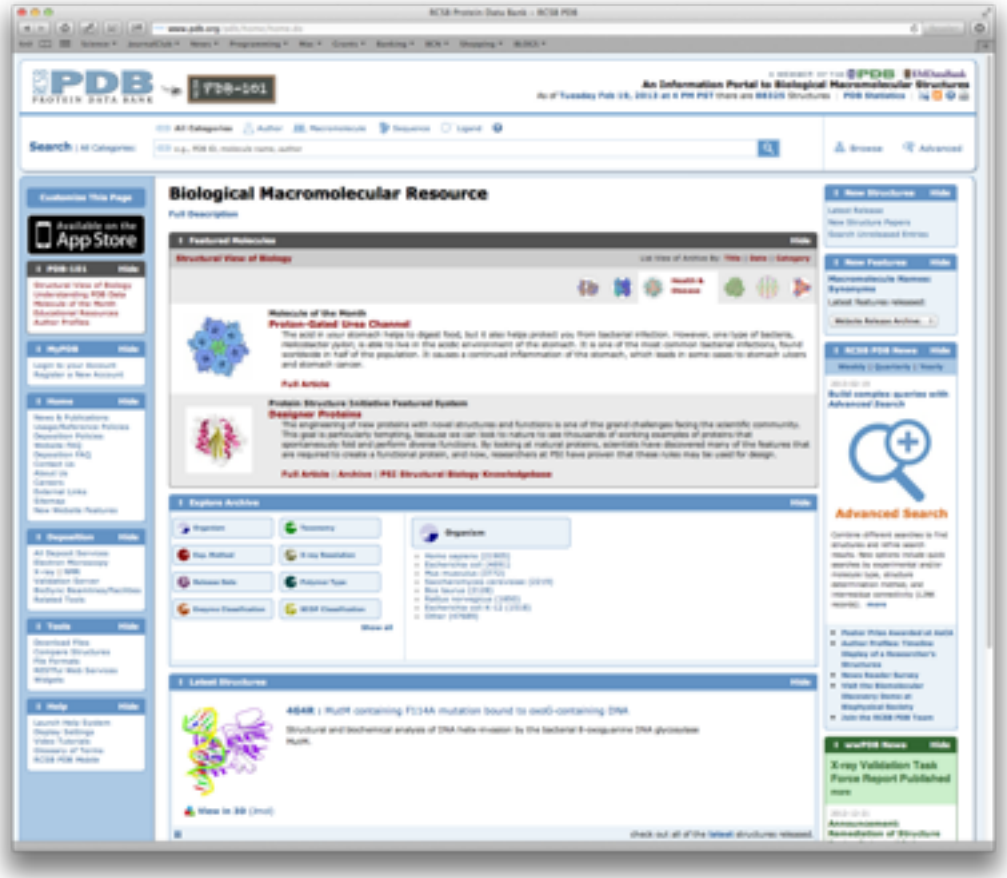
Databases
Alignment & structure classification

GOALS

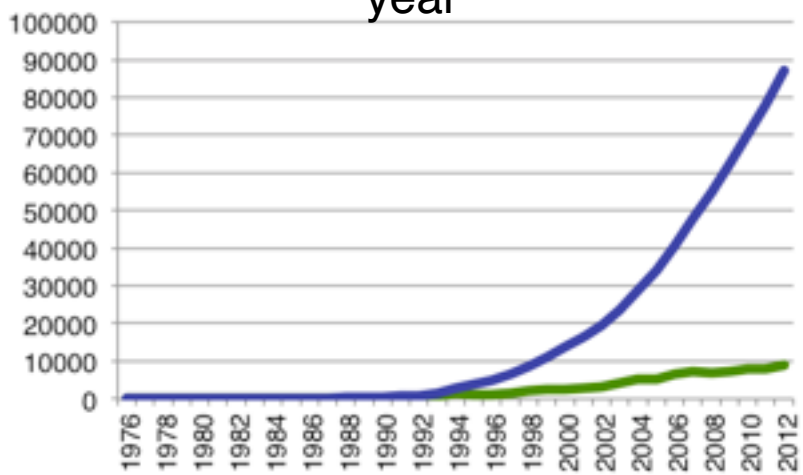
1. Known structures
2. Structure comparison
3. Structure classification
4. Number of folds in nature
5. Sequences VS fold structures

1. Known structures

PDB



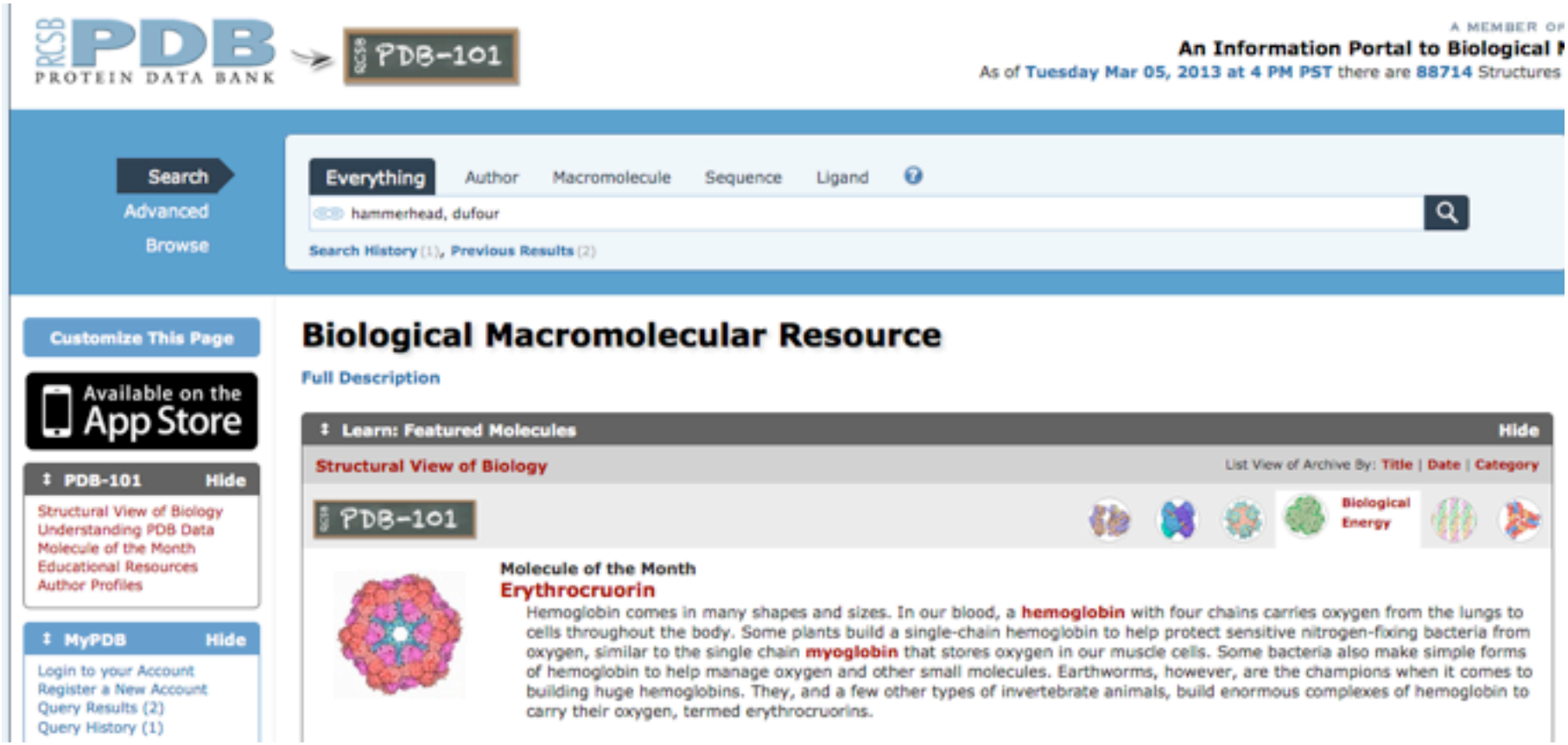
Yearly and total PDB structures per year



Yearly

Total

PDB search


RCSB PDB
PROTEIN DATA BANK
PDB-101
An Information Portal to Biological
As of Tuesday Mar 05, 2013 at 4 PM PST there are 88714 Structures
Search
Advanced
Browse
Everything
Author
Macromolecule
Sequence
Ligand
hammerhead, dufour
Search History (1), Previous Results (2)
Customize This Page
Available on the App Store
PDB-101 Hide
Structural View of Biology
Understanding PDB Data
Molecule of the Month
Educational Resources
Author Profiles
MyPDB Hide
Login to your Account
Register a New Account
Query Results (2)
Query History (1)
Biological Macromolecular Resource
Full Description
Learn: Featured Molecules
Hide
Structural View of Biology
List View of Archive By: Title | Date | Category
PDB-101
Biological Energy
Molecule of the Month
Erythrocruorin
Hemoglobin comes in many shapes and sizes. In our blood, a hemoglobin with four chains carries oxygen from the lungs to cells throughout the body. Some plants build a single-chain hemoglobin to help protect sensitive nitrogen-fixing bacteria from oxygen, similar to the single chain myoglobin that stores oxygen in our muscle cells. Some bacteria also make simple forms of hemoglobin to help manage oxygen and other small molecules. Earthworms, however, are the champions when it comes to building huge hemoglobins. They, and a few other types of invertebrate animals, build enormous complexes of hemoglobin to carry their oxygen, termed erythrocruorins.

PDB search

Showing 1 - 2 of 2 Results

Results : 25 Page: 1 of 1

Filter: Check All View: Detailed Reports: Select one... Sort: ↓ Relevance

2RO2 **Solution structure of domain I of the negative polarity CChMVd hammerhead ribozyme**

Authors: Gallego, J., Dufour, D., Gago, S., de la Pena, M., Flores, R.

Release: 2008-12-30 **Classification:** RNA

Experiment: SOLUTION NMR **Residue Count:** 23

Compound: 1 Polymer [Display Full Polymer Details | Display for All Results]

Citation: **Structure-function analysis of the ribozymes of chrysanthemum chlorotic mottle viroid: a loop-loop interaction motif conserved in most natural hammerheads**
(2009) Nucleic Acids Res. **37:** 368-381 [Display Full Abstract | Display for All Results]

Search Hit: Title: Solution structure of domain I of the negative polarity CChMVd **hammerhead** ribozyme

2RPK **Solution Structure of Domain II of the Positive Polarity CCHMVD Hammerhead Ribozyme**

Authors: Gallego, J., Dufour, D., de la Pena, M., Gago, S., Flores, R.

Release: 2008-12-30 **Classification:** RNA

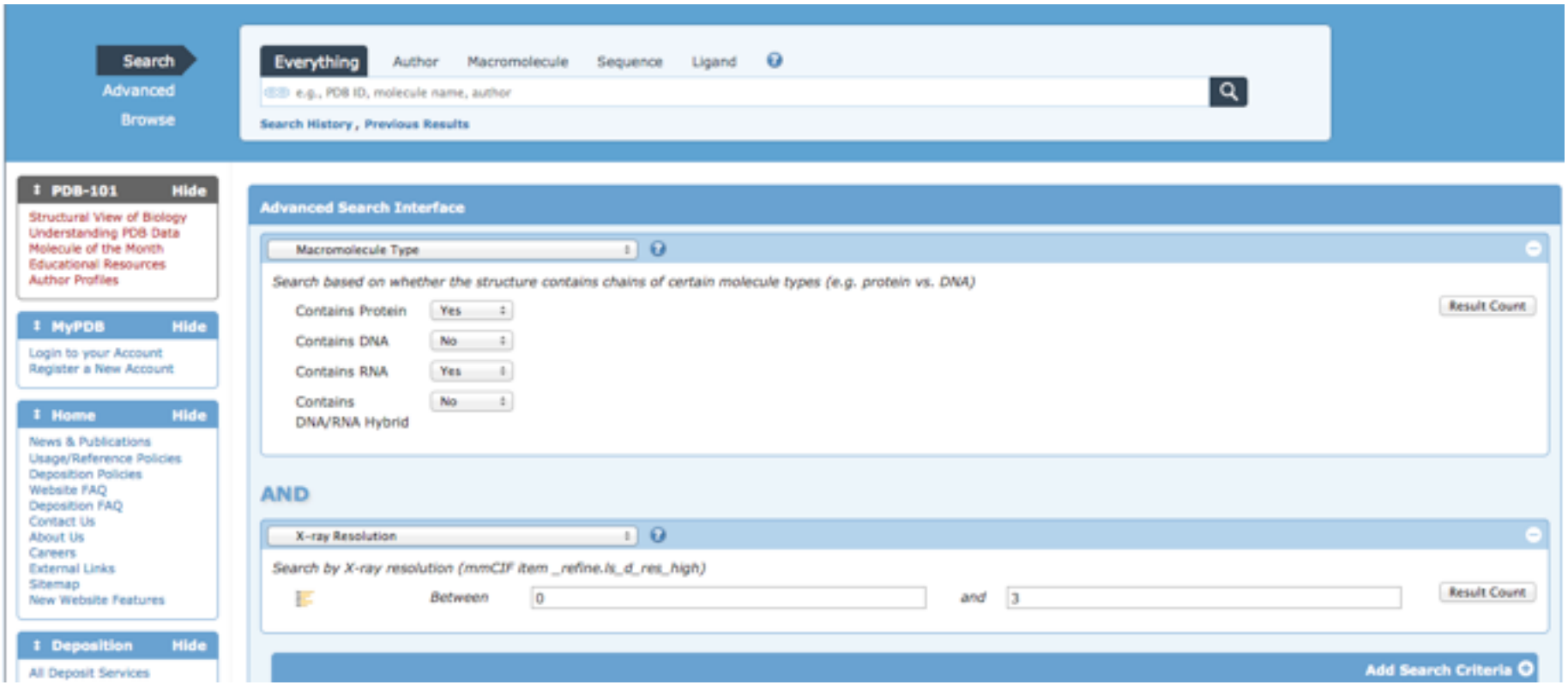
Experiment: SOLUTION NMR **Residue Count:** 20

Compound: 1 Polymer [Display Full Polymer Details | Display for All Results]

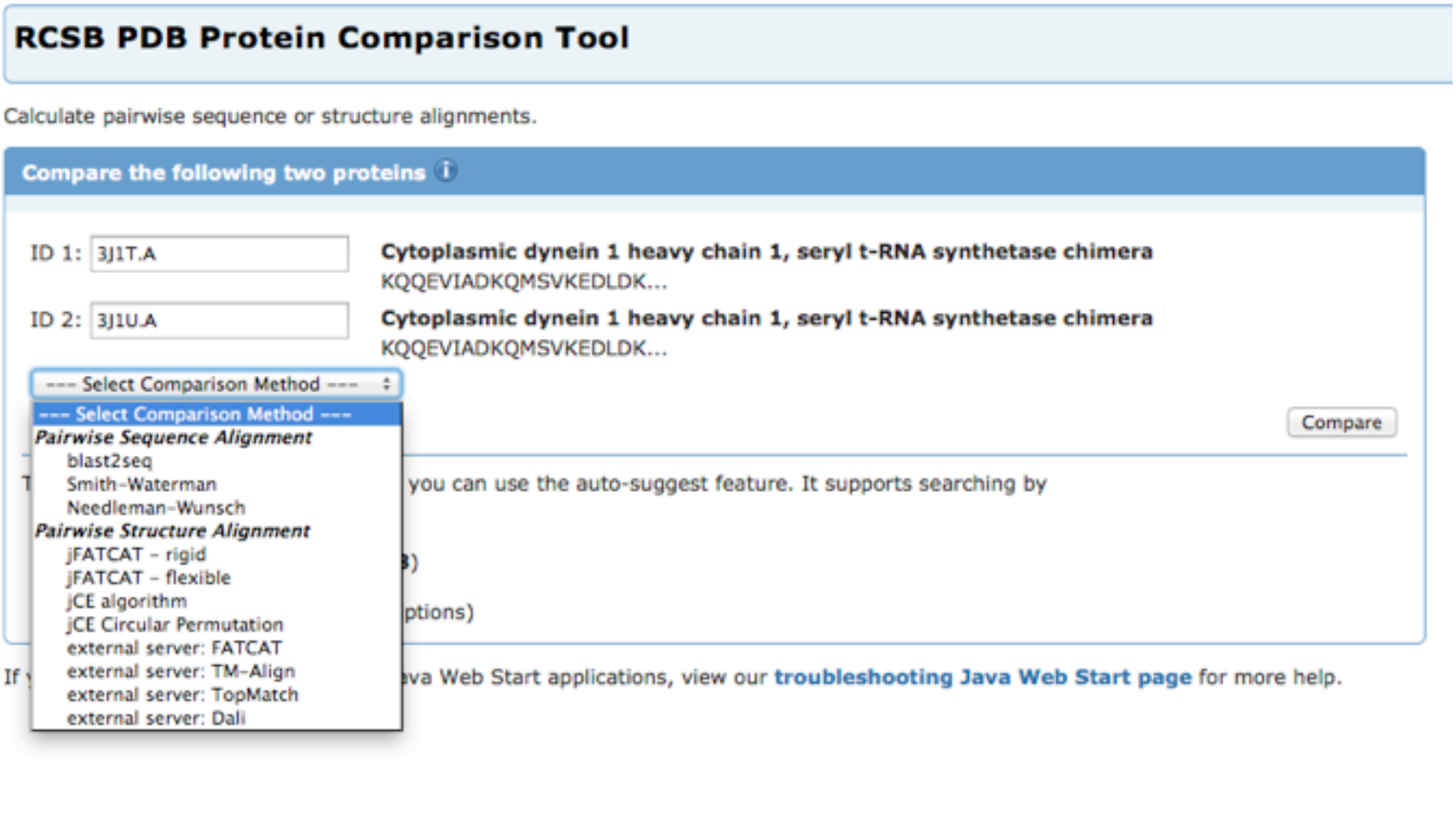
Citation: **Structure-function analysis of the ribozymes of chrysanthemum chlorotic mottle viroid: a loop-loop interaction motif conserved in most natural hammerheads**
(2009) Nucleic Acids Res. **37:** 368-381 [Display Full Abstract | Display for All Results]

Search Hit: Title: Solution Structure of Domain II of the Positive Polarity CCHMVD **Hammerhead** Ribozyme

Advanced search



PDB comparison tool

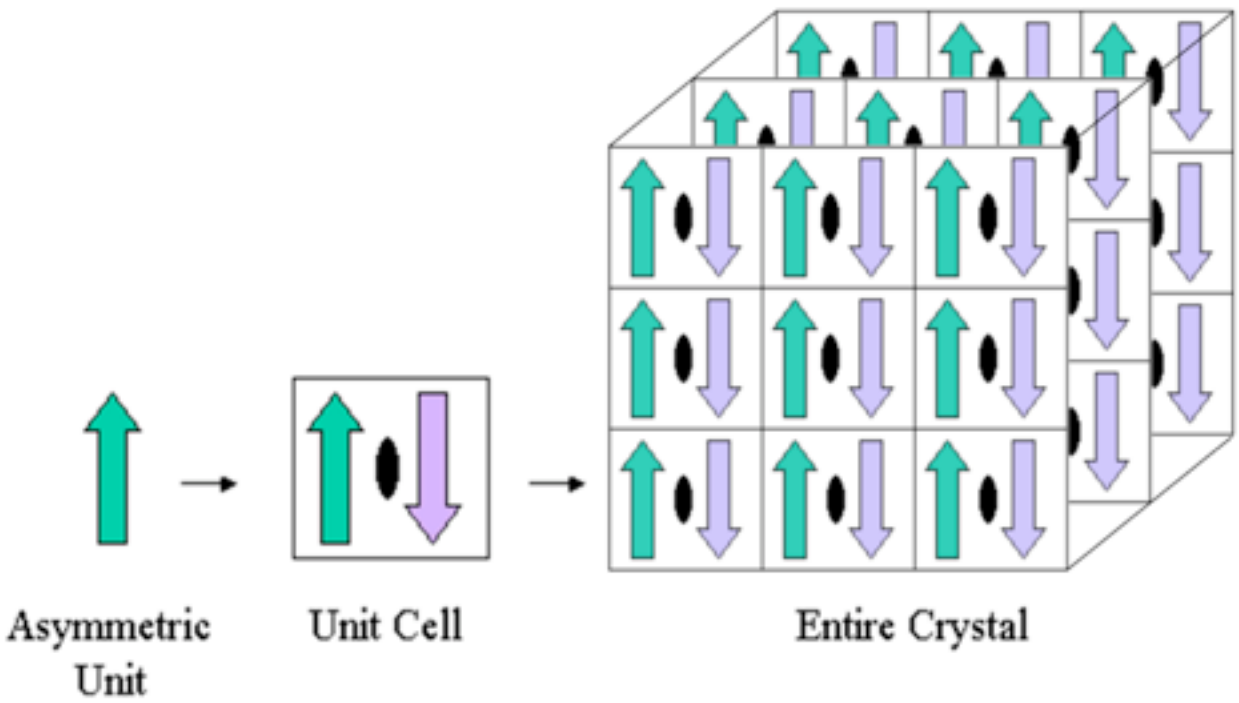


PDB format

http://www.wwpdb.org/documentation/format33/v3.3.html

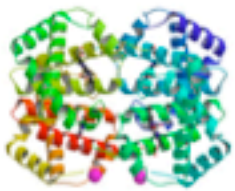
```
HEADER    MOTOR PROTEIN/STRUCTURAL PROTEIN        25-JUN-12   3J1U
TITLE     LOW AFFINITY DYNEIN MICROTUBULE BINDING DOMAIN - TUBULIN COMPLEX
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: CYTOPLASMIC DYNEIN 1 HEAVY CHAIN 1, SERYL T-RNA SYNTHETASE
COMPND   3 CHIMERA;
COMPND   4 CHAIN: A;
COMPND   5 FRAGMENT: SEE REMARK 999;
COMPND   6 SYNONYM: CYTOPLASMIC DYNEIN HEAVY CHAIN 1, DYNEIN HEAVY CHAIN,
COMPND   7 CYTOSOLIC;
COMPND   8 ENGINEERED: YES;
COMPND   9 MOL_ID: 2;
COMPND  10 MOLECULE: TUBULIN ALPHA-1B CHAIN;
COMPND  11 CHAIN: B;
COMPND  12 SYNONYM: ALPHA-TUBULIN UBIQUITOUS, TUBULIN K-ALPHA-1, TUBULIN ALPHA-
COMPND  13 UBIQUITOUS CHAIN;
COMPND  14 MOL_ID: 3;
COMPND  15 MOLECULE: TUBULIN BETA-2B CHAIN;
COMPND  16 CHAIN: C;
COMPND  17 SYNONYM: BETA TUBULIN
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: MUS MUSCULUS;
SOURCE   3 ORGANISM_COMMON: MOUSE;
SOURCE   4 ORGANISM_TAXID: 10090;
SOURCE   5 GENE: DYNC1H1, DHC1, DNCH1, DNCHC1, DYHC;
SOURCE   6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 562;
SOURCE   8 MOL_ID: 2;
SOURCE   9 ORGANISM_SCIENTIFIC: BOS TAURUS;
SOURCE  10 ORGANISM_COMMON: BOVINE;
SOURCE  11 ORGANISM_TAXID: 9913;
SOURCE  12 MOL_ID: 3;
SOURCE  13 ORGANISM_SCIENTIFIC: BOS TAURUS;
SOURCE  14 ORGANISM_COMMON: BOVINE;
SOURCE  15 ORGANISM_TAXID: 9913
KEYWDS    MOTOR PROTEIN-STRUCTURAL PROTEIN COMPLEX
EXPDTA    ELECTRON MICROSCOPY
AUTHOR    W.B.REDWINE,R.HERNANDEZ-LOPEZ,S.ZOU,J.HUANG,S.L.RECK-PETERSON,
AUTHOR   2 A.E.LESCHZINER
```

Assymetric Unit VS Biological



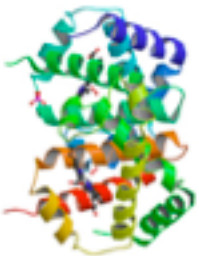
Assymetric Unit VS Biological

Asymmetric unit with one biological assembly



Entry **2hhb** contains **one** hemoglobin molecule (**4 chains**) in the asymmetric unit.

Asymmetric unit with a portion of a biological assembly



Entry **1hho** contains **half** a hemoglobin molecule (**2 chains**) in the asymmetric unit. A crystallographic two-fold axis generates the other 2 chains of the hemoglobin molecule.

2. Structure comparison

Structure-Structure alignments

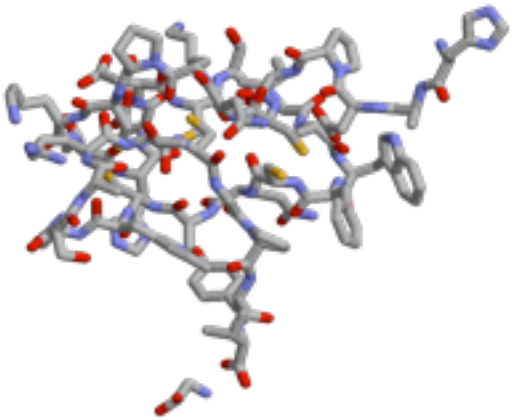
General steps in a bioinformatics procedure:

Representation

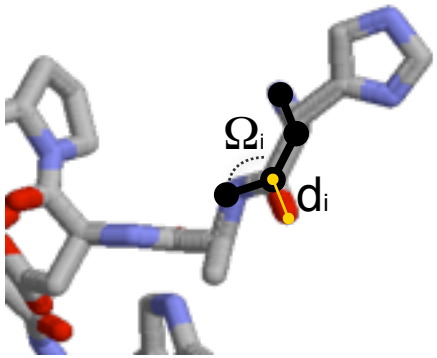
Scoring

Optimizer

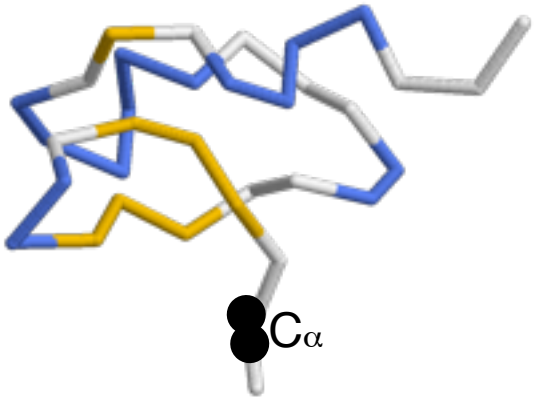
Structures



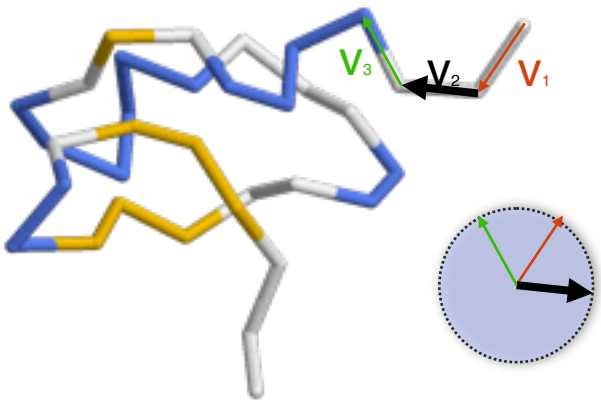
All atoms and coordinates



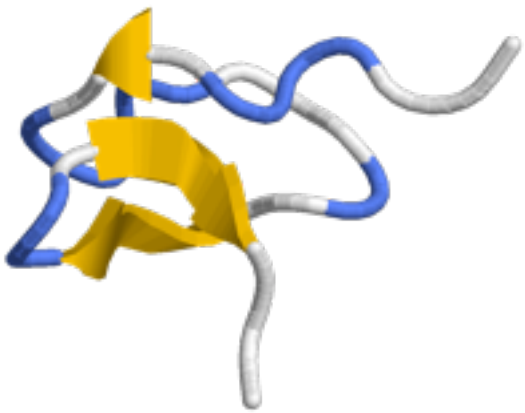
Dihedral space or distance space



Reduced atom representation



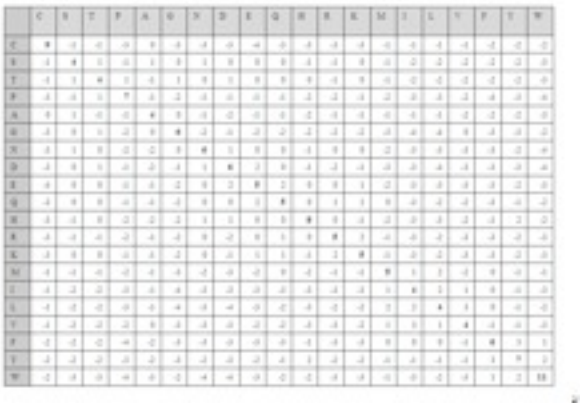
Vector representation



Secondary Structure

Accessible surface (and others)

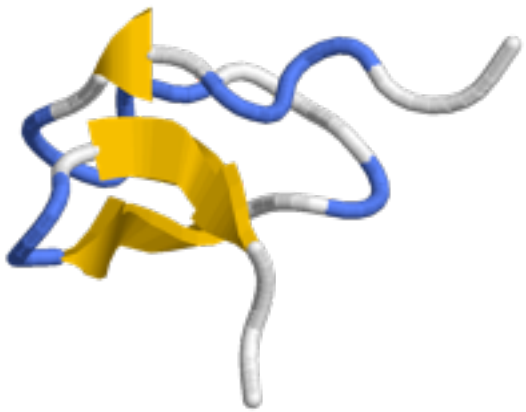
Raw scores



Aminoacid substitutions

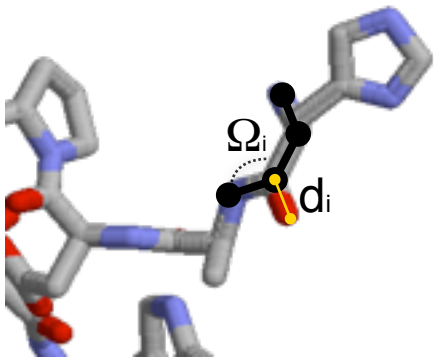
$$RMSD(x,y) = \sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i) - \mathbf{y}(i)\right\|^2\right)}$$

Root Mean Square Deviation



Secondary Structure (H,B,C)

Accessible surface (B,A [%])

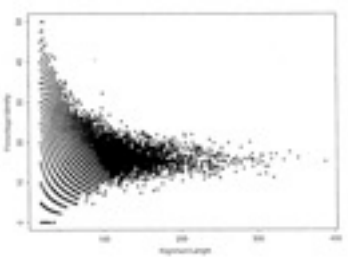


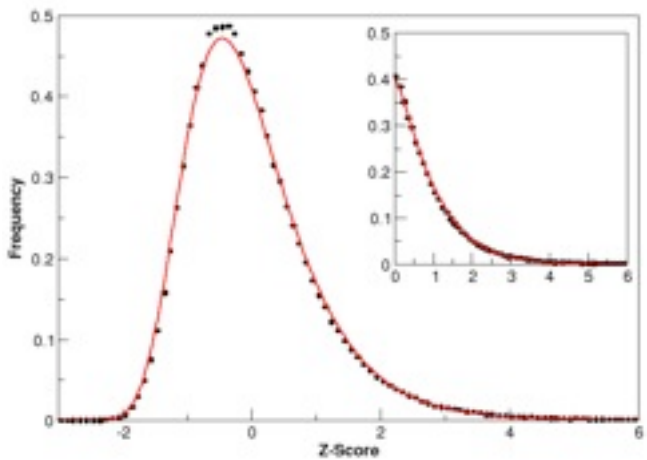
Angles or distances

Significance of an alignment (score)

Probability that the optimal alignment of two random sequences/structures of the same length and composition as the aligned sequences/structures have at least as good a score as the evaluated alignment.

Empirical





Sometimes approximated by Z-score (normal distribution).

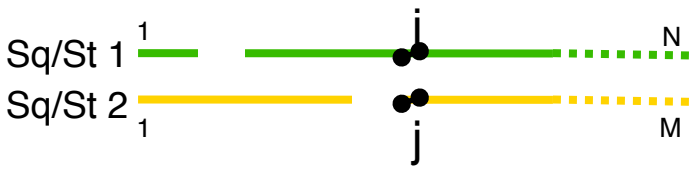
Analytic

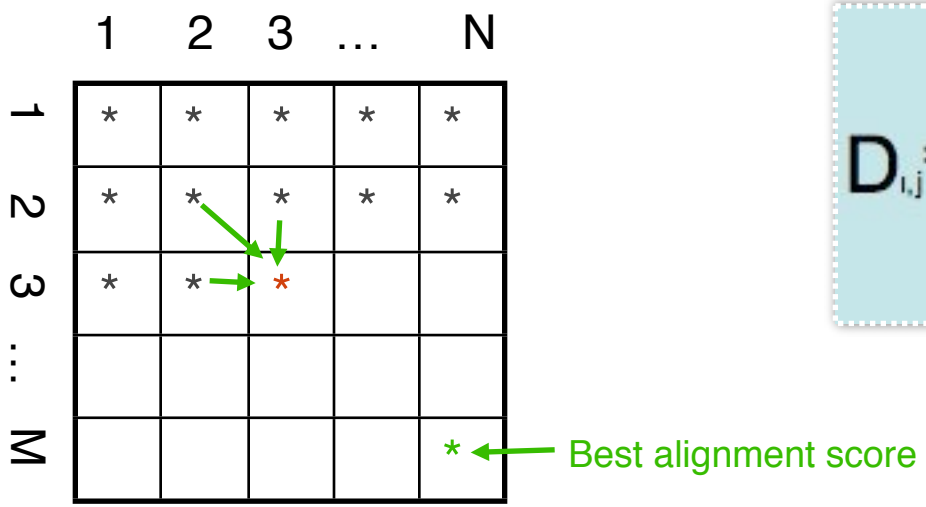
$$P(s) = e^{-\lambda (s-\mu)}$$

$$P(s \geq x) = 1 - \exp\left(e^{-\lambda (s-\mu)} \right)$$

Karlin and Altschul, 1990 PNAS 87, pp2264

Global dynamic programming alignment



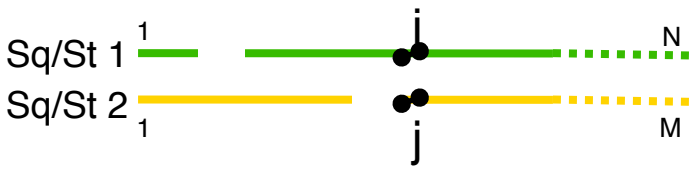


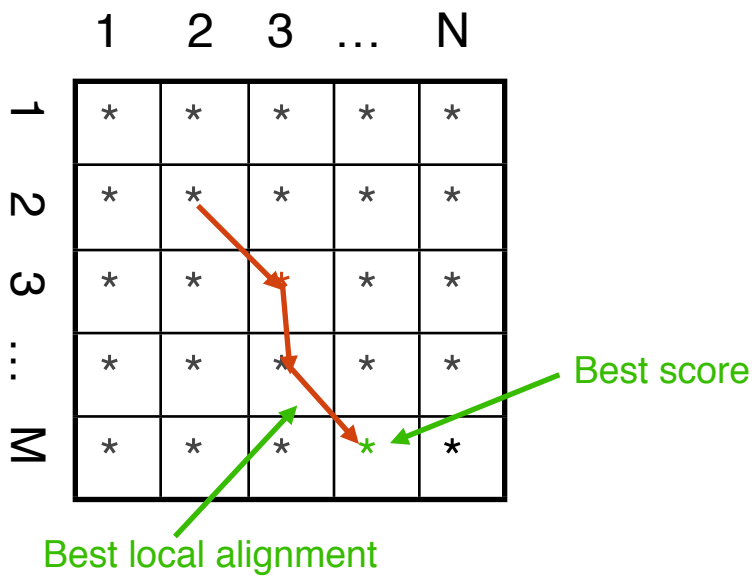
$$D_{i,j}=\min\begin{cases} D_{i,j-1}+Score_{(\Delta,rj)} \\ D_{i-1,j-1}+Score_{(ri,rj)} \\ D_{i-1,j}+Score_{(ri,\Delta)} \end{cases}$$

Backtracking to get the best alignment

Needleman and Wunsch (1970) J. Mol Biol, 3 pp443

Local dynamic programming alignment



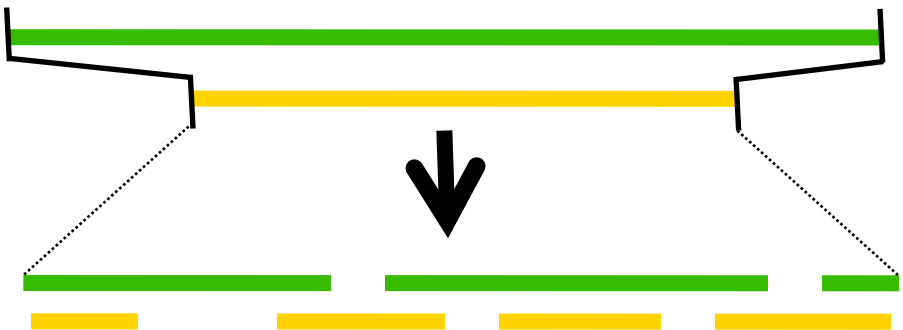


$$D_{i,j} = \min \begin{cases} D_{i,j-1} + Score_{(\Delta, rj)} \\ D_{i-1,j-1} + Score_{(ri, rj)} \\ D_{i-1,j} + Score_{(ri, \Delta)} \\ 0 \end{cases}$$

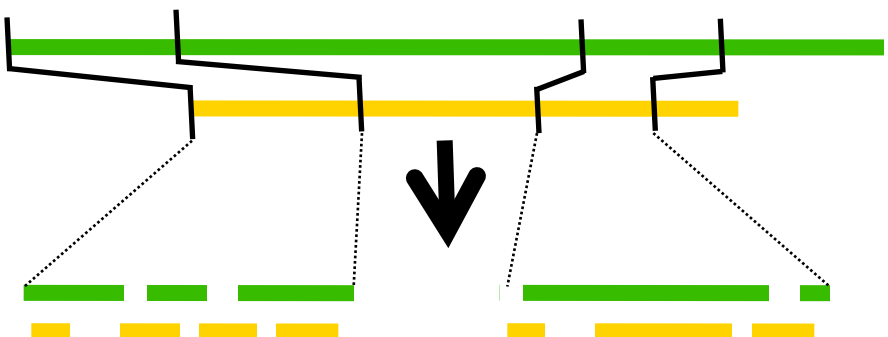
Backtracking to get the best alignment

Smith and Waterman (1981) J. Mol Biol, 147 pp195

Global .vs. local alignment



Global alignment



Local alignment

Multiple alignment

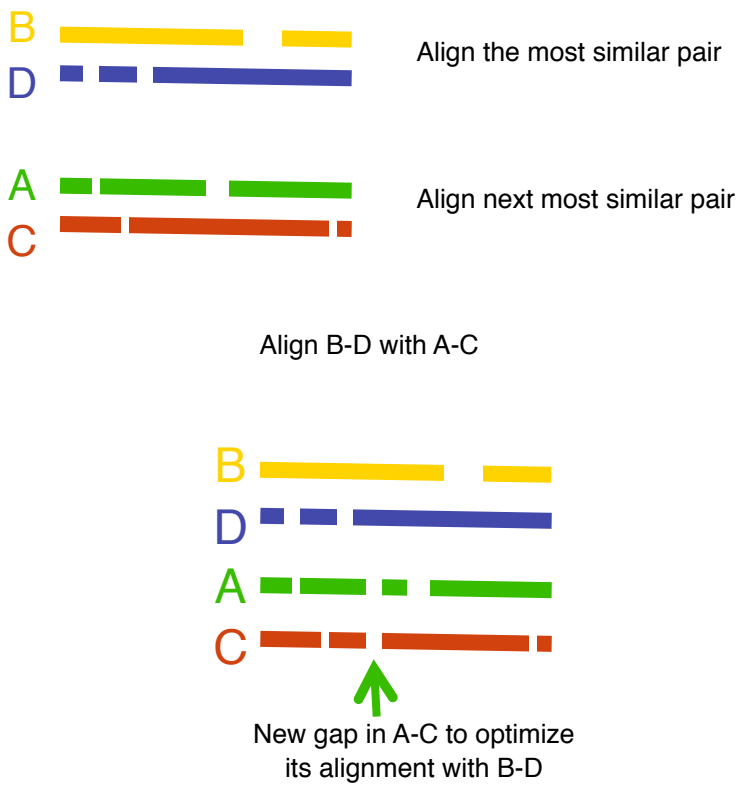
Pairwise alignments

Example – 4 sequences A, B, C, D.

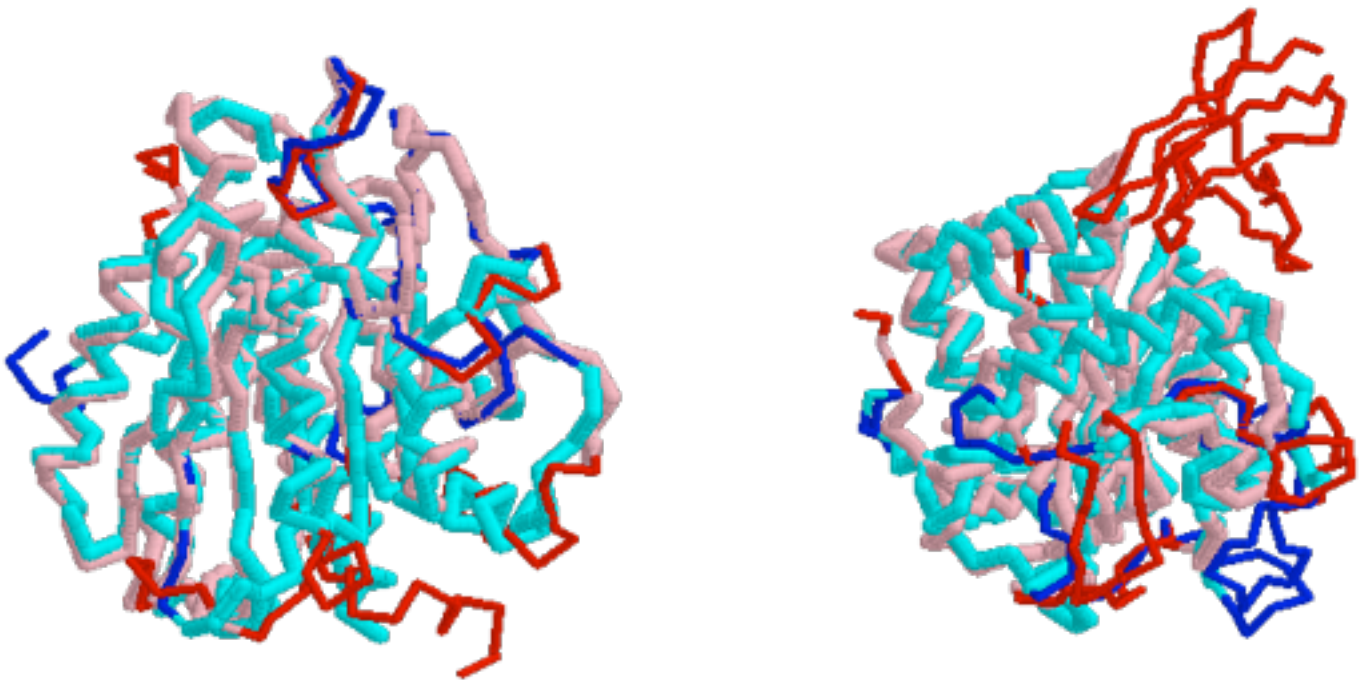
6 pairwise comparisons
then cluster analysis

Multiple alignments

Following the tree from step 1



Coverage .vs. Accuracy

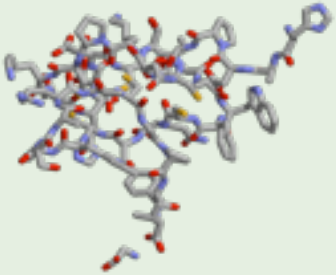


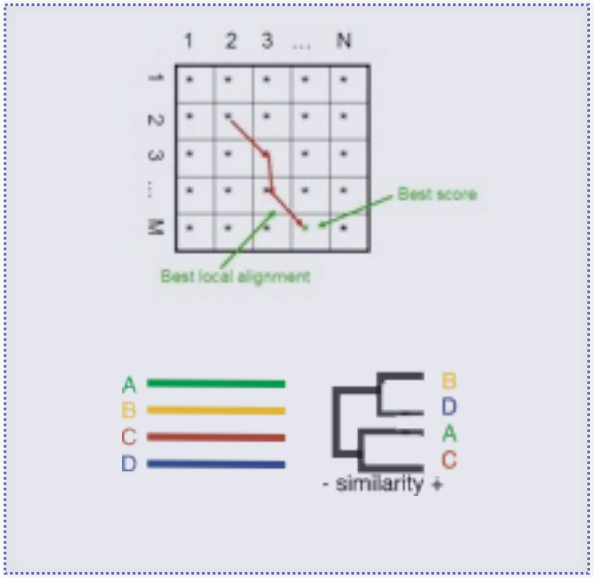
Same RMSD ~ 2.5Å

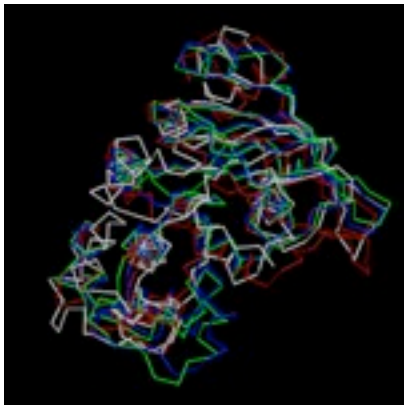
Coverage ~90% Cα

Coverage ~75% Cα

Structural alignment by properties conservation (SALIGN-MODELLER)

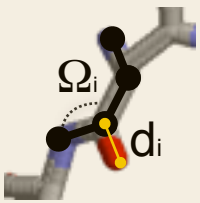


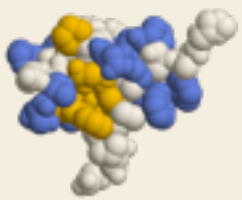




✓ Uses all available structural information
✓ Provides the optimal alignment

Computationally expensive





$$RMSD(x,y) = \sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i) - \mathbf{y}(i)\right\|^2\right)}$$

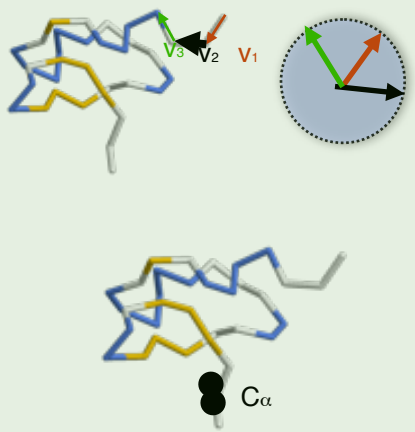
$R_{i,j}$ $D_{,i(3),j(3)}$ $S_{i,j}$ $B_{i,j}$ $I_{i,j}$

M. S. Madhusudhan, B. M. Webb, M. A. Marti-Renom, N. Eswar, A. Sali, Protein Eng Des Sel, (Jul 8, 2009).

Structural alignment by properties conservation (SALIGN-MODELLER)

http://salilab.org/salign

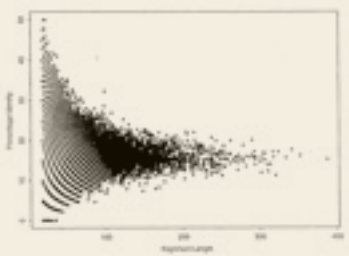
Vector Alignment Search Tool (VAST)



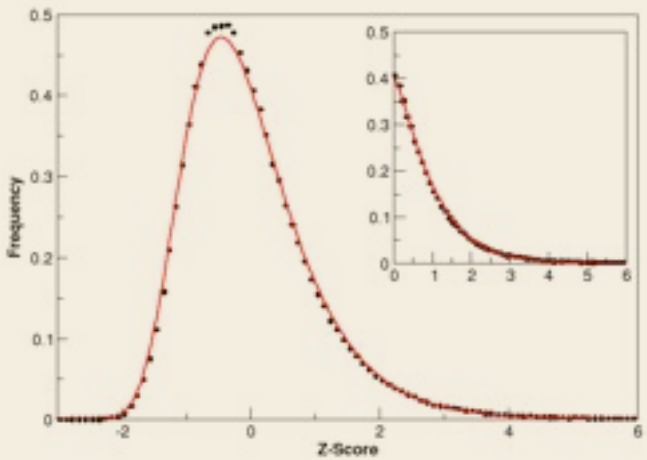
Graph theory search
of similar SSE
Refining by Monte Carlo
at all atom resolution

✓ Good scoring system with significance

Reduces the protein representation



$$RMSD(x,y) = \sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\|\mathbf{x}(i)-\mathbf{y}(i)\|^2\right)}$$

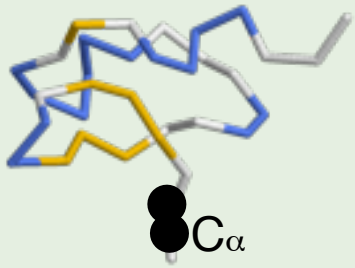


Gibrat JF et al. (1996) Curr Opin Struct Biol 3 pp377

Vector Alignment Search Tool (VAST)

http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml

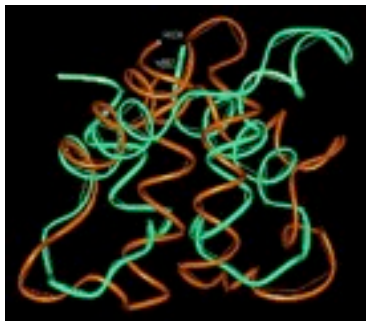
Incremental combinatorial extension (CE)



Exhaustive combination of fragments

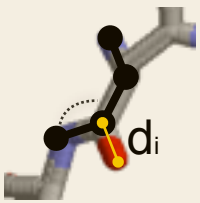
Longest combination of AFPs

Heuristic similar to PSI-BLAST



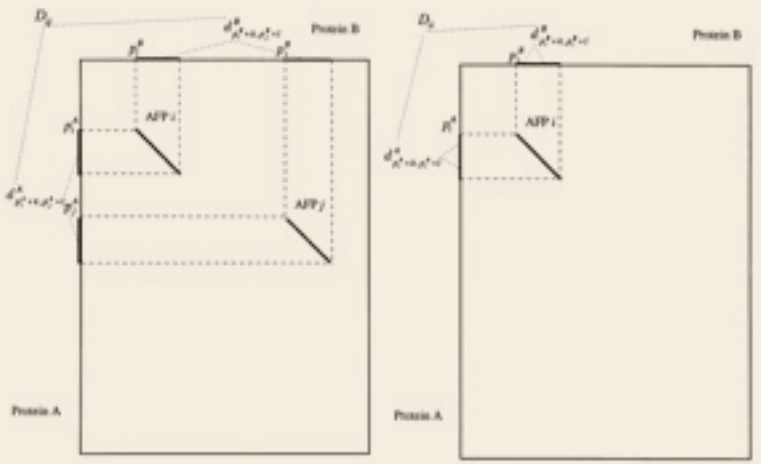
✓ FAST!
✓ Good quality of local alignments

Complicated scoring and heuristics



8 residues peptides

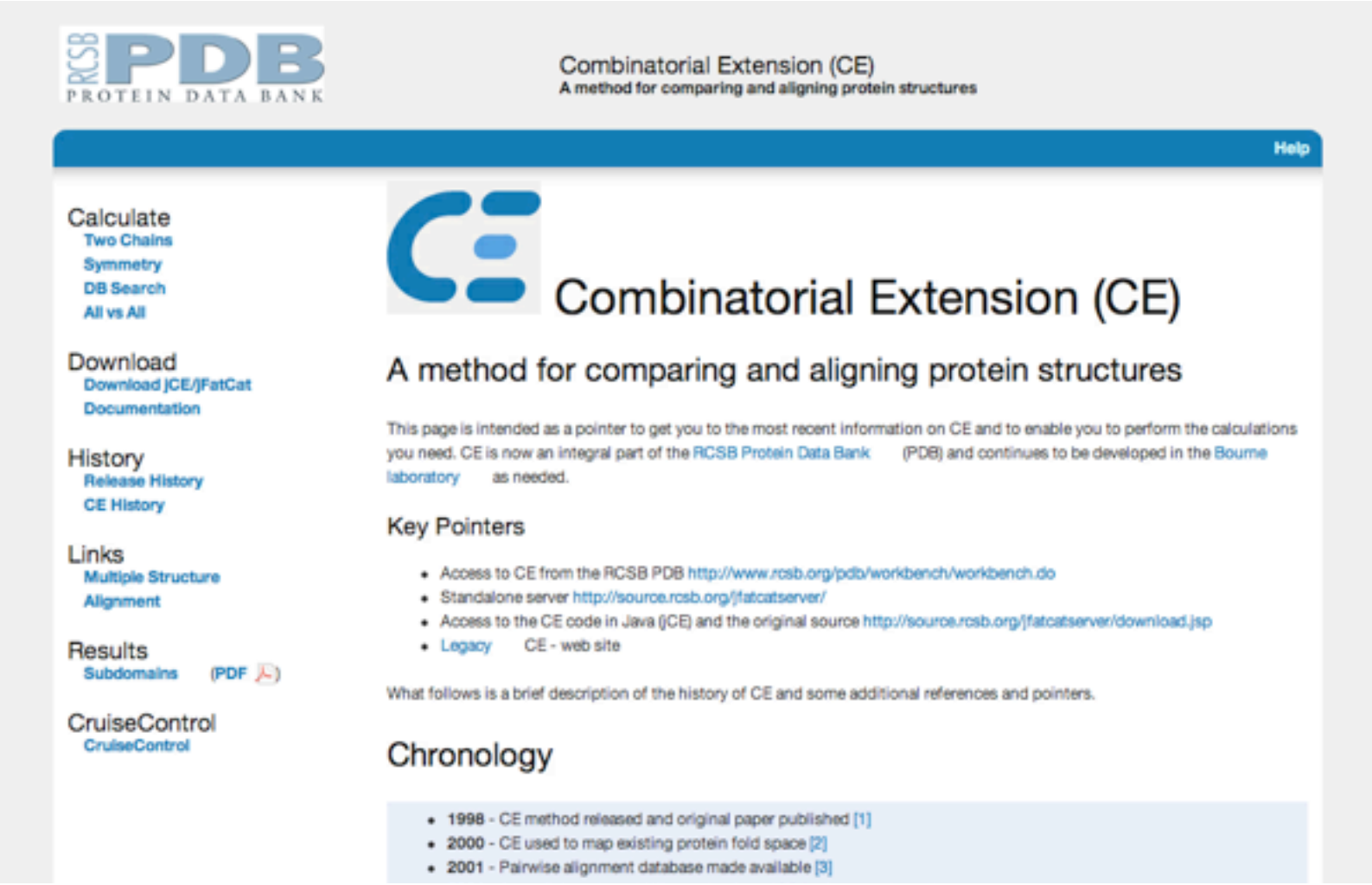
$$RMSD(x,y) = \sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i)-\mathbf{y}(i)\right\|^{2}\right)}$$



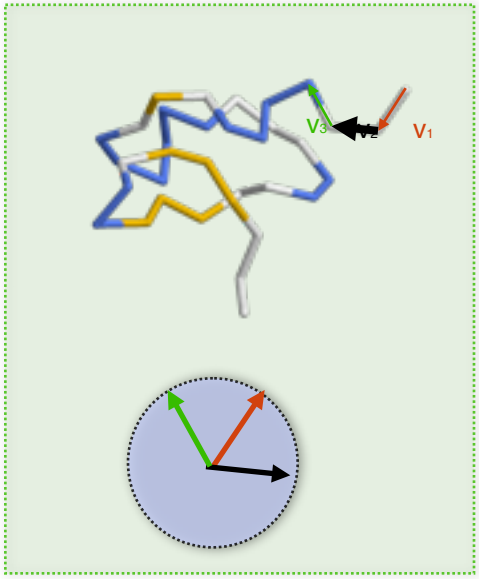
Shindyalov IN, amd Bourne PE. (1998) Protein Eng. 9 pp739

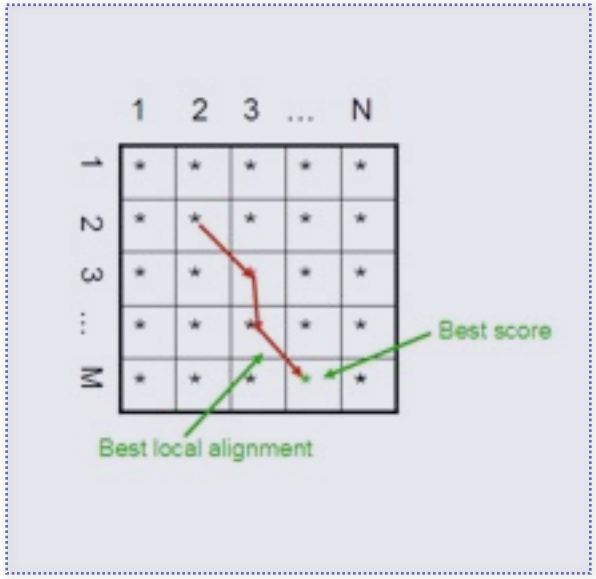
Incremental combinatorial extension (CE)

http://source.rcsb.org/jfatcatserver/ceHome.jsp



Matching molecular models obtained from theory (MAMMOTH)



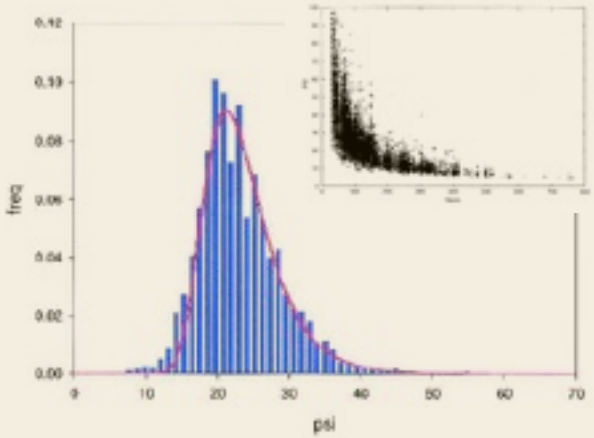


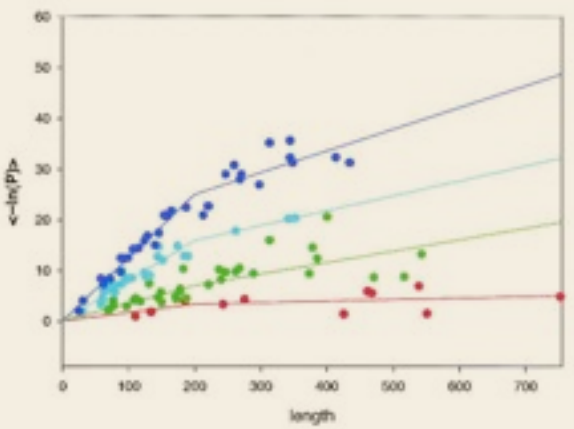
- ✓ VERY FAST!
- ✓ Good scoring system with significance

Reduces the protein representation

$$URMS^{R} = \sqrt{2.0 - \frac{2.84}{\sqrt{n}}}$$

$$S_{AB} = \frac{(URMS^{R} - URMS^{AB})D}{URMS^{R}}$$





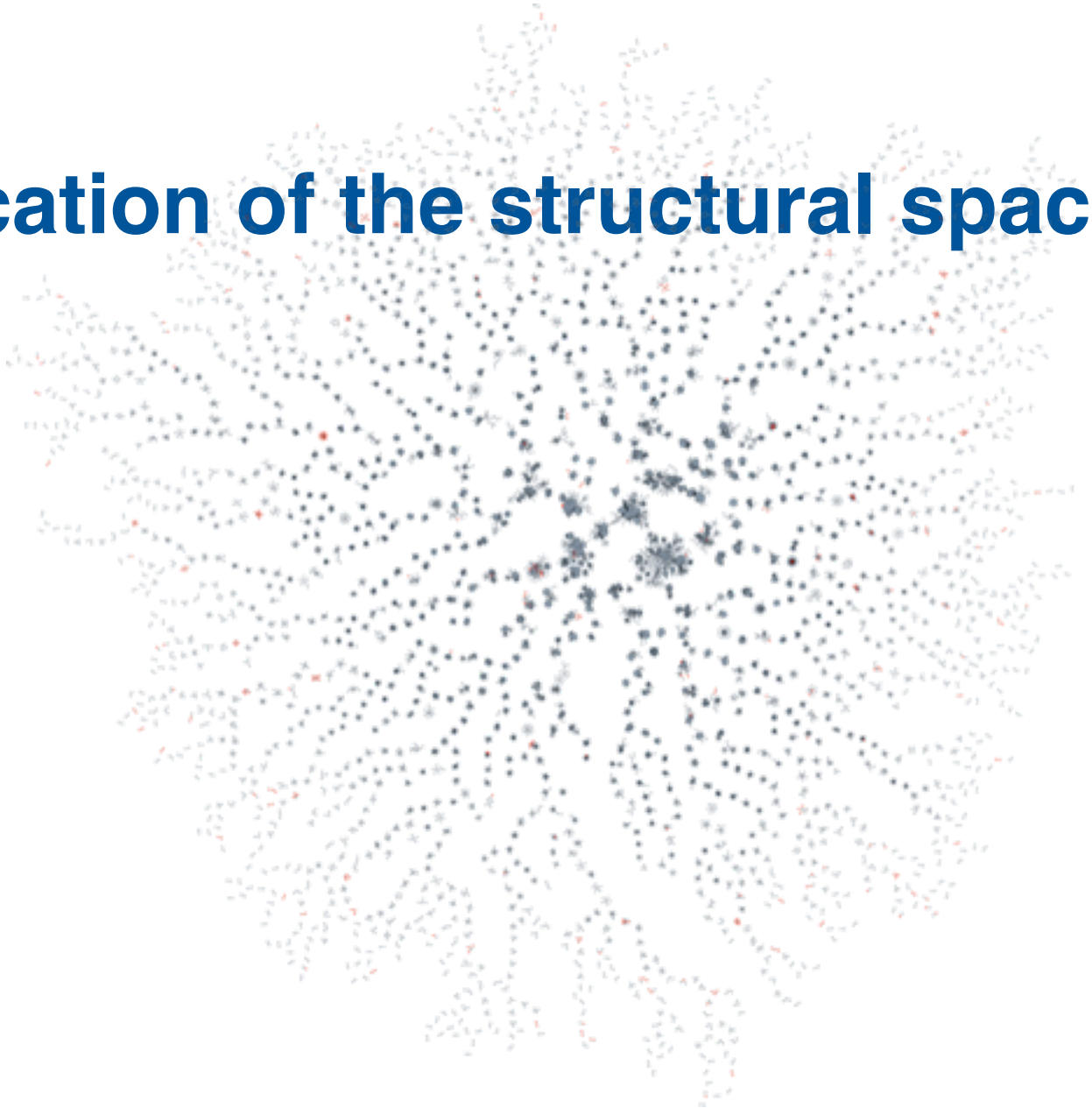
Ortiz AR, (2002) Protein Sci. 11 pp2606

Matching molecular models obtained from theory (MAMMOTH)

http://ub.cbm.uam.es/software/online/mammoth.php

3. Structure classification

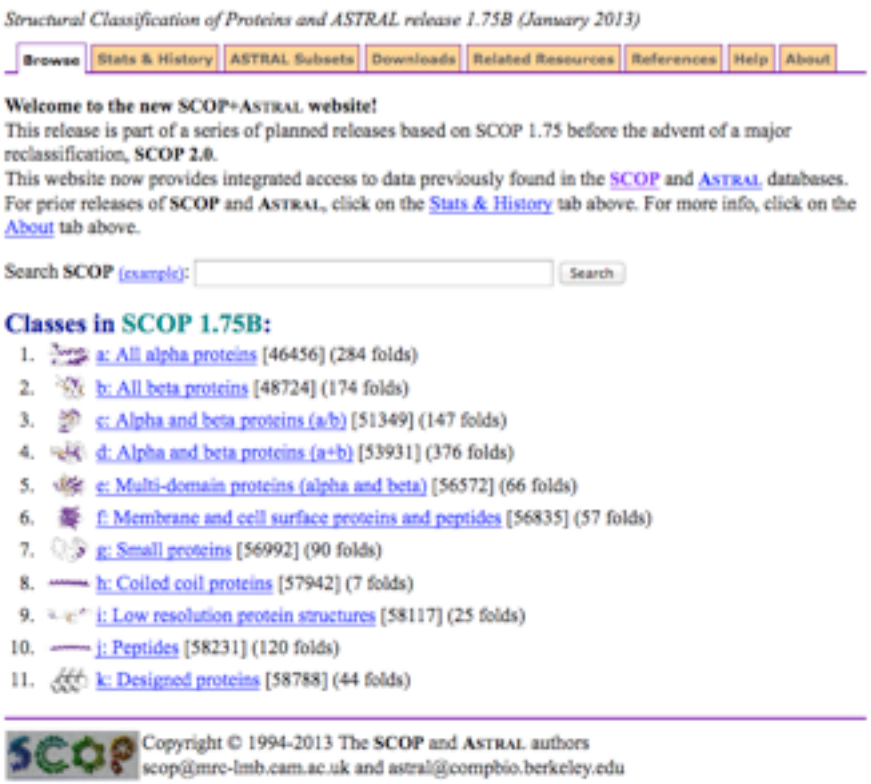
Classification of the structural space



SCOP$_{1.75}$ database

http://scop.berkeley.edu/

Murzin A. G.,el at. (1995). J. Mol. Biol. ***247****, 536-540.*



- ✓ **Largely recognized as "standard of gold"**
- ✓ **Manually classification**
- ✓ **Clear classification of structures in:**

CLASS
FOLD
SUPER-FAMILY
FAMILY

- ✓ **Some large number of tools already available**

Manually classification
Not 100% up-to-date
Domain boundaries definition

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	284	507	928
All beta proteins	174	354	815
Alpha and beta proteins (a/b)	147	244	902
Alpha and beta proteins (a+b)	376	552	1170
Multi-domain proteins	66	66	100
Membrane and cell surface proteins	57	109	127
Small proteins	90	129	230
Total	1194	1961	4272

a: All alpha proteins -> a.3: Cytochrome c -> a.3.1: Cytochrome c ->
(class) (fold) (superfamily)
a.3.1.4: Two-domain cytochrome c
(family)

Structural Classification of Proteins and ASTRAL release 1.75B (January 2013)

Browse | Stats & History | ASTRAL Subsets | Downloads | Related Resources | References | Help | About

Search **SCOP** (example): Search

Lineage for Family a.3.1.4: Two-domain cytochrome c

1. Root: SCOP 1.75B
2. Class a: All alpha proteins [46456] (284 folds)
3. Fold a.3: Cytochrome c [46625] (1 superfamily)
   core: 3 helices; folded leaf, opened
4. Superfamily a.3.1: Cytochrome c [46626] (9 families) $\mathcal{S}$
   covalently-bound heme completes the core
5. Family a.3.1.4: Two-domain cytochrome c [46680] (2 protein domains)
   duplication: consists of two cytochrome c type domains

Protein Domains:

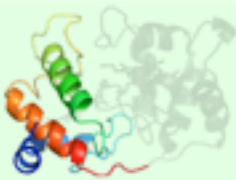
1. Cytochrome c4 [46681] (2 species)
   1. Species Pseudomonas stutzeri [TaxId:316] [46682] (3 PDB entries)
   2. Species Thiobacillus ferrooxidans [TaxId:920] [88972] (1 PDB entry)
2. Flavocytochrome c sulfide dehydrogenase, FCSD, cytochrome subunit [46683] (1 species)
   Species Chromatium vinosum [TaxId:1049] [46684] (1 PDB entry)

More info for Family a.3.1.4: Two-domain cytochrome c

Timeline for Family a.3.1.4: Two-domain cytochrome c:

Family a.3.1.4: Two-domain cytochrome c appears in SCOP 1.75A

Lineage for d1etpa1 (1etp A:1-92)

1. Root: SCOP 1.75B
2. Class a: All alpha proteins [46456] (284 folds)
3. Fold a.3: Cytochrome c [46625] (1 superfamily)
   core: 3 helices; folded leaf, opened
4. Superfamily a.3.1: Cytochrome c [46626] (9 families) S
   covalently-bound heme completes the core
5. Family a.3.1.4: Two-domain cytochrome c [46680] (2 protein domains)
   duplication: consists of two cytochrome c type domains
6. Protein Cytochrome c4 [46681] (2 species)
7. Species Pseudomonas stutzeri [TaxId:316] [46682] (3 PDB entries)
8. 
   Domain d1etpa1: 1etp A:1-92 [15962]
   complexed with hem

Details for d1etpa1

PDB Entry: 1etp (more details)
PDB Description: crystal structure of cytochrome c4 from pseudomonas stutzeri
PDB Compounds: (A:) cytochrome c4

SCOP Domain Sequences for d1etpa1:

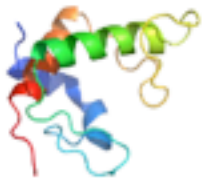
Sequence; same for both **SEQRES** and **ATOM** records: (download)

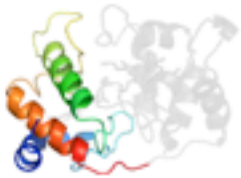
```
>d1etpa1 a.3.1.4 (A:1-92) Cytochrome c4 {Pseudomonas stutzeri [TaxId: 316]}
agdaeagqgkvavcgachgvdgnspapnfpklagqgeryllkqlqdikagstpgapegvg
rkvlemtgmldplsdqdlediaayfsqkgsv
```

SCOP Domain Coordinates for d1etpa1:

Click to download the PDB-style file with coordinates for d1etpa1.
(The format of our PDB-style files is described here.)

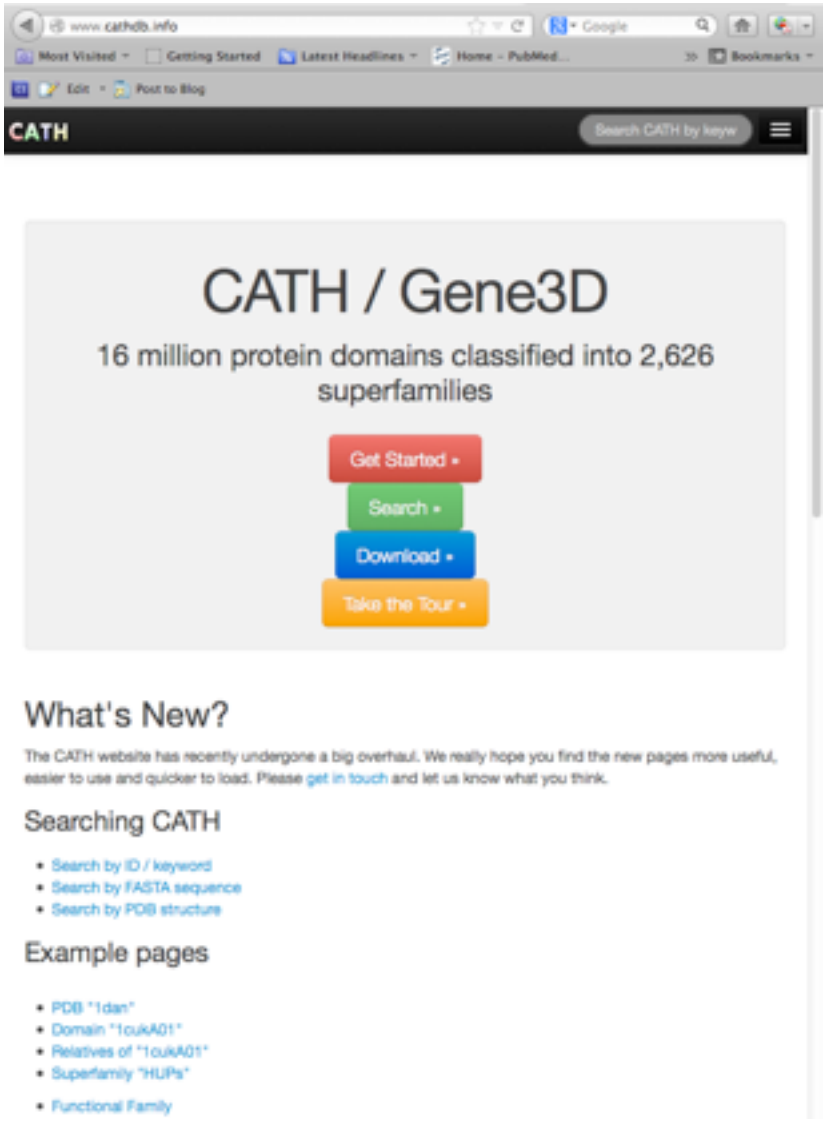
Timeline for d1etpa1:


d1etpa1
(click for larger image)


d1etpa1 in context of chain
Domains from same chain:
d1etpa2

CATH$_{3.5}$ database

http://www.cathdb.info



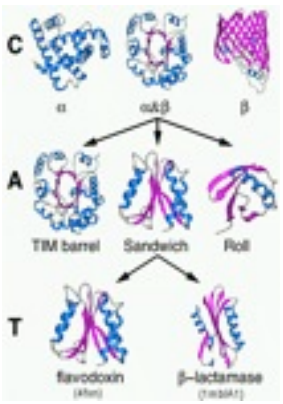
Uses FSSP for superimposition

- ✓ Recognized as “standard of gold”
- ✓ Semi-automatic classification
- ✓ Clear classification of structures in:

CLASS
ARCHITECTURE
TOPOLOGY
HOMOLOGOUS SUPERFAMILIES

- ✓ Some large number of tools already available
- ✓ Easy to navigate

Semi-automatic classification
Domain boundaries definition



173,536 CATH Domains
2,626 CATH Superfamilies
51,334 PDBs

*Orengo, C.A., et al. (1997) Structure. **5**. 1093-1108.*

Browse - tree

Details on the currently selected CATH node are displayed in the panel below

H Nop10-like SnoRNP

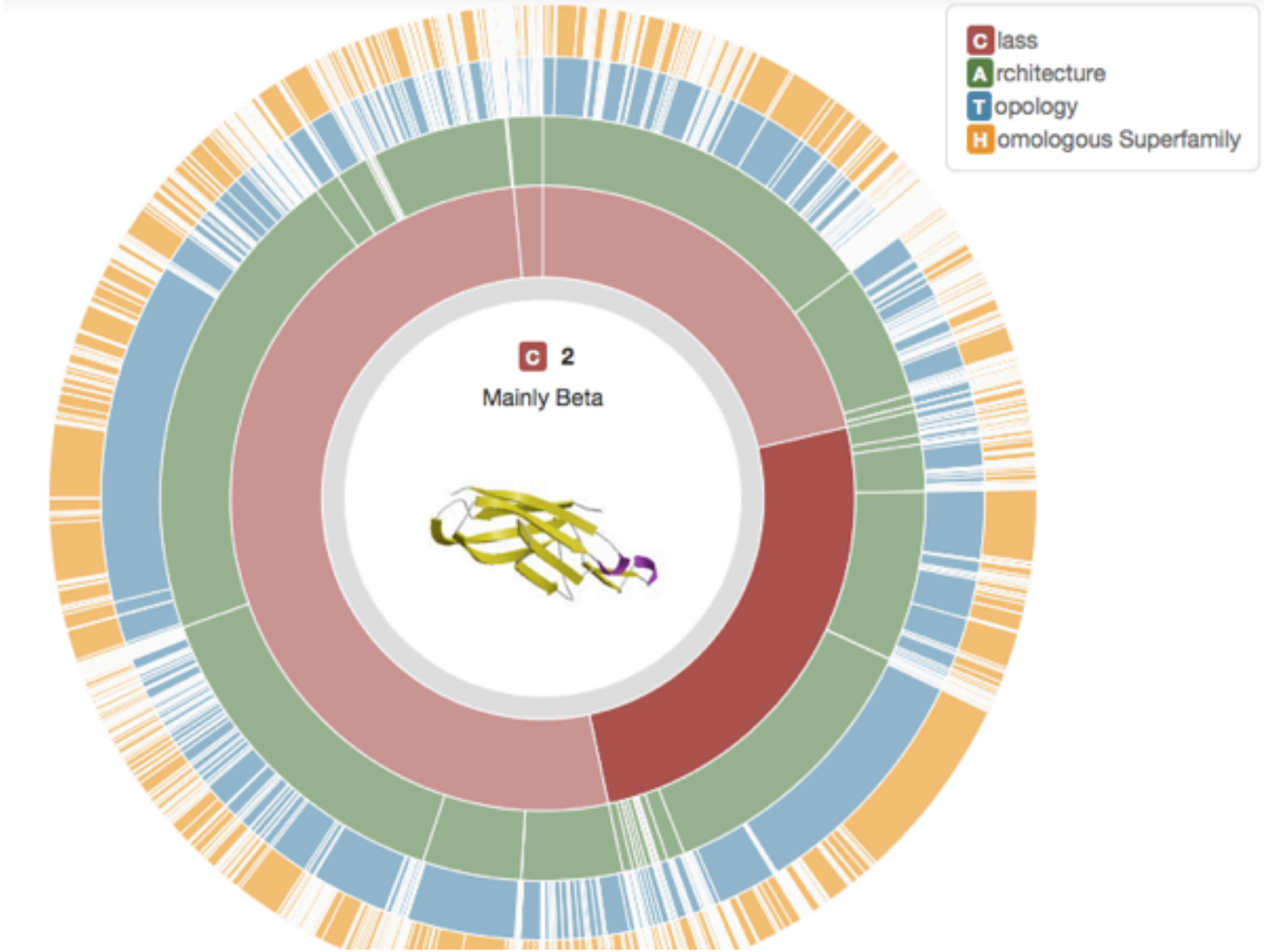
View Superfamily

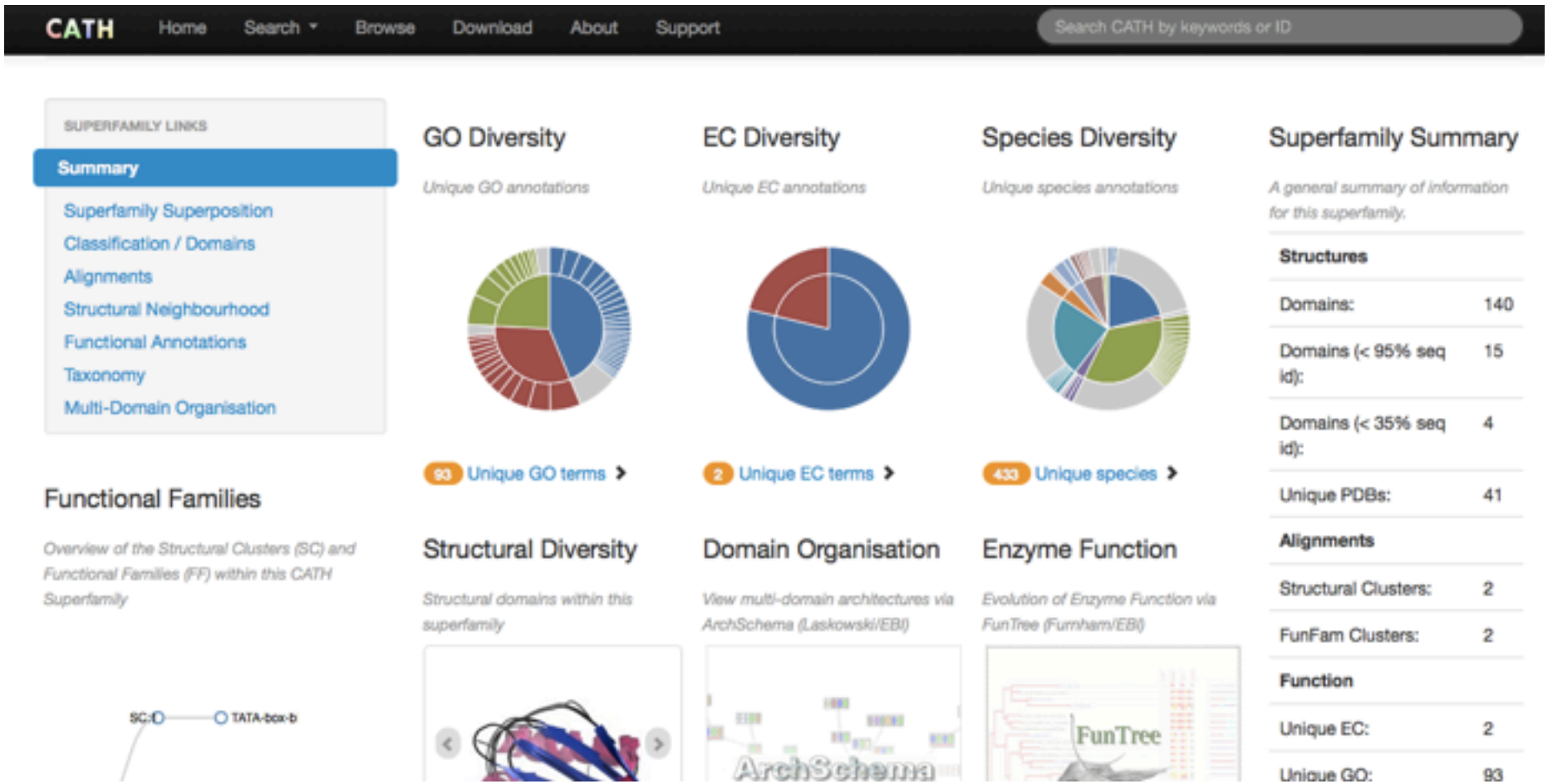
CATH ID	2.20.28.40
Non-redundant Sequences (<35% seq id)	1
Domains	3
Example Domain	2apoB00 [PDB]

Top of CATH Hierarchy (4 Classes)

- C 1 Mainly Alpha
- C 2 Mainly Beta — *5 Architectures, 386 Folds, 875 Superfamilies, 37038 Domains*
  - A 2.10 Ribbon — *20 Architectures, 229 Folds, 520 Superfamilies, 43881 Domains*
  - A 2.20 Single Sheet — *25 Folds, 31 Superfamilies, 1733 Domains*
    - T 2.20.20 Anthopleurin-A — *20 Folds, 31 Superfamilies, 683 Domains*
    - T 2.20.25 N-terminal domain of TfIIb — *3 Superfamilies, 21 Domains*
    - T 2.20.28 Rubrerythrin, domain 2 — *6 Superfamilies, 342 Domains*
      - H 2.20.28.10 *Not yet named* — *5 Superfamilies, 209 Domains*
      - H 2.20.28.20 *Not yet named* — *131 Domains*
      - H 2.20.28.30 RNA polymerase ii, chain L — *11 Domains*
      - H 2.20.28.40 Nop10-like SnoRNP — *61 Domains*
      - H 2.20.28.50 degv family protein — *3 Domains*
    - T 2.20.50 Outer Surface Protein A; domain 2 — *3 Domains*
    - T 2.20.60 Heparin-binding Growth Factor, Midkine; Chain A — *1 Superfamilies, 1 Domains*
    - T 2.20.70 Ubiquitin Ligase Nedd4; Chain: W; — *1 Superfamilies, 1 Domains*
    - T 2.20.80 Lipovitellin-phosvitin complex, chain A, domain 4 — *1 Superfamilies, 60 Domains*
    - T 2.20.90 Lipovitellin-phosvitin complex; beta-sheet shell regions — *1 Superfamilies, 1 Domains*
    - T 2.20.100 TSP-1 type 1 repeat — *1 Superfamilies, 1 Domains*
    - T 2.20.110 Histone H3 K4-specific methyltransferase SET7/9 N-terminal domain — *1 Superfamilies, 3 Domains*
    - T 2.20.120 Multimodular pneumococcal cell wall endolysin, domain 3 — *1 Superfamilies, 3 Domains*
    - T 2.20.130 S-adenosyl-L-methionine-dependent methyltransferases — *1 Superfamilies, 6 Domains*

Browse - sunburst




CATH
Home
Search
Browse
Download
About
Support
Search CATH by keywords or ID
SUPERFAMILY LINKS
Summary
Superfamily Superposition
Classification / Domains
Alignments
Structural Neighbourhood
Functional Annotations
Taxonomy
Multi-Domain Organisation
GO Diversity
Unique GO annotations
93 Unique GO terms
EC Diversity
Unique EC annotations
2 Unique EC terms
Species Diversity
Unique species annotations
433 Unique species
Superfamily Summary
A general summary of information for this superfamily.
Structures
Domains: 140
Domains (< 95% seq id): 15
Domains (< 35% seq id): 4
Unique PDBs: 41
Alignments
Structural Clusters: 2
FunFam Clusters: 2
Function
Unique EC: 2
Unique GO: 93
Functional Families
Overview of the Structural Clusters (SC) and Functional Families (FF) within this CATH Superfamily
SC:0
TATA-box-b
Structural Diversity
Structural domains within this superfamily
Domain Organisation
View multi-domain architectures via ArchSchema (Laskowski/EBI)
ArchSchema
Enzyme Function
Evolution of Enzyme Function via FunTree (Furnham/EBI)
FunTree

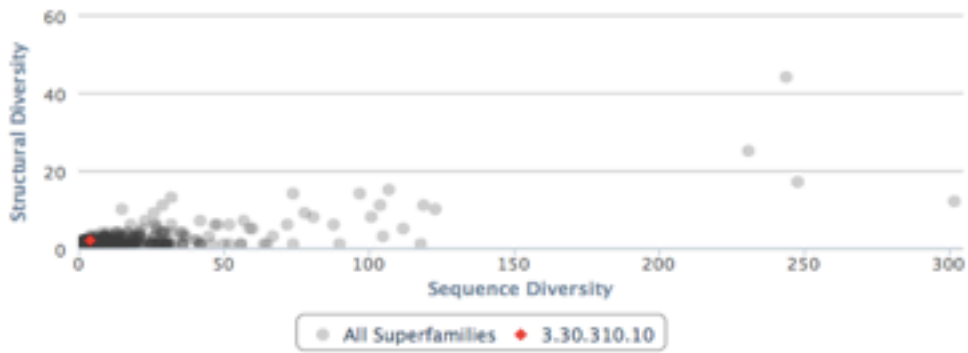
ArchSchema (requires Java)

FunTree (opens new window)

Unique EC:	2
Unique GO:	93
Taxonomy	
Unique Species:	433

Sequence/Structure Diversity

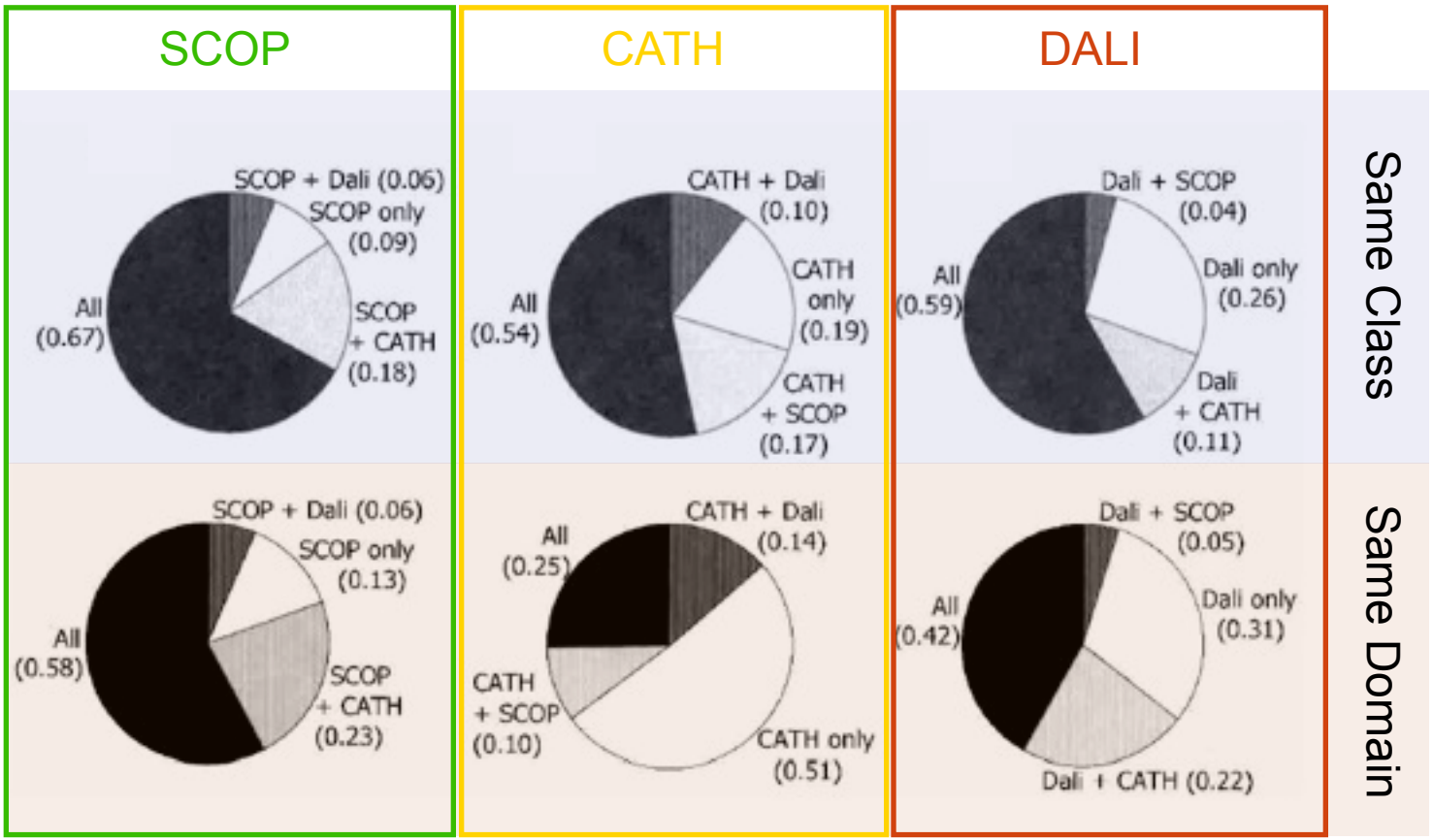
Overview of the sequence / structure diversity of this superfamily compared to other superfamilies in CATH. Click on the chart to view the data in more detail.



Classification of the structural space

Not an easy task!

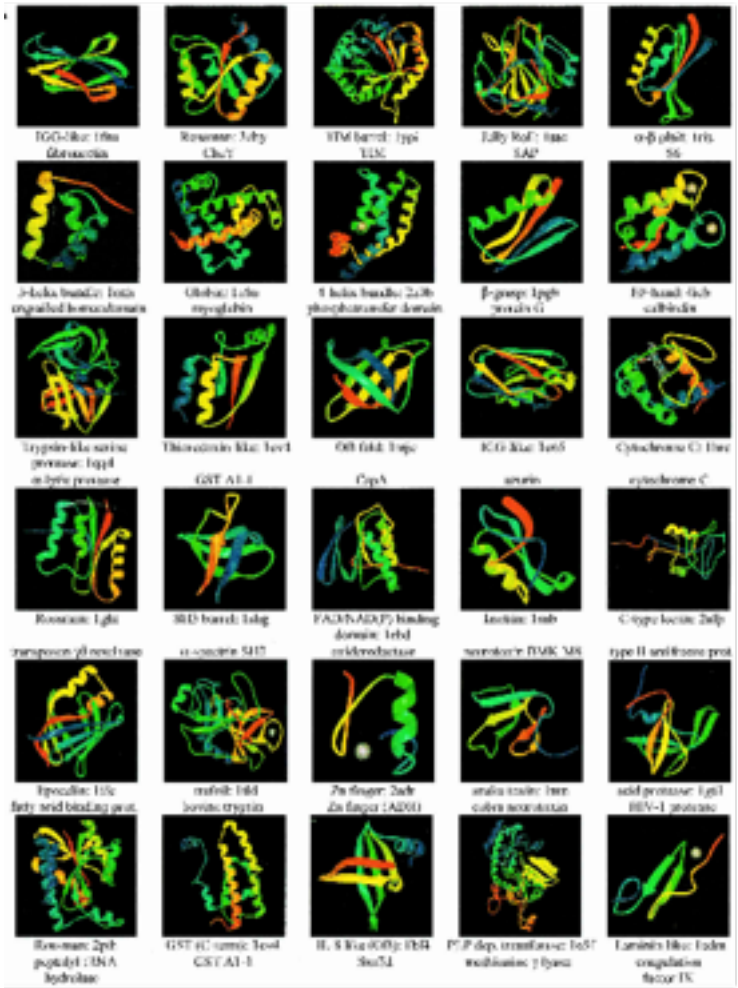
Domain definition AND domain classification

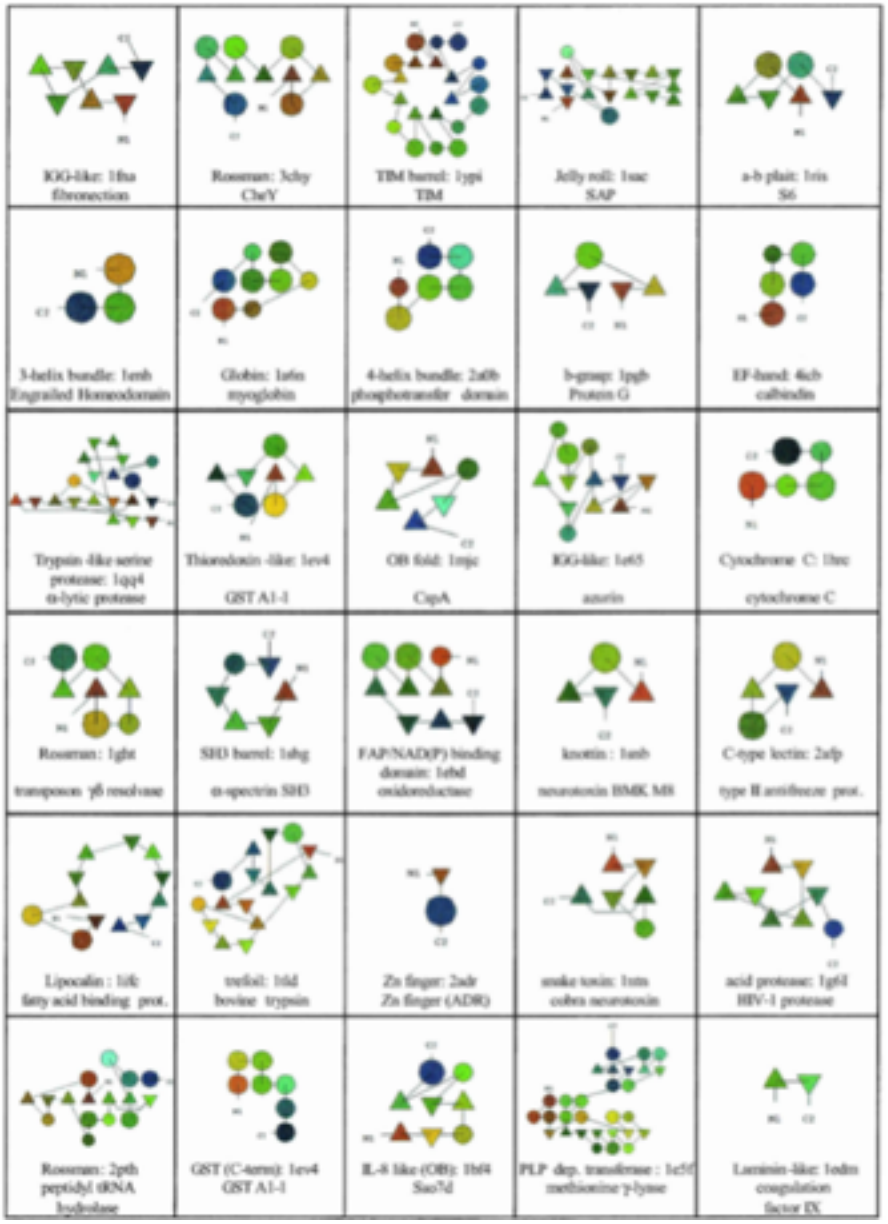


*Day, et al. (2003) Protein Sciences, **12** pp2150*

Table 2. *SCOP, CATH, and Dali codes associated with the 30 most populated metafolds*

4. Number of folds in nature




IGG-like: 1fha
fibronectin
Rossman: 3chy
CheY
TIM barrel: 1ypi
TIM
Jelly roll: 1sac
SAP
a-b plait: 1ris
S6
3-helix bundle: 1enh
Engrailed Homeodomain
Globin: 1a6n
myoglobin
4-helix bundle: 2a0b
phosphotransfer domain
b-grasp: 1pgb
Protein G
EF-hand: 4icb
calbindin
Trypsin-like serine
protease: 1qq4
a-lytic protease
Thioredoxin -like: 1ev4
GST A1-1
OB fold: 1mjc
CspA
IGG-like: 1a65
azurin
Cytochrome C: 1hrc
cytochrome C
Rossman: 1ghr
transposon γδ resolvase
SH3 barrel: 1shg
α-spectrin SH3
FAP/NAD(P) binding
domain: 1ebd
oxidoreductase
knottin : 1anh
neurotoxin BMK M8
C-type lectin: 2afp
type II antifreeze prot.
Lipocