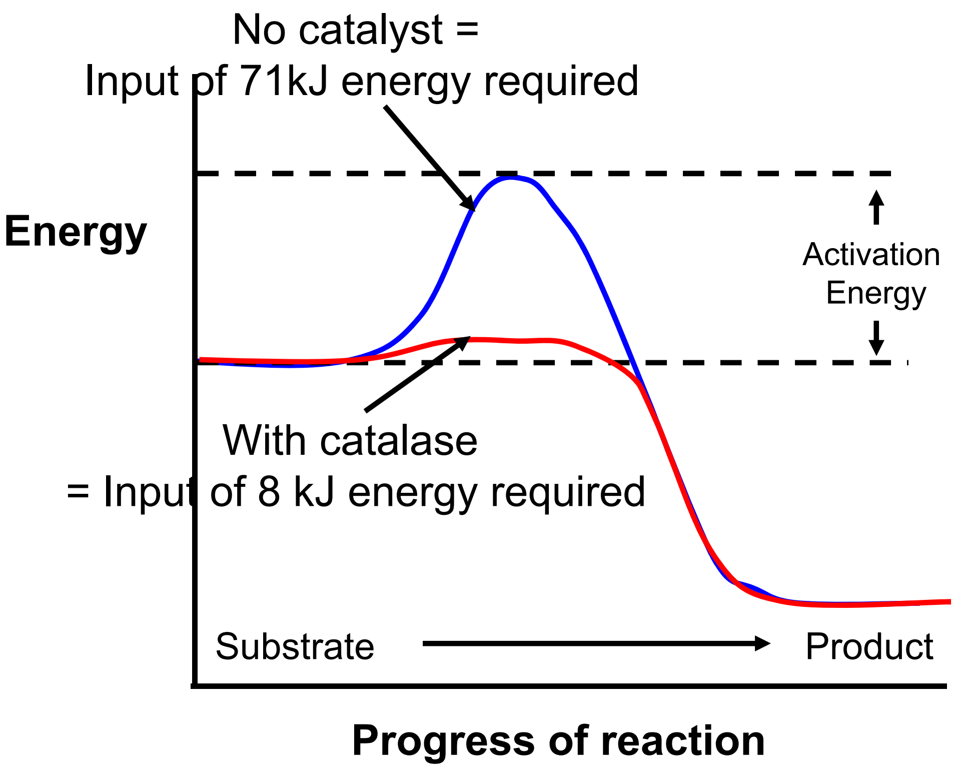
Examples of proteins include hormones acting as messengers; enzymes speeding up reactions; cell receptors acting as ‘antennae’; antibodies fighting foreign invaders; membrane channels allowing specific molecules to enter or leave a cell; they make up the muscles for moving; let you grow hair, ligaments and fingernails; and let you see (the lens of your eye is pure crystallized protein).

Proteins speed up reactions - Enzymes



Catalase speeds up the breakdown of hydrogen peroxide, (H_2O_2) a toxic by product of metabolic reactions, to the harmless substances, water and oxygen.

The reaction is extremely rapid as the enzyme lowers the energy needed to kick-start the reaction (activation energy)



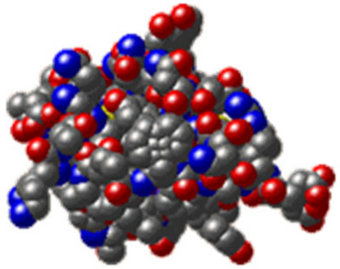
Proteins can regulate metabolism – hormones

When your body detects an increase in the sugar content of blood after a meal, the hormone insulin is released from cells in the pancreas.

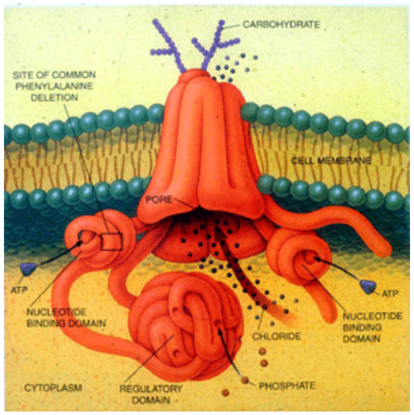
Insulin binds to cell membranes and this triggers the cells to absorb glucose for use or for storage as glycogen in the liver.

Insulin

$C_{254}H_{377}N_{65}O_{76}S_6$



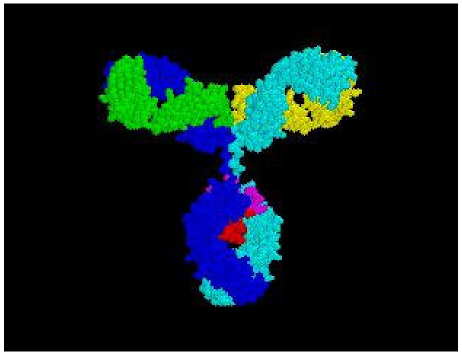
Proteins span membranes –protein channels


CARBOHYDRATE
SITE OF COMMON PHENYLALANINE DELETION
CELL MEMBRANE
PORE
ATP
NUCLEOTIDE BINDING DOMAIN
ATP
NUCLEOTIDE BINDING DOMAIN
CHLORIDE
CYTOPLASM
REGULATORY DOMAIN
PHOSPHATE

The CFTR membrane protein is an ion channel that regulates the flow of chloride ions.

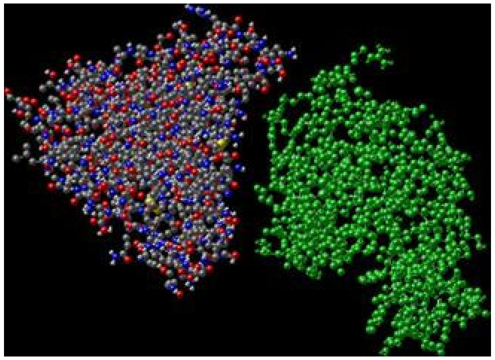
Not enough of this protein gets inserted into the membranes of people suffering Cystic fibrosis. This causes secretions to become thick as they are not hydrated. The lungs and secretory ducts become blocked as a consequence.

Proteins Defend us against pathogens –antibodies



Left: **Antibodies** like IgG found in humans, recognise and bind to groups of molecules or **epitopes** found on foreign invaders.

Right: The binding site of an **antigen** protein (left) interacting with the epitope of a foreign antigen (green)



Protein Folding

- Proteins are created linearly and then assume their tertiary structure by “folding.”
  - Exact mechanism is still unknown
  - Mechanistic simulations can be illuminating
- Proteins assume the lowest energy structure
  - Or sometimes an ensemble of low energy structures.
- Hydrophobic collapse drives process
- Local (secondary) structure proclivities
- Internal stabilizers:
  - Hydrogen bonds, disulphide bonds, salt bridges.

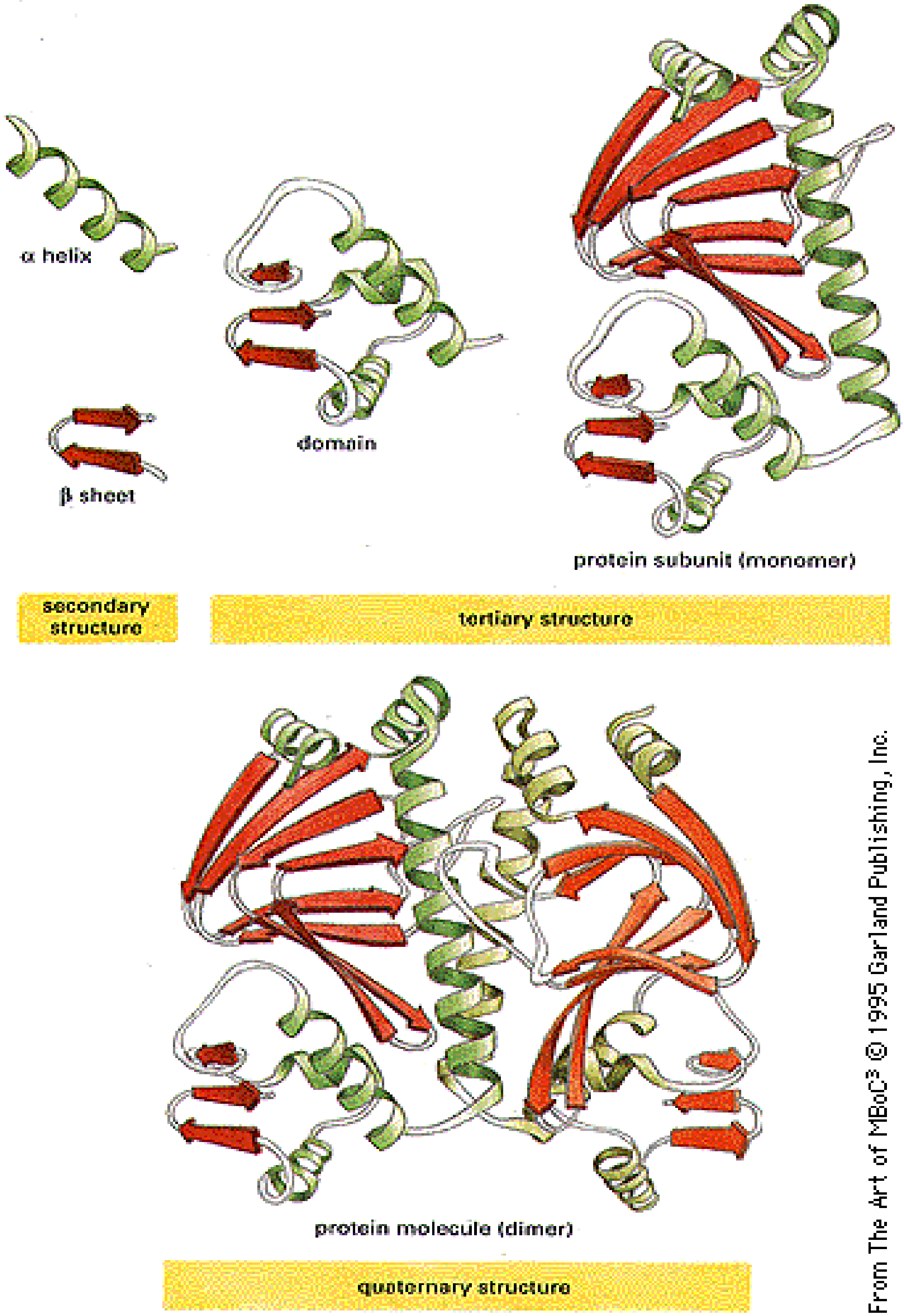
Protein structure

- Most proteins will fold spontaneously in water, so amino acid sequence alone should be enough to determine protein structure
- However, the physics are daunting:
  - 20,000+ protein atoms, plus equal amounts of water
  - Many non-local interactions
  - Can takes seconds (most chemical reactions take place ~10^{12} --1,000,000,000,000x faster)
- Empirical determinations of protein structure are advancing rapidly.

Protein Structure Levels

- Protein structure is described in four levels
  - Primary structure: amino acid sequence
  - Secondary structure: local (in sequence) ordering into
    - (α)Helices: compressed, corkscrew structures
    - (β)Strands: extended, nearly straight structures
    - (β)Sheets: paired strands, reinforced by hydrogen bonds
      - parallel (same direction) or antiparallel sheets
    - Coils, Turns & Loops: changes in direction
  - Tertiary structure: global ordering (all angles/atoms)
  - Quaternary structures: multiple, disconnected amino acid chains interacting to form a larger structure

Protein structure cartoons



From The Art of MBoC^3 © 1995 Garland Publishing, Inc.

Why do we need structure prediction?

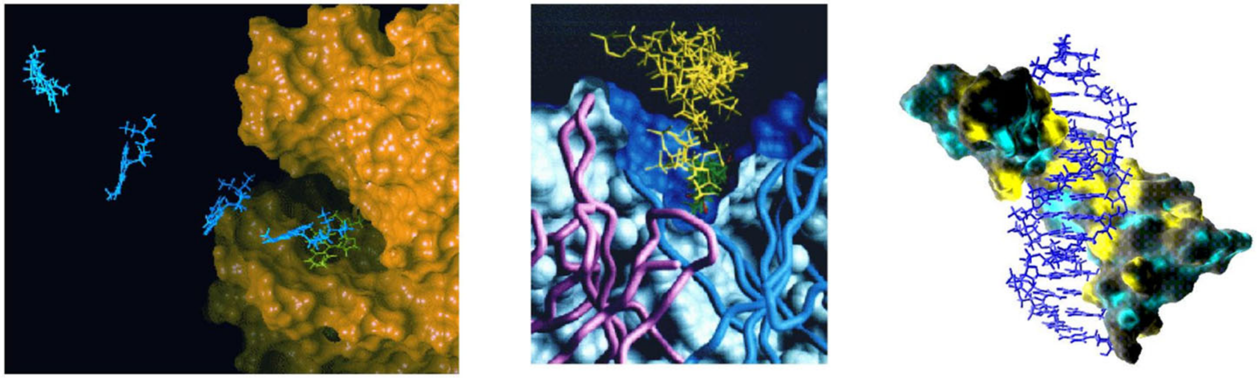
- 3D structure give clues to function:
  - active sites, binding sites, conformational changes...
  - structure and function conserved more than sequence
  - 3D structure determination is difficult, slow and expensive
  - Intellectual challenge, Nobel prizes etc...
  - Engineering new proteins

The Use of Structure

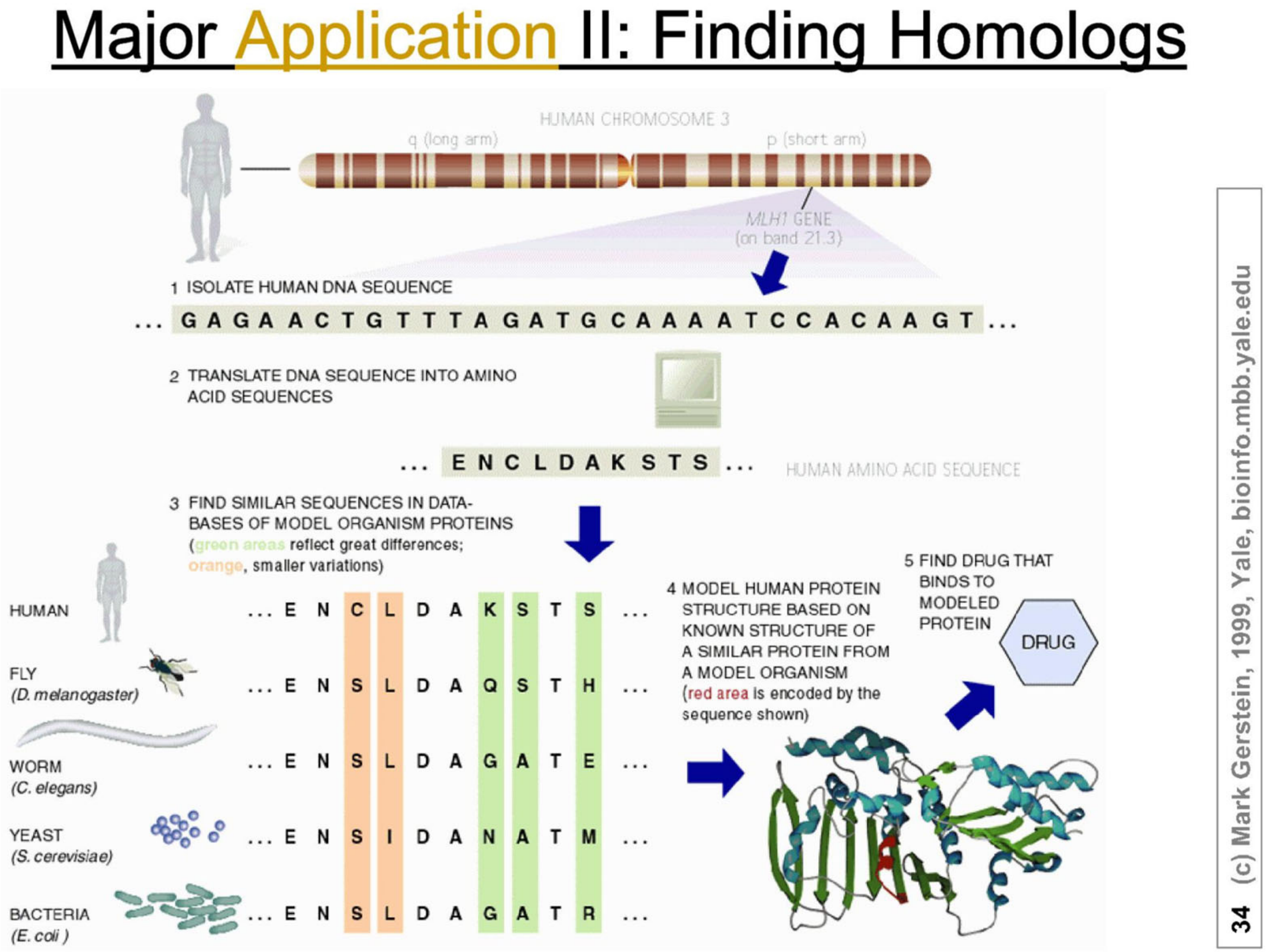
Major Application I: Designing Drugs

- 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).



The Use of Structure



Major Application II: Finding Homologs

1 ISOLATE HUMAN DNA SEQUENCE

2 TRANSLATE DNA SEQUENCE INTO AMINO ACID SEQUENCES

3 FIND SIMILAR SEQUENCES IN DATA-BASES OF MODEL ORGANISM PROTEINS (green areas reflect great differences; orange, smaller variations)

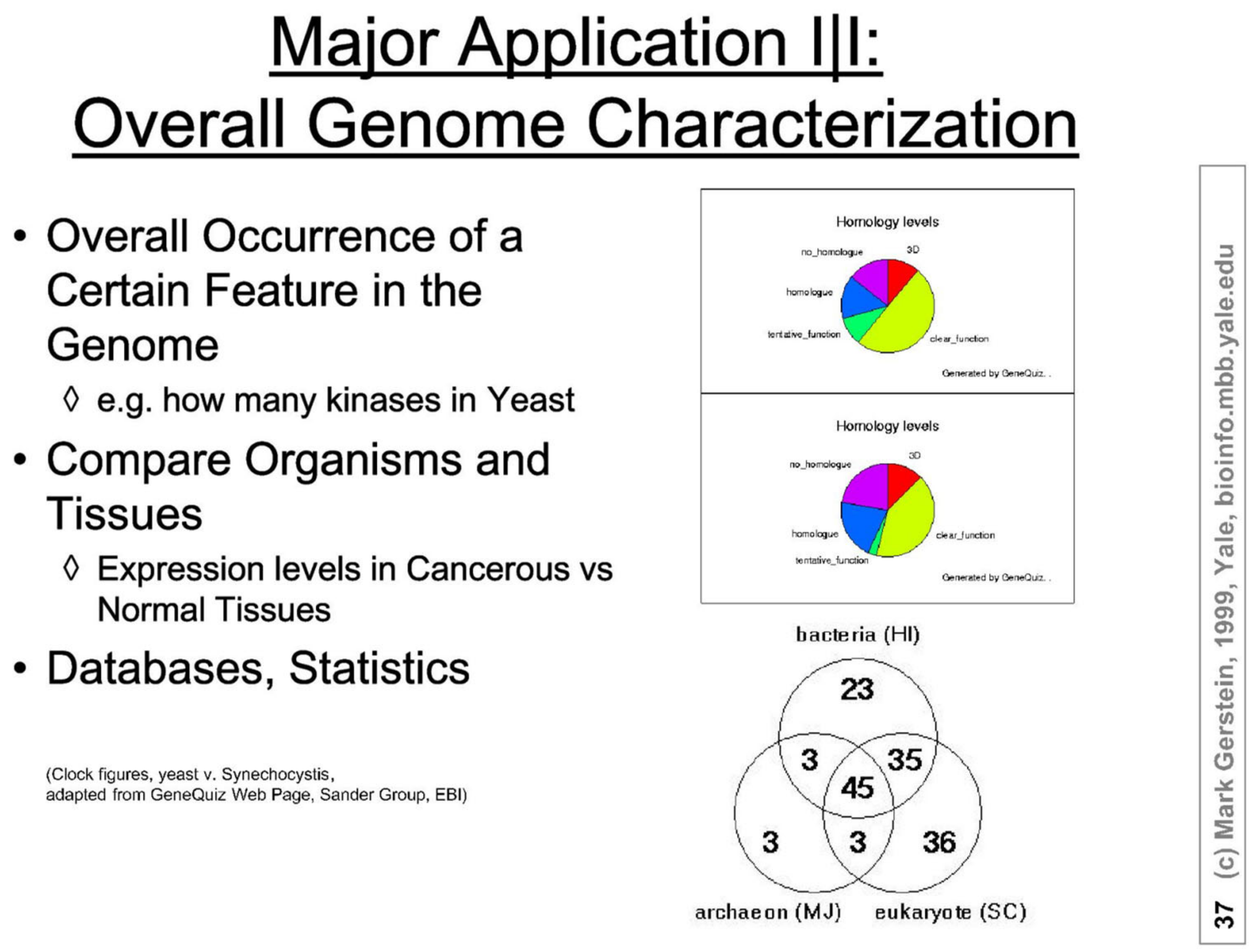
4 MODEL HUMAN PROTEIN STRUCTURE BASED ON KNOWN STRUCTURE OF A SIMILAR PROTEIN FROM A MODEL ORGANISM (red area is encoded by the sequence shown)

5 FIND DRUG THAT BINDS TO MODELED PROTEIN

HUMAN CHROMOSOME 3
q (long arm)
p (short arm)
MLH1 GENE
(on band 21.3)
... G A G A A C T G T T T A G A T G C A A A A T C C A C A A G T ...
... E N C L D A K S T S ...
HUMAN AMINO ACID SEQUENCE
HUMAN ... E N C L D A K S T S ...
FLY (D. melanogaster) ... E N S L D A Q S T H ...
WORM (C. elegans) ... E N S L D A G A T E ...
YEAST (S. cerevisiae) ... E N S I D A N A T M ...
BACTERIA (E. coli) ... E N S L D A G A T R ...
DRUG

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

The Use of Structure



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

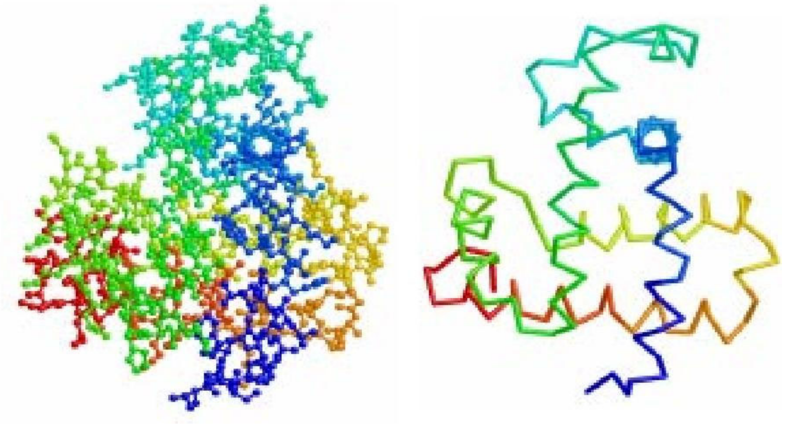
(Clock figures, yeast v. Synechocystis, adapted from GeneQuiz Web Page, Sander Group, EBI)

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

It's not that simple...

- Amino acid sequence contains all the information for 3D structure (experiments of Anfinsen, 1970's)
- But, there are thousands of atoms, rotatable bonds, solvent and other molecules to deal with...
- Levinthal's paradox

Sperm Whale Myoglobin

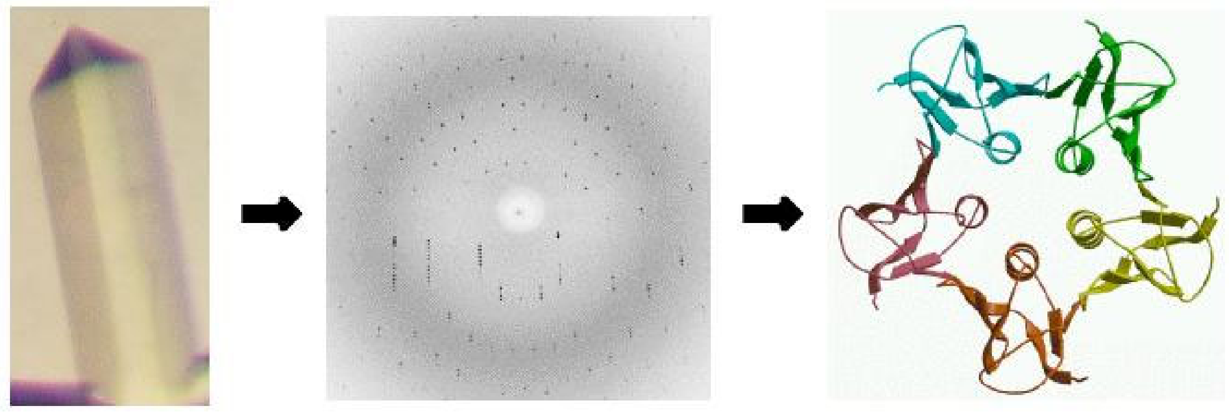


Empirical structure determination

- Two major experimental methods for determining protein structure
- X-ray Crystallography
  - Requires growing a crystal of the protein (impossible for some, never easy)
  - Diffraction pattern can be inverse-Fourier transformed to characterize electron densities (Phase problem)
- Nuclear Magnetic Resonance (NMR) imaging
  - Provides distance constraints, but can be hard to find a corresponding structure
  - No crystal of proteins needed, can observe protein dynamics
  - Works only for relatively small proteins (so far)

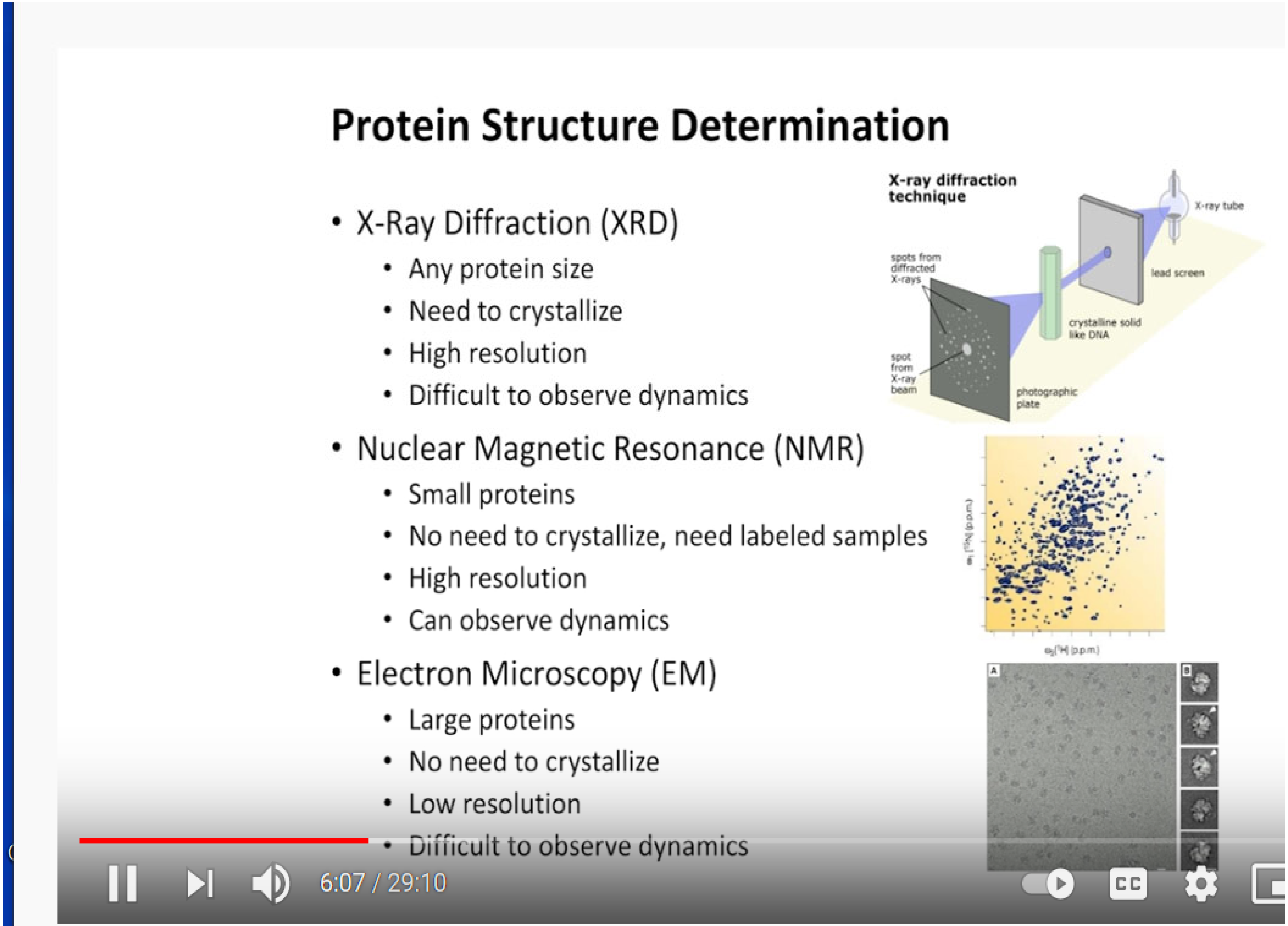
X-ray crystallography

- X-rays, since wavelength is near the distance between bonded carbon atoms
- Maps electron density, not atoms directly
- Crystal to get a lot of spatially aligned atoms
- Have to invert Fourier transform to get structure, but only have amplitudes, not phases



NMR structure determination

- NMR can detect certain features of hydrogen atoms:
  - NOESY measures distances between non-bonded H's within about 5A
  - COSY and TOCSY described relations through bonds
- Combination of distance and angle constraints, plus knowledge of covalent bonds (amino acid sequence) determines a unique (sometimes) structure.
- Overlapping measurement limits size ~120AA



Protein Structure Determination

- X-Ray Diffraction (XRD)
  - Any protein size
  - Need to crystallize
  - High resolution
  - Difficult to observe dynamics
- Nuclear Magnetic Resonance (NMR)
  - Small proteins
  - No need to crystallize, need labeled samples
  - High resolution
  - Can observe dynamics
- Electron Microscopy (EM)
  - Large proteins
  - No need to crystallize
  - Low resolution
  - Difficult to observe dynamics

https://www.nature.com/articles/d41586-020-00341-9

Why predict protein structure?

- Neither crystallography nor NMR can keep pace with genome sequencing efforts
  - Only 10566 (3641 with <90% identity) human proteins in PDB, although growing fast
  - Computer scientists love this problem
  - Understandable with minimal biology
  - Seems like a good discrimination task
- Understand the mechanisms of folding (?)

Protein Structure Classification – SCOP2

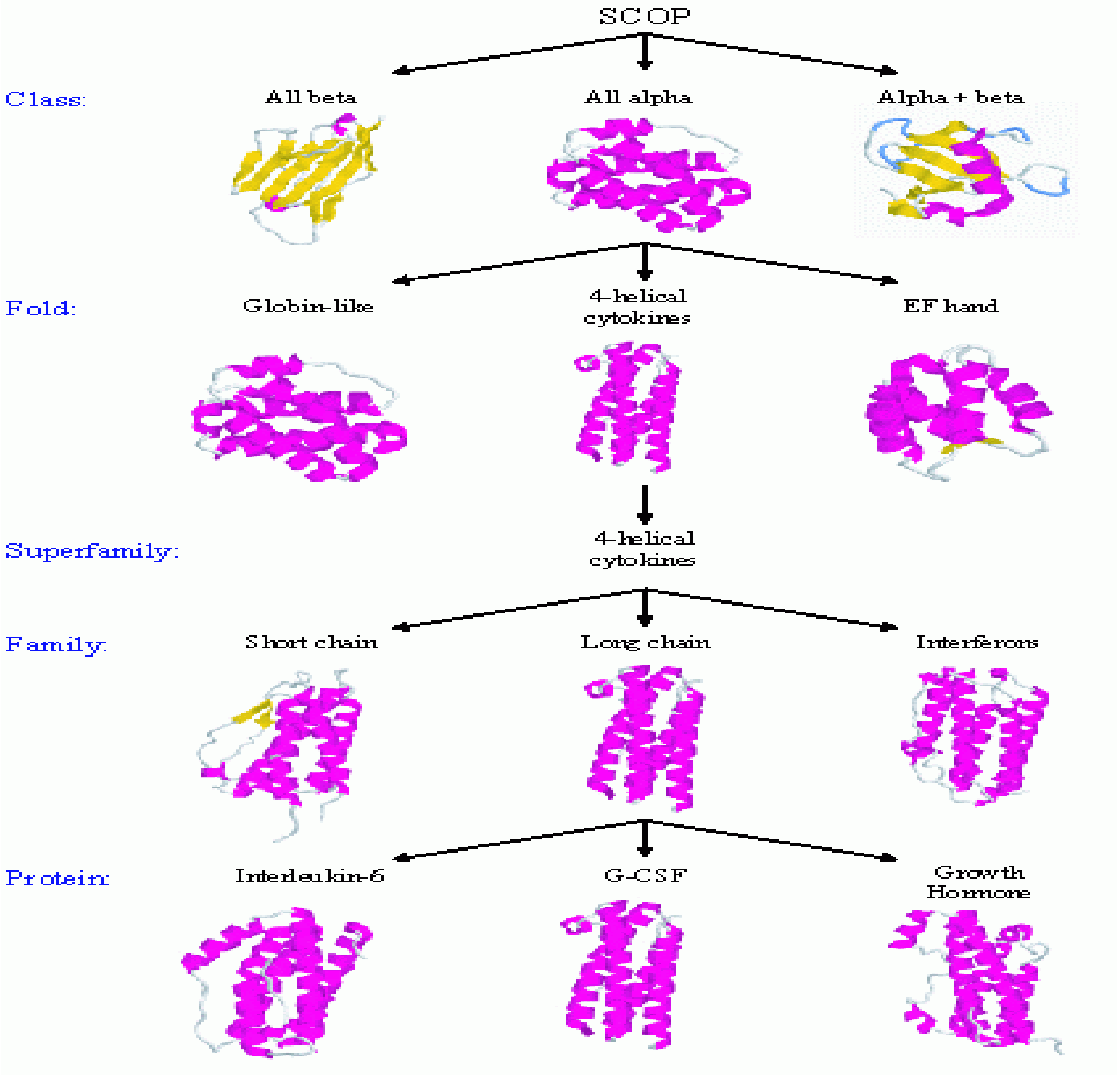
- Structure Classification Of Proteins database https://scop2.mrc-lmb.cam.ac.uk/

- Hierarchical Clustering
  - Family – clear evolutionarily relationship
  - Superfamily – probable common evolutionary origin
  - Fold – major structural similarity
  - Class– common structural component

- Boundaries between levels are more or less subjective

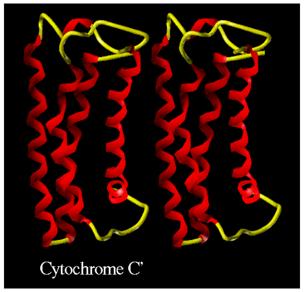
- Conservative evolutionary classification leads to many new divisions at the family and superfamily levels, therefore it is recommended to first focus on higher levels in the classification tree.

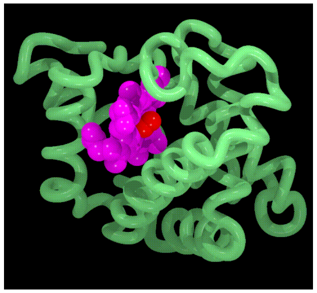
Protein Structure Classification - SCOP



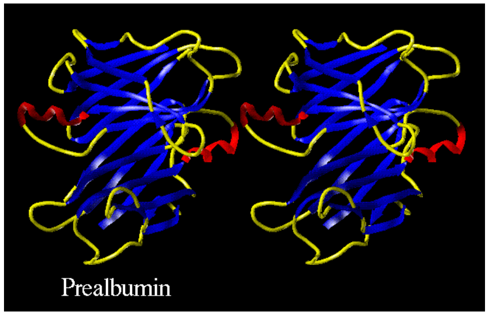
Protein Structure Classification - SCOP

- α/α

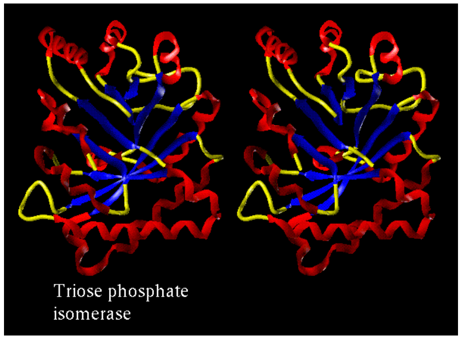




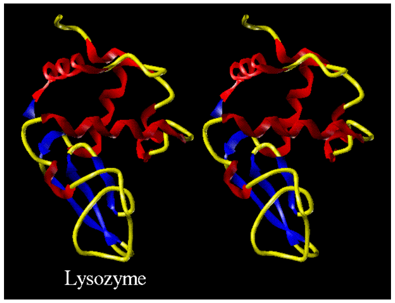
- β/β



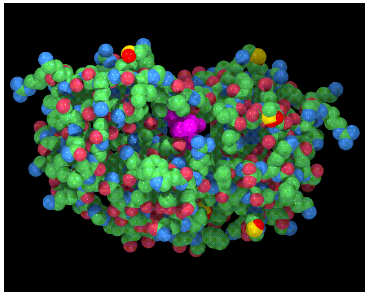
- α/β



- α+β



- **Misc**



Protein Structure Classification - SCOP

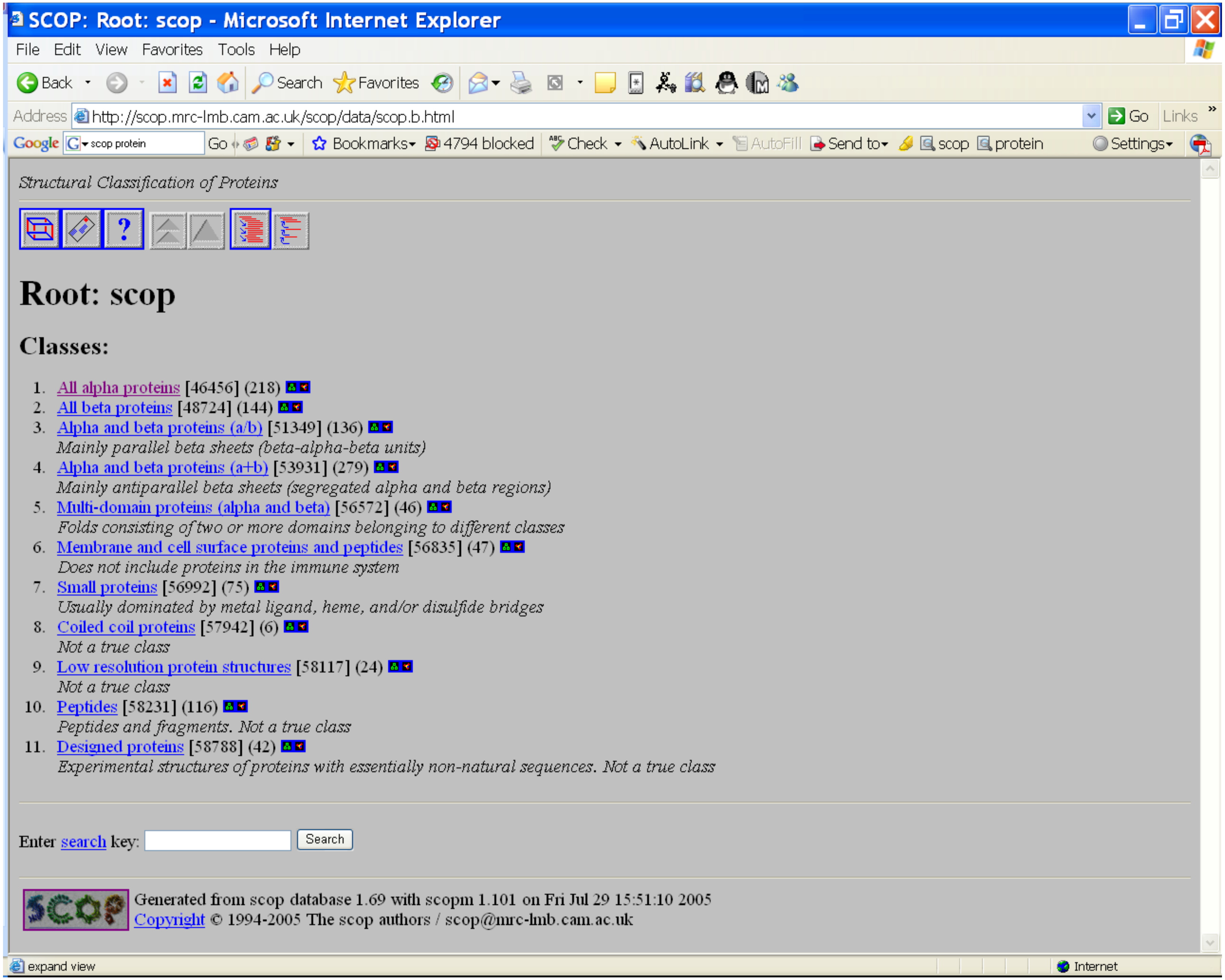
Scop Classification Statistics

SCOP: Structural Classification of Proteins. 1.75 release

38221 PDB Entries. 1 Literature Reference. 110800 Domains. (excluding nucleic acids and theoretical models).

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	284	507	871
All beta proteins	174	354	742
Alpha and beta proteins (a/b)	147	244	803
Alpha and beta proteins (a+b)	376	552	1055
Multi-domain proteins	66	66	89
Membrane and cell surface proteins	58	110	123
Small proteins	90	129	219
Total	1195	1962	3902

Protein Structure Classification - SCOP



SCOP: Root: scop - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html
Go Links
Google scop protein Go Bookmarks 4794 blocked Check AutoLink AutoFill Send to scop protein Settings
expand view
Internet

Structural Classification of Proteins

Root: scop

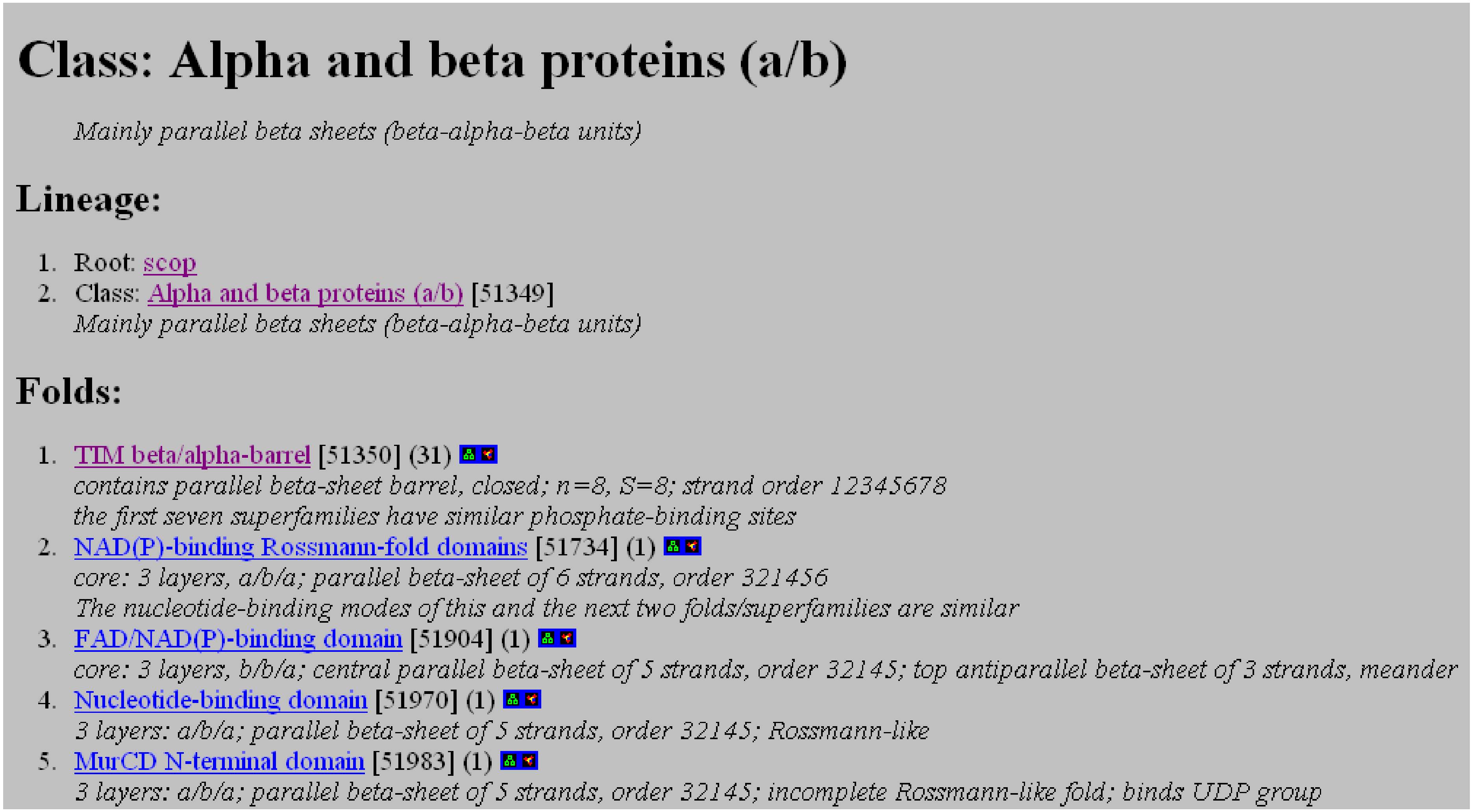
Classes:

1. All alpha proteins [46456] (218)
2. All beta proteins [48724] (144)
3. Alpha and beta proteins (a/b) [51349] (136)
   Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) [53931] (279)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) [56572] (46)
   Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides [56835] (47)
   Does not include proteins in the immune system
7. Small proteins [56992] (75)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins [57942] (6)
   Not a true class
9. Low resolution protein structures [58117] (24)
   Not a true class
10. Peptides [58231] (116)
    Peptides and fragments. Not a true class
11. Designed proteins [58788] (42)
    Experimental structures of proteins with essentially non-natural sequences. Not a true class

Enter search key: Search

Generated from scop database 1.69 with scopm 1.101 on Fri Jul 29 15:51:10 2005
Copyright © 1994-2005 The scop authors / scop@mrc-lmb.cam.ac.uk

Protein Structure Classification - SCOP



Class: Alpha and beta proteins (a/b)

Mainly parallel beta sheets (beta-alpha-beta units)

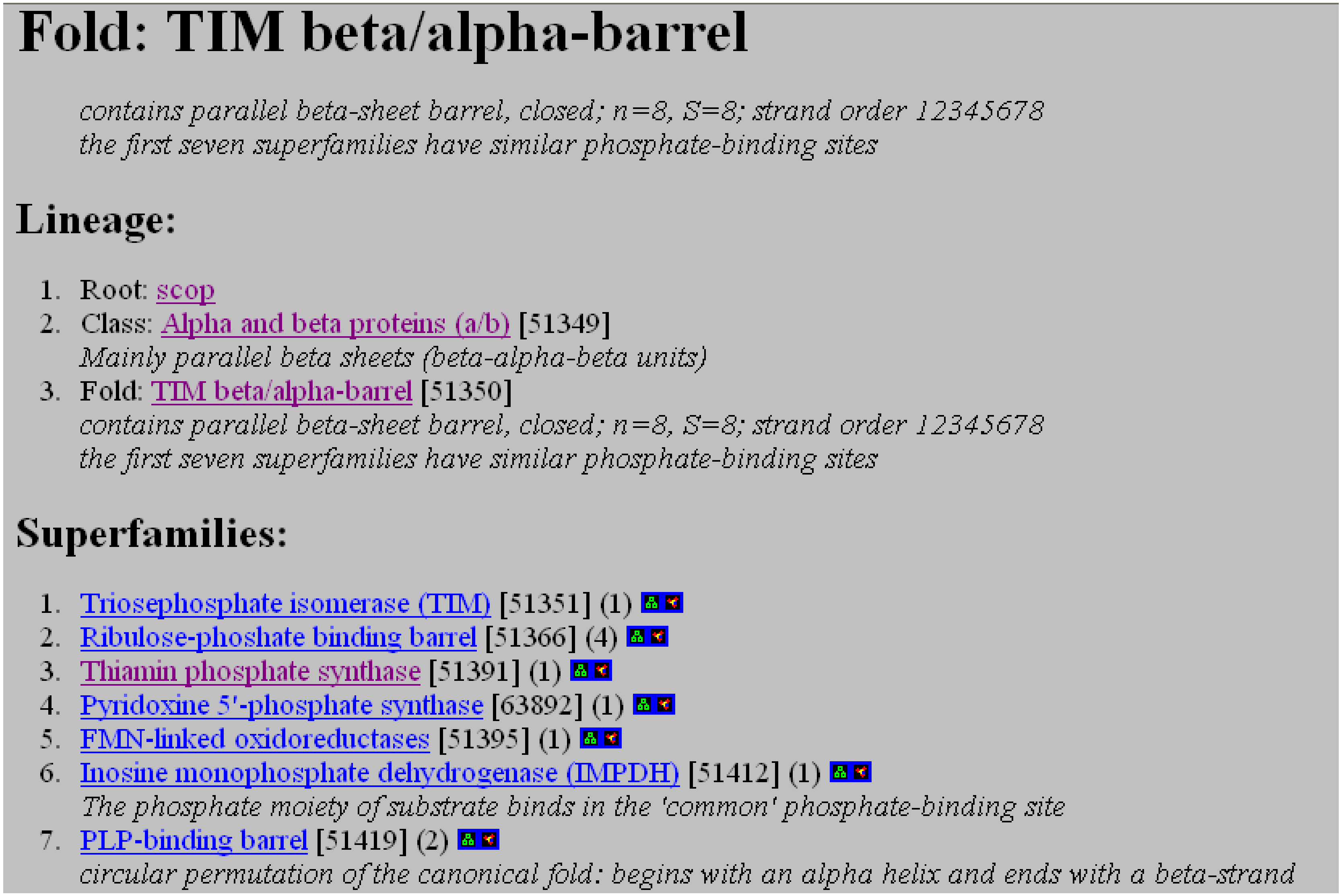
Lineage:

1. Root: scop
2. Class: Alpha and beta proteins (a/b) [51349]
   Mainly parallel beta sheets (beta-alpha-beta units)

Folds:

1. TIM beta/alpha-barrel [51350] (31)
   contains parallel beta-sheet barrel, closed; n=8, S=8; strand order 12345678
   the first seven superfamilies have similar phosphate-binding sites
2. NAD(P)-binding Rossmann-fold domains [51734] (1)
   core: 3 layers, a/b/a; parallel beta-sheet of 6 strands, order 321456
   The nucleotide-binding modes of this and the next two folds/superfamilies are similar
3. FAD/NAD(P)-binding domain [51904] (1)
   core: 3 layers, b/b/a; central parallel beta-sheet of 5 strands, order 32145; top antiparallel beta-sheet of 3 strands, meander
4. Nucleotide-binding domain [51970] (1)
   3 layers: a/b/a; parallel beta-sheet of 5 strands, order 32145; Rossmann-like
5. MurCD N-terminal domain [51983] (1)
   3 layers: a/b/a; parallel beta-sheet of 5 strands, order 32145; incomplete Rossmann-like fold; binds UDP group

Protein Structure Classification - SCOP



Fold: TIM beta/alpha-barrel

contains parallel beta-sheet barrel, closed; n=8, S=8; strand order 12345678
the first seven superfamilies have similar phosphate-binding sites

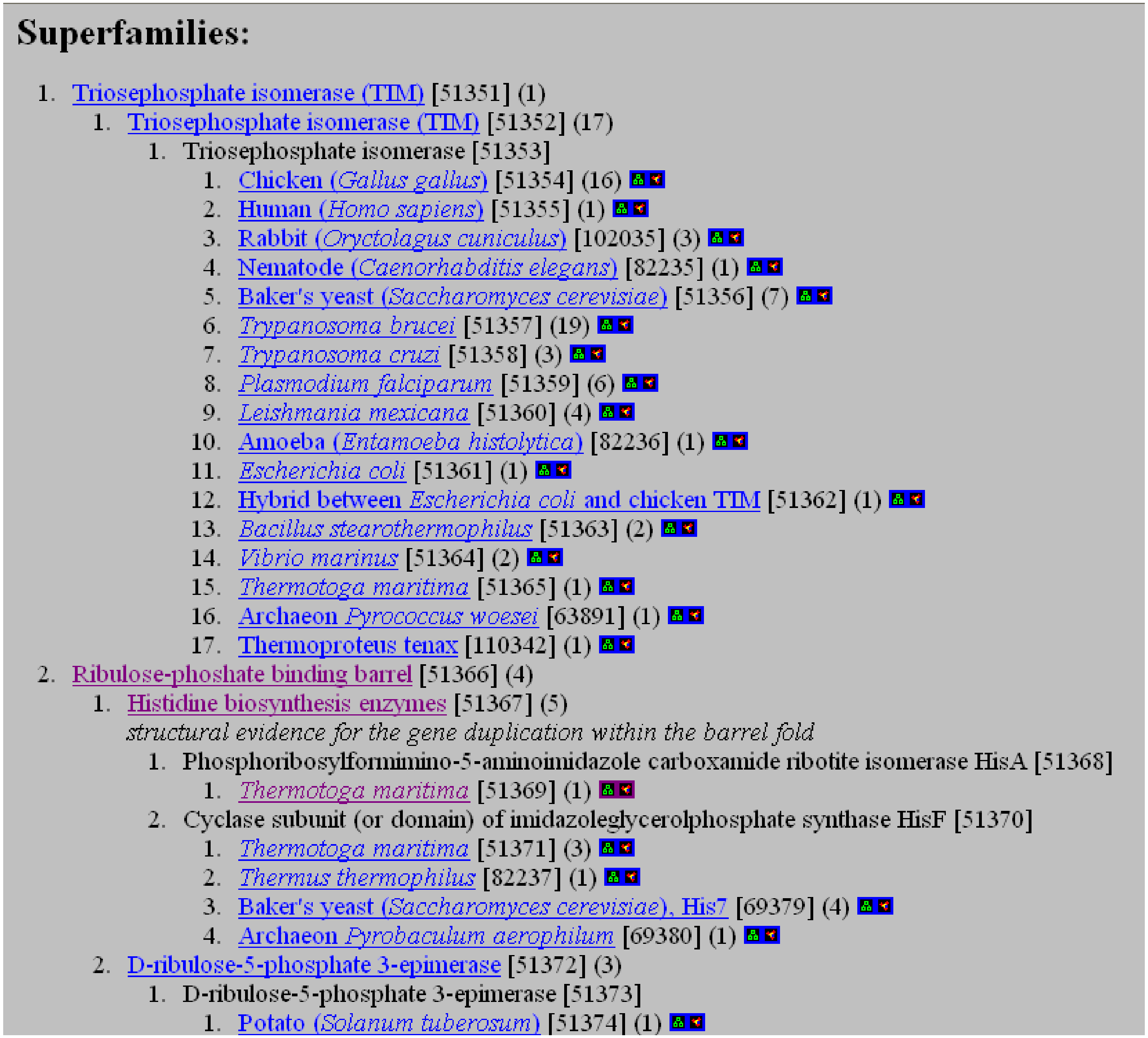
Lineage:

1. Root: scop
2. Class: Alpha and beta proteins (a/b) [51349]
   Mainly parallel beta sheets (beta-alpha-beta units)
3. Fold: TIM beta/alpha-barrel [51350]
   contains parallel beta-sheet barrel, closed; n=8, S=8; strand order 12345678
   the first seven superfamilies have similar phosphate-binding sites

Superfamilies:

1. Triosephosphate isomerase (TIM) [51351] (1)
2. Ribulose-phoshate binding barrel [51366] (4)
3. Thiamin phosphate synthase [51391] (1)
4. Pyridoxine 5'-phosphate synthase [63892] (1)
5. FMN-linked oxidoreductases [51395] (1)
6. Inosine monophosphate dehydrogenase (IMPDH) [51412] (1)
   The phosphate moiety of substrate binds in the 'common' phosphate-binding site
7. PLP-binding barrel [51419] (2)
   circular permutation of the canonical fold: begins with an alpha helix and ends with a beta-strand

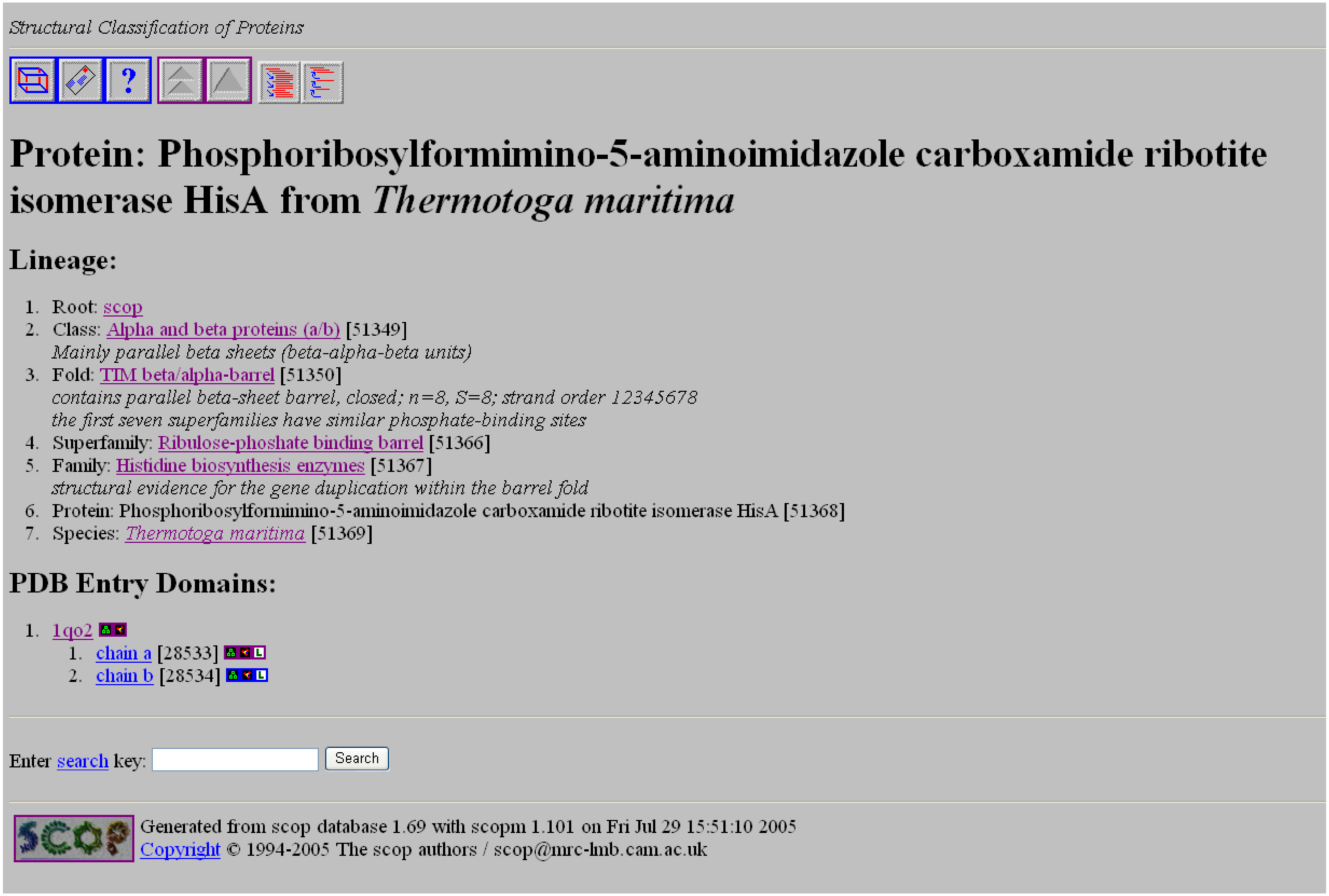
Protein Structure Classification - SCOP



Superfamilies:

1. Triosephosphate isomerase (TIM) [51351] (1)
    1. Triosephosphate isomerase (TIM) [51352] (17)
        1. Triosephosphate isomerase [51353]
            1. Chicken (*Gallus gallus*) [51354] (16)
            2. Human (*Homo sapiens*) [51355] (1)
            3. Rabbit (*Oryctolagus cuniculus*) [102035] (3)
            4. Nematode (*Caenorhabditis elegans*) [82235] (1)
            5. Baker's yeast (*Saccharomyces cerevisiae*) [51356] (7)
            6. *Trypanosoma brucei* [51357] (19)
            7. *Trypanosoma cruzi* [51358] (3)
            8. *Plasmodium falciparum* [51359] (6)
            9. *Leishmania mexicana* [51360] (4)
            10. Amoeba (*Entamoeba histolytica*) [82236] (1)
            11. *Escherichia coli* [51361] (1)
            12. Hybrid between *Escherichia coli* and chicken TIM [51362] (1)
            13. *Bacillus stearothermophilus* [51363] (2)
            14. *Vibrio marinus* [51364] (2)
            15. *Thermotoga maritima* [51365] (1)
            16. Archaeon *Pyrococcus woesei* [63891] (1)
            17. Thermoproteus tenax [110342] (1)
2. Ribulose-phoshate binding barrel [51366] (4)
    1. Histidine biosynthesis enzymes [51367] (5)
       structural evidence for the gene duplication within the barrel fold
        1. Phosphoribosylformimino-5-aminoimidazole carboxamide ribotite isomerase HisA [51368]
            1. *Thermotoga maritima* [51369] (1)
        2. Cyclase subunit (or domain) of imidazoleglycerolphosphate synthase HisF [51370]
            1. *Thermotoga maritima* [51371] (3)
            2. *Thermus thermophilus* [82237] (1)
            3. Baker's yeast (*Saccharomyces cerevisiae*), His7 [69379] (4)
            4. Archaeon *Pyrobaculum aerophilum* [69380] (1)
    2. D-ribulose-5-phosphate 3-epimerase [51372] (3)
        1. D-ribulose-5-phosphate 3-epimerase [51373]
            1. Potato (*Solanum tuberosum*) [51374] (1)

Protein Structure Classification - SCOP



Structural Classification of Proteins

Protein: Phosphoribosylformimino-5-aminoimidazole carboxamide ribotite isomerase HisA from *Thermotoga maritima*

Lineage:

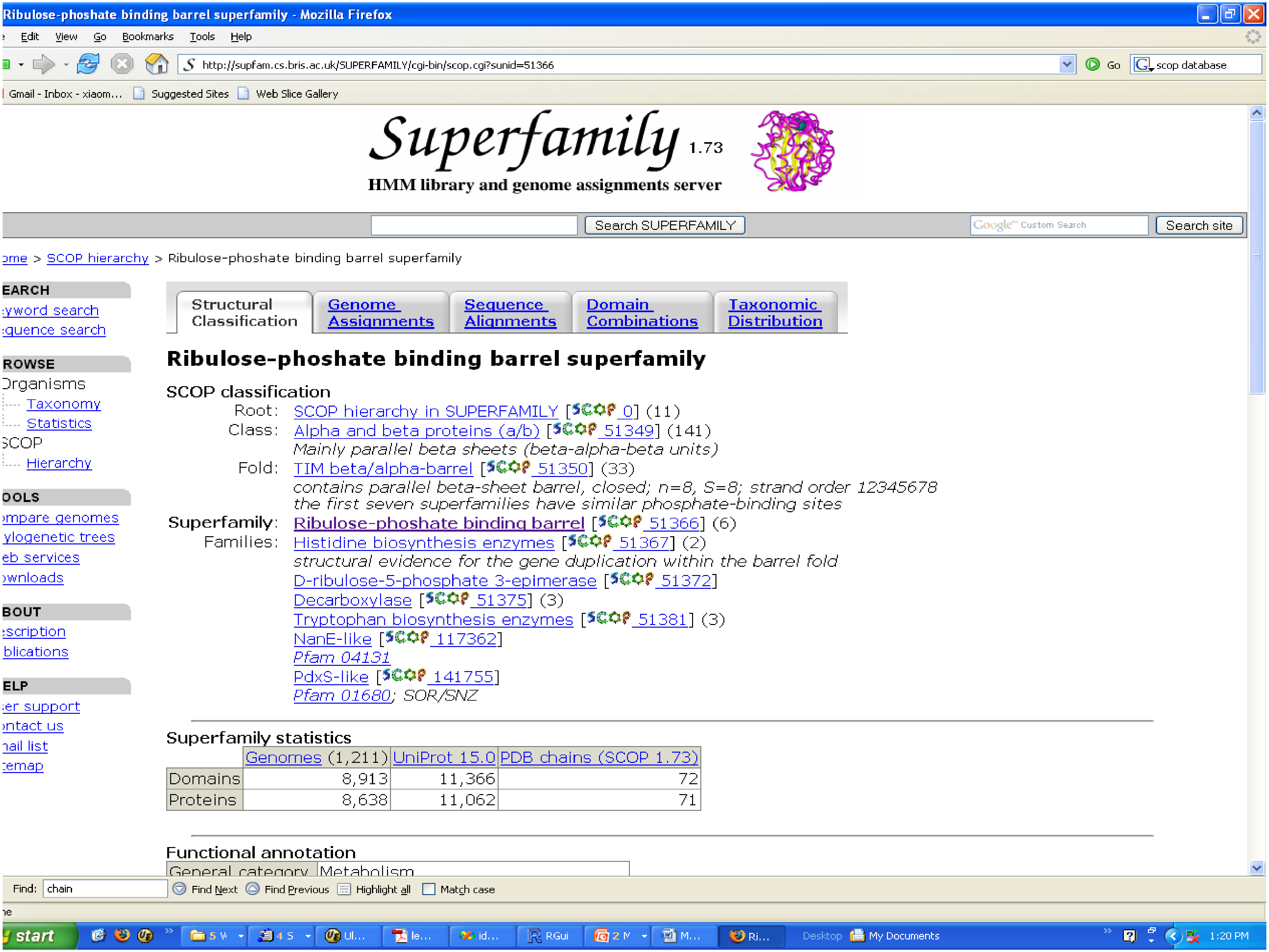
1. Root: scop
2. Class: Alpha and beta proteins (a/b) [51349]
   Mainly parallel beta sheets (beta-alpha-beta units)
3. Fold: TIM beta/alpha-barrel [51350]
   contains parallel beta-sheet barrel, closed; n=8, S=8; strand order 12345678
   the first seven superfamilies have similar phosphate-binding sites
4. Superfamily: Ribulose-phoshate binding barrel [51366]
5. Family: Histidine biosynthesis enzymes [51367]
   structural evidence for the gene duplication within the barrel fold
6. Protein: Phosphoribosylformimino-5-aminoimidazole carboxamide ribotite isomerase HisA [51368]
7. Species: *Thermotoga maritima* [51369]

PDB Entry Domains:

1. 1qo2
   1. chain a [28533] L
   2. chain b [28534] L

Enter search key: Search

Generated from scop database 1.69 with scopm 1.101 on Fri Jul 29 15:51:10 2005
Copyright © 1994-2005 The scop authors / scop@mrc-lmb.cam.ac.uk



ome > SCOP hierarchy > Ribulose-phoshate binding barrel superfamily

Ribulose-phoshate binding barrel superfamily

SCOP classification

Root: SCOP hierarchy in SUPERFAMILY [SCOP 0] (11)

Class: Alpha and beta proteins (a/b) [SCOP 51349] (141)
Mainly parallel beta sheets (beta-alpha-beta units)

Fold: TIM beta/alpha-barrel [SCOP 51350] (33)
contains parallel beta-sheet barrel, closed; n=8, S=8; strand order 12345678
the first seven superfamilies have similar phosphate-binding sites

Superfamily: Ribulose-phoshate binding barrel [SCOP 51366] (6)

Families: Histidine biosynthesis enzymes [SCOP 51367] (2)
structural evidence for the gene duplication within the barrel fold
D-ribulose-5-phosphate 3-epimerase [SCOP 51372]
Decarboxylase [SCOP 51375] (3)
Tryptophan biosynthesis enzymes [SCOP 51381] (3)
NanE-like [SCOP 117362]
Pfam 04131
PdxS-like [SCOP 141755]
Pfam 01680; SOR/SNZ

Superfamily statistics

	Genomes (1,211)	UniProt 15.0	PDB chains (SCOP 1.73)
Domains	8,913	11,366	72
Proteins	8,638	11,062	71

Functional annotation

General category	Metabolism

Structural Classification of Proteins 2

- http://scop2.mrc-lmb.cam.ac.uk/

- SCOP2 is a successor of Structural classification of proteins (SCOP). Similarly to SCOP, the main focus of SCOP2 is on proteins that are structurally characterized and deposited in the PDB. Proteins are organized according to their structural and evolutionary relationships, but, in contrast to SCOP, instead of a simple tree-like hierarchy these relationships form a complex network of nodes. Each node represents a relationship of a particular type and is exemplified by a region of protein structure and sequence.

Relationships in SCOP2

- The relationships in SCOP2 fall into four major categories: Protein types, Evolutionary events, Structural classes and Protein relationships. The first two categories do not have counterparts in SCOP.
- Protein types category groups proteins according to their type as soluble, membrane, fibrous and intrinsically disordered; each type to a large extent correlates with characteristic sequence and structural features.
- Evolutionary events category provides annotation of various structural rearrangements and peculiarities that have been observed amongst related proteins and which have given rise to substantial structural differences.
- Structural classes, organizes protein folds according to their secondary structural content.
- The Protein relationships, consists of three subcategories: Structural, Evolutionary and ‘Other’ relationships.

Protein Structure Classification - CATH

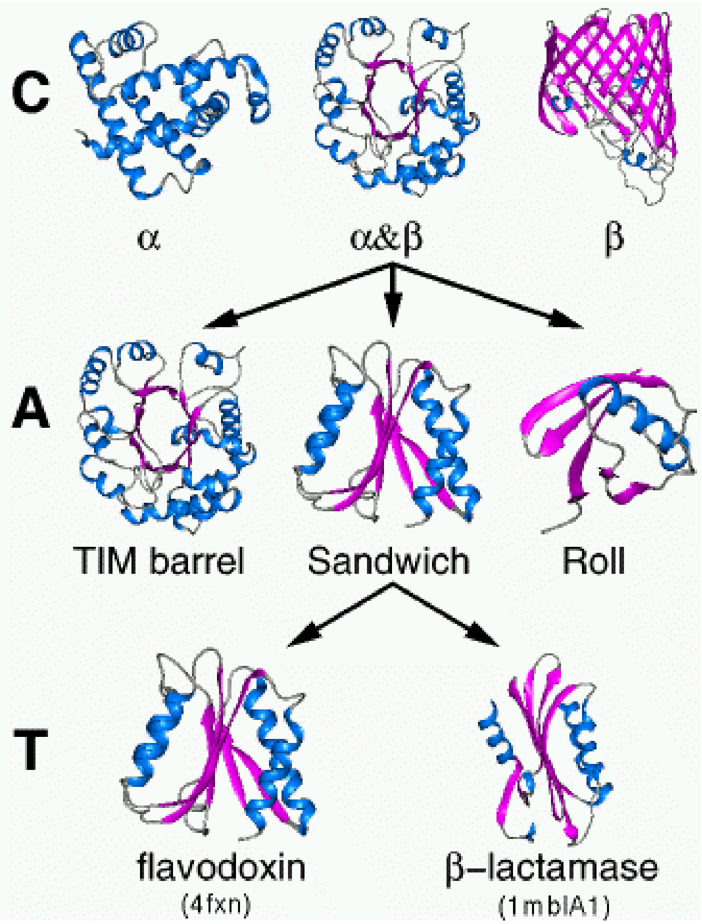
- **CATH Protein Structure Classification**
  - http://www.cathdb.info/

- CATH is a manually curated classification of protein domain structures. Each protein has been chopped into structural domains and assigned into homologous superfamilies (groups of domains that are related by evolution). This classification procedure uses a combination of automated and manual techniques which include computational algorithms, empirical and statistical evidence, literature review and expert analysis.

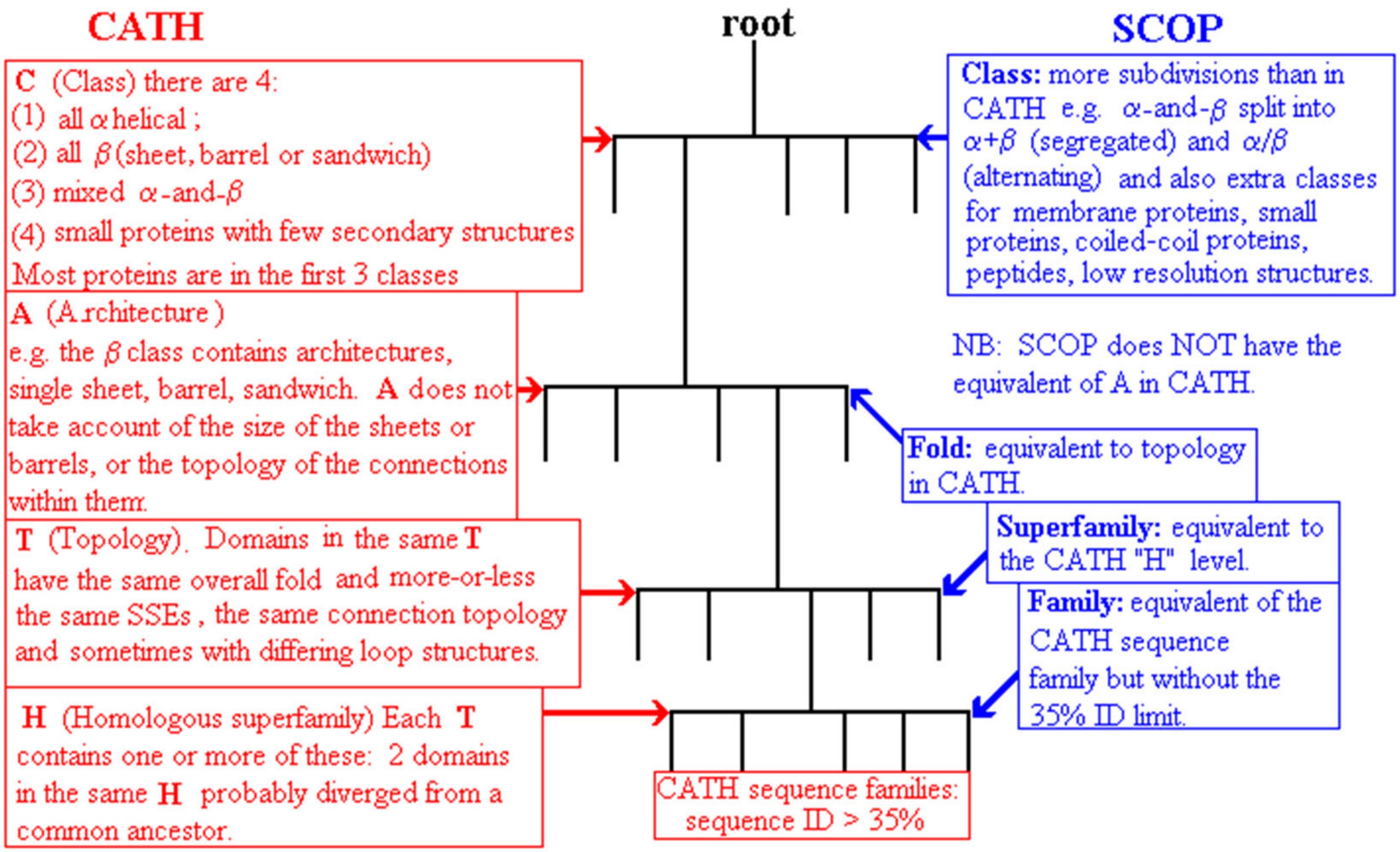
Protein Structure Classification - CATH

- **CATH** is a hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H).
  - Class, derived from secondary structure content, is assigned for more than 90% of protein structures automatically.
  - Architecture, which describes the gross orientation of secondary structures, independent of connectivities, is currently assigned manually.
  - The topology level clusters structures into fold groups according to their topological connections and numbers of secondary structures.
  - The homologous superfamilies cluster proteins with highly similar structures and functions. The assignments of structures to fold groups and homologous superfamilies are made by sequence and structure comparisons.

Protein Structure Classification - CATH



CATH vs. SCOP



CATH

C (Class) there are 4:
(1) all α helical ;
(2) all β (sheet, barrel or sandwich)
(3) mixed α-and-β
(4) small proteins with few secondary structures
Most proteins are in the first 3 classes

A (Architecture)
e.g. the β class contains architectures, single sheet, barrel, sandwich. **A** does not take account of the size of the sheets or barrels, or the topology of the connections within them.

T (Topology). Domains in the same **T** have the same overall fold and more-or-less the same SSEs , the same connection topology and sometimes with differing loop structures.

H (Homologous superfamily) Each **T** contains one or more of these: 2 domains in the same **H** probably diverged from a common ancestor.

CATH sequence families:
sequence ID > 35%

SCOP

Class: more subdivisions than in CATH e.g. α-and-β split into α+β (segregated) and α/β (alternating) and also extra classes for membrane proteins, small proteins, coiled-coil proteins, peptides, low resolution structures.

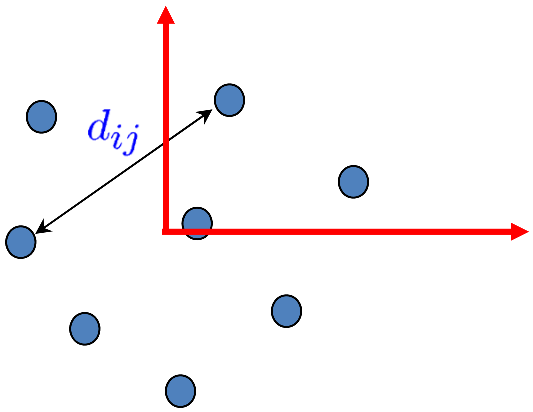
NB: SCOP does NOT have the equivalent of A in CATH.

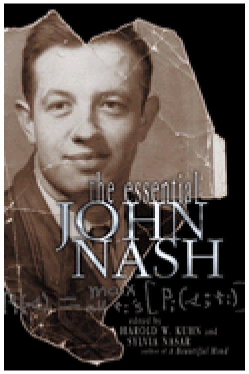
Fold: equivalent to topology in CATH.

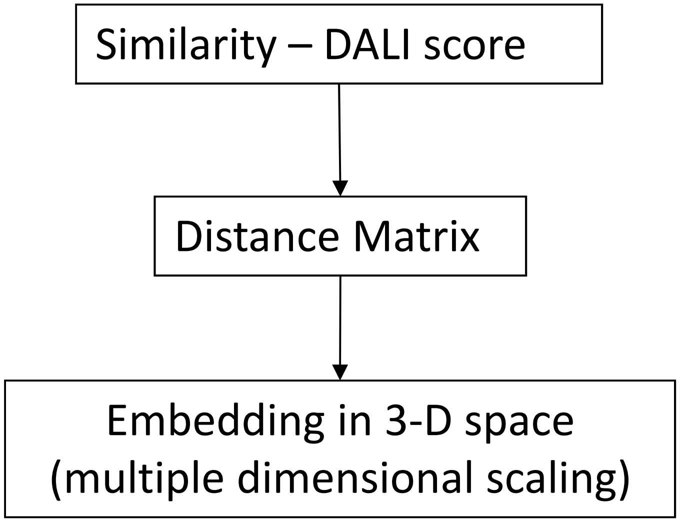
Superfamily: equivalent to the CATH "H" level.

Family: equivalent of the CATH sequence family but without the 35% ID limit.

Protein Fold Space Map



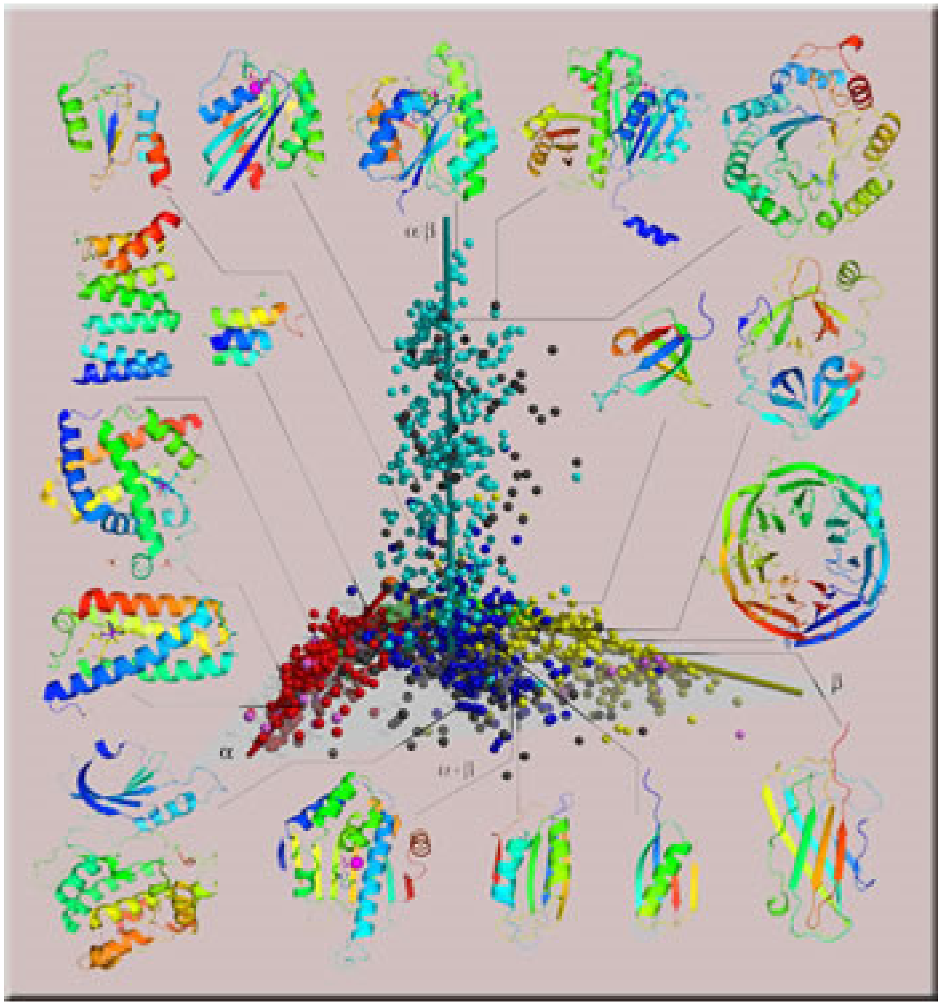




Similarity – DALI score

Distance Matrix

Embedding in 3-D space
(multiple dimensional scaling)



Kim, PNAS, Mar 4, 2003

Structure prediction

Summary of the four main approaches to structure prediction. Note that there are overlaps between nearly all categories.

Method	Knowledge	Approach	Difficulty	Usefulness
Secondary structure prediction	Sequence-structure statistics	Forget 3D arrangement and predict where the helices/strands are	Medium	Can improve alignments, fold recognition, *ab initio*
Comparative modelling (Homology modelling)	Proteins of known structure	Identify related structure with sequence methods, copy 3D coords and modify where necessary	Relatively easy	Very, if sequence identity drug design
Fold recognition	Proteins of known structure	Same as above, but use more sophisticated methods to find related structure	Medium	Limited due to poor models
ab initio tertiary structure prediction	Energy functions, statistics	Simulate folding, or generate lots of structures and try to pick the correct one	Very hard	Not really early time

Secondary Structure Prediction

AGADIR – An algorithm to predict the helical content of peptides
APSSP – Advanced Protein Secondary Structure Prediction Server
GOR – Garnier et al, 1996
HNN – Hierarchical Neural Network method (Guermeur, 1997)
Jpred – A consensus method for protein secondary structure prediction at University of Dundee
JUFO – Protein secondary structure prediction from sequence (neural network)
nnPredict – University of California at San Francisco (UCSF)
Porter – University College Dublin
PredictProtein – PHDsec, PHDacc, PHDhtm, PHDtopology, PHDthreader, MaxHom, EvalSec from Columbia University
Prof – Cascaded Multiple Classifiers for Secondary Structure Prediction
PSA – BioMolecular Engineering Research Center (BMERC) / Boston
PSIpred – Various protein structure prediction methods at Brunel University
SOPMA – Geourjon and Deléage, 1995
SSpro – Secondary structure prediction using bidirectional recurrent neural networks at University of California
DLP – Domain linker prediction at RIKEN

http://us.expasy.org/tools/#secondary

Secondary Structure Prediction – HNN

- http://npsa-pbil.ibcp.fr/cgi-bin/secpred_hnn.pl

-
  ```
  >gi|78099986|sp|P0ABK2|CYDB_ECOLI Cytochrome d ubiquinol oxidase subunit 2
  (Cytochrome d ubiquinol oxidase subunit II) (Cytochrome bd-I oxidase
  subunit II)
  MIDYEVLRFIWWLLVGVLLIGFAVTDGFDMGVGMLTRFLGRNDTERRIMINSIAPHWDGNQVWLITAGGA
  LFAAWPMVYAAAFSGFYVAMILVLASLFFRPVGFDYRSKIEETRWRNMWDWGIFIGSFVPPLVIGVAFGN
  LLQGVPFNVDEYLRLYYTGNFFQLLNPFGLLAGVVSVGMIITQGATYLQMRTVGELHLRTRATAQVAALV
  TLVCFALAGVWVMYGIDGYVVKSTMDHYAASNPLNKEVVREAGAWLVNFNNTPILWAIPALGVVLPLLTI
  LTARMDKAAWAFVFSSLTLACIILTAGIAMFPFVMPSSTMMNASLTMWDATSSQLTLNVMTWVAVVLVPIILLY
  TAWCYWKMFGRITKEDIERNTHSLY
  ```

Secondary Structure Prediction – HNN

Sequence length : 379

HNN :

Alpha helix (Hh) : 209 is 55.15%

3_{10} helix (Gg) : 0 is 0.00%

Pi helix (Ii) : 0 is 0.00%

Beta bridge (Bb) : 0 is 0.00%

Extended strand (Ee) : 55 is 14.51%

Beta turn (Tt) : 0 is 0.00%

Bend region (Ss) : 0 is 0.00%

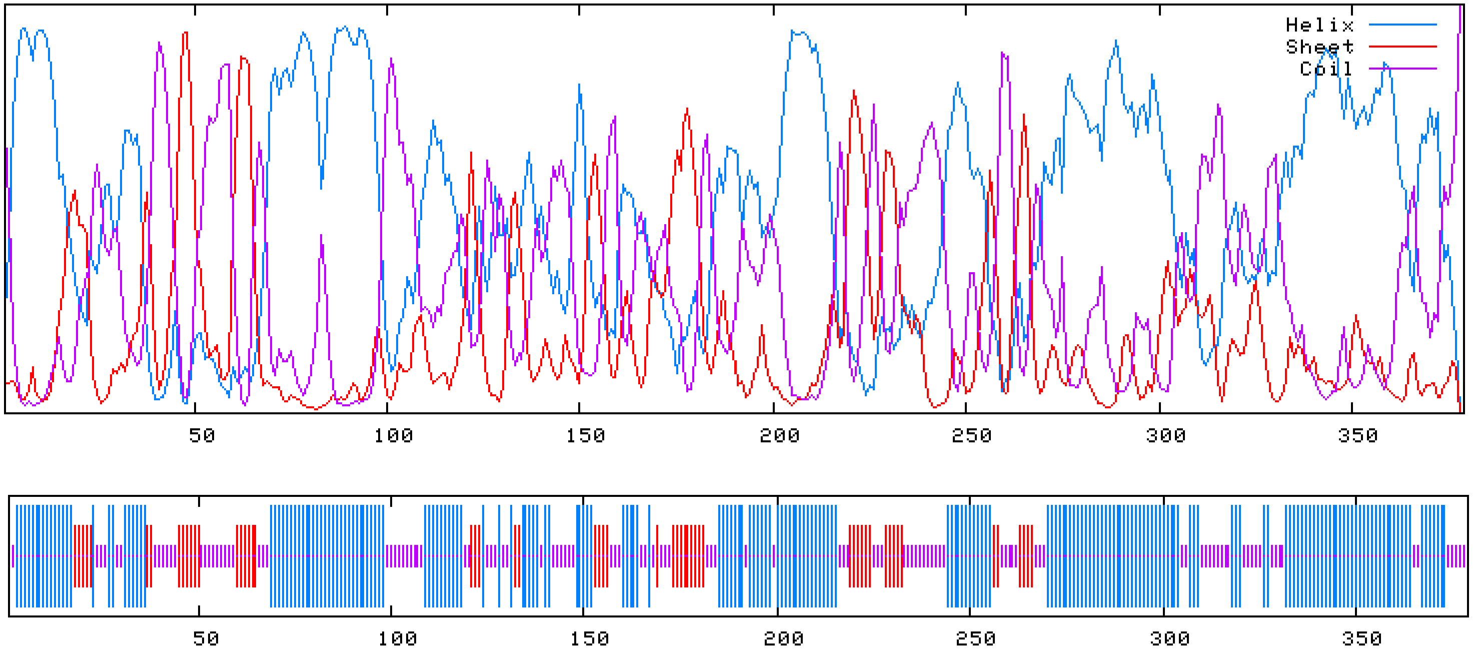
Random coil (Cc) : 115 is 30.34%

Ambigous states (?) : 0 is 0.00%

Other states : 0 is 0.00%

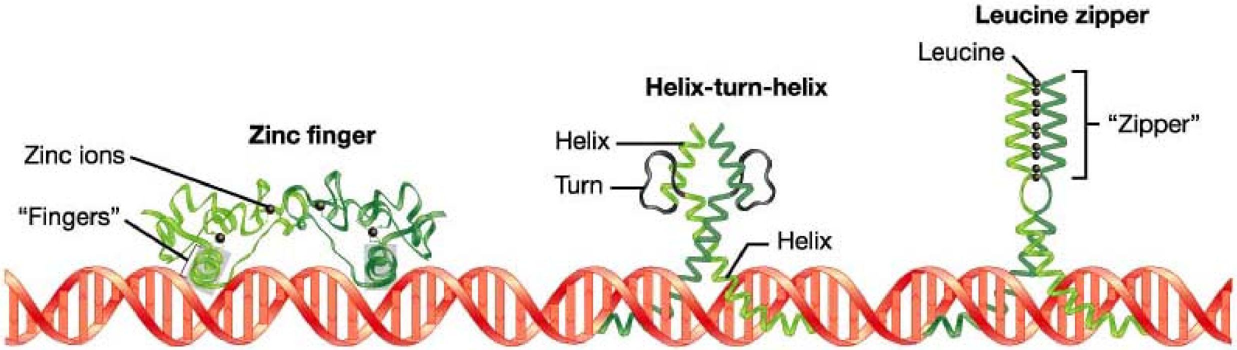
```
        10        20        30        40        50        60        70
         |         |         |         |         |         |         |
MIDYEVLRFIWWLLVGVLLIGFAVTDGFDMGVGMLTRFLGRNDTERRIMINSIAPHWDGNQVWLITAGGA
ccchhhhhhhhhhhhhhheeeeehccchhcchhhhhheecccccceeeeeeccccccccceeeeeeccch
LFAAWPMVYAAAFSGFYVAMILVLASLFFRPVGFDYRSKIEETRWRNMWDWGIFIGSFVPPLVIGVAFGN
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhcccccccccchhhhhhhhhhcceeehccchccheehhhhhc
LLQGVPFNVDEYLRLYYTGNFFQLLNPFGLLAGVVSVGMIITQGATYLQMRTVGELHLRTRATAQVAALV
hhcccccchhhhheeeeeccchhhhcchceccceeeeeeeeeccchhhhhhchhhhhhchhhhhhhhhhh
TLVCFALAGVWVMYGIDGYVVKSTMDHYAASNPLNKEVVREAGAWLVNFNNTPILWAIPALGVVLPLLTI
hhhhhhccceeeeeeccceeeeecccccccccchhhhhhhhhhhhheeccccceeeeccchhhhhhhhhh
LTARMDKAAWAFVFSSLTLACIILTAGIAMFPFVMPSSTMMNASLTMWDATSSQLTLNVMTWVAVVLVPI
hhhhhhhhhhhhhhhhhhhhhhhhhcchhcccccccchhhccccchhccccchhhhhhhhhhhhhhhhhh
ILLYTAWCYWKMFGRITKEDIERNTHSLY
hhhhhhhhhhhhhhhcchhhhhhhccccc
```

Secondary Structure Prediction – HNN



Motifs Readily Identified from Sequence

- Zinc Finger - order and spacing of a pattern for cysteine and histidine.
- Leucine zippers – two antiparallel alpha helices held together by interactions between hybrophobic leucine residues at every seventh position in each helix.
- Coiled coils – 2-3 helices coiled around each other in a left-handed supercoil (3.5 residue/turn instead of 3.6 – 7/two turns); first and fourth are always hydrophobic, others hydrophilic; 5-10 heptads.
- Transmembrane-spanning proteins – alpha helices comprising amino acids with hydrophobic side chains, typically 20-30 residues.



Topology Prediction

PSORT - Prediction of protein subcellular localization
TargetP - Prediction of subcellular location
DAS - Prediction of transmembrane regions in prokaryotes using the Dense Alignment Surface method (Stockholm University)
HMMTOP - Prediction of transmembrane helices and topology of proteins (Hungarian Academy of Sciences)
PredictProtein - Prediction of transmembrane helix location and topology (Columbia University)
SOSUI - Prediction of transmembrane regions (Nagoya University, Japan)
TMAP - Transmembrane detection based on multiple sequence alignment (Karolinska Institut; Sweden)
TMHMM - Prediction of transmembrane helices in proteins (CBS; Denmark)
TMpred - Prediction of transmembrane regions and protein orientation (EMBnet-CH)
TopPred - Topology prediction of membrane proteins (France)

http://us.expasy.org/tools

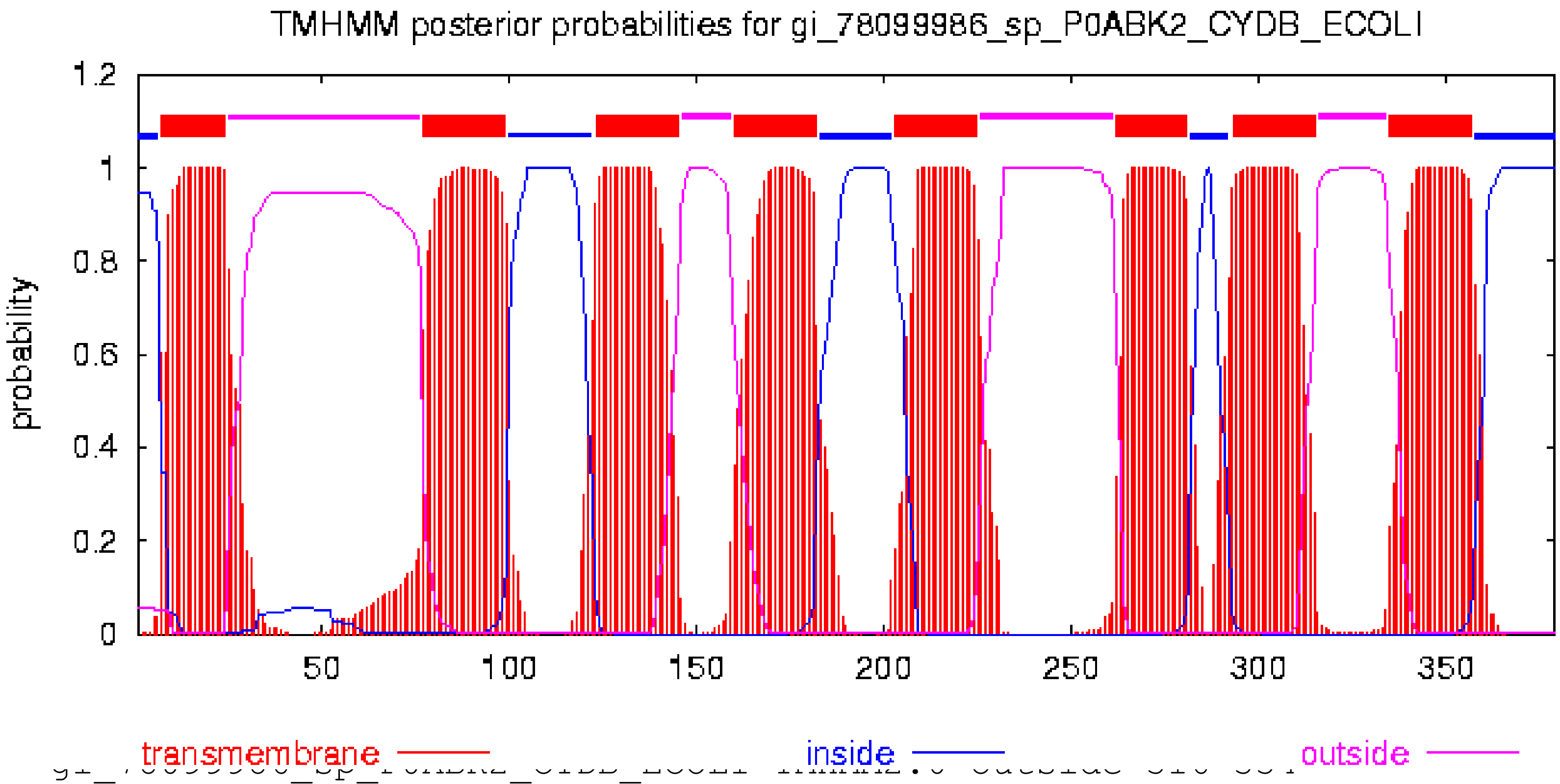
Transmembrane Helix – TMHMM

- http://www.cbs.dtu.dk/services/TMHMM-2.0/

-
```
>gi|78099986|sp|P0ABK2|CYDB_ECOLI Cytochrome d ubiquinol oxidase subunit 2
(Cytochrome d ubiquinol oxidase subunit II) (Cytochrome bd-I oxidase
subunit II)
MIDYEVLRFIWWLLVGVLLIGFAVTDGFDMGVGMLTRFLGRNDTERRIMINSIAPHWDGNQVWLITAGGA
LFAAWPMVYAAAFSGFYVAMILVLASLFFRPVGFDYRSKIEETRWRNMWDWGIFIGSFVPPLVIGVAFGN
LLQGVPFNVDEYLRLYYTGNFFQLLNPFGLLAGVVSVGMIITQGATYLQMRTVGELHLRTRATAQVAALV
TLVCFALAGVWVMYGIDGYVVKSTMDHYAASNPLNKEVVREAGAWLVNFNNTPILWAIPALGVVLPLLTI
LTARMDKAAWAFVFSSLTLACIILTAGIAMFPFVMPSSTMMNASLTMWDATSSQLTLNVMTWVAVVLVPIILLY
TAWCYWKMFGRITKEDIERNTHSLY
```

Transmembrane Helix – TMHMM

```
# gi_78099986_sp_P0ABK2_CYDB_ECOLI Length: 379 #
gi_78099986_sp_P0ABK2_CYDB_ECOLI Number of predicted TMHs: 8 #
gi_78099986_sp_P0ABK2_CYDB_ECOLI Exp number of AAs in TMHs:
```



```
gi_78099986_sp_P0ABK2_CYDB_ECOLI TMHMM2.0 outside 310 334
gi_78099986_sp_P0ABK2_CYDB_ECOLI TMHMM2.0 TMhelix 335 357
gi_78099986_sp_P0ABK2_CYDB_ECOLI TMHMM2.0 inside 358 379
```

Tertiary Structure Prediction

Comparative modeling

SWISS-MODEL – An automated knowledge-based protein modelling server

3Djigsaw – Three-dimensional models for proteins based on homologues of known structure

CPHmodels – Automated neural-network based protein modelling server

ESyPred3D – Automated homology modeling program using neural networks

Geno3d – Automatic modeling of protein three-dimensional structure

SDSC1 – Protein Structure Homology Modeling Server

Threading

3D-PSSM – Protein fold recognition using 1D and 3D sequence profiles coupled with secondary structure information (Foldfit)

Fugue – Sequence-structure homology recognition

HHpred – Protein homology detection and structure prediction by HMM-HMM comparison

Libellula – Neural network approach to evaluate fold recognition results

LOOPP – Sequence to sequence, sequence to structure, and structure to structure alignment

SAM-T02 – HMM-based Protein Structure Prediction

Threader – Protein fold recognition

ProSup – Protein structure superimposition

SWEET – Constructing 3D models of saccharides from their sequences

Ab initio

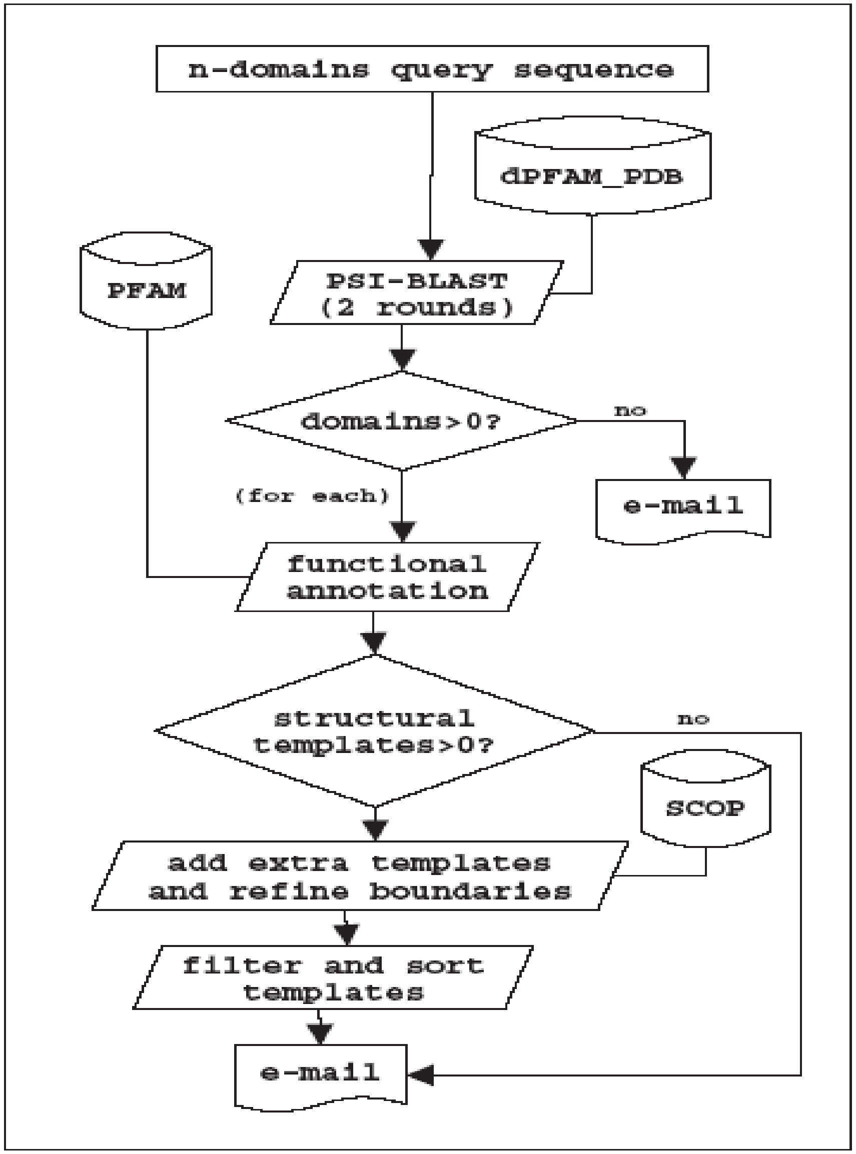
HMMSTR/Rosetta – Prediction of protein structure from sequence

http://us.expasy.org/tools

Tertiary Structure Prediction

Comparative modeling

3Djigsaw – Three-dimensional models for proteins based on homologues of known structure



Contreras-Moreira,B., Bates,P.A. (2002) **Domain Fishing: a first step in protein comparative modelling**. *Bioinformatics* **18**: 1141-1142.

Tertiary Structure Prediction

Threading

3D-PSSM – Protein fold recognition using 1D and 3D sequence profiles coupled with secondary structure information (Foldfit)

Fugue – Sequence-structure homology recognition

HHpred – Protein homology detection and structure prediction by HMM-HMM comparison

Libellula – Neural network approach to evaluate fold recognition results

LOOPP – Sequence to sequence, sequence to structure, and structure to structure alignment

SAM-T02 – HMM-based Protein Structure Prediction

Threader – Protein fold recognition

ProSup – Protein structure superimposition

SWEET – Constructing 3D models of saccharides from their sequences

6th in-class question

Please tell your thoughts about how to choose a research topic based on last lecture.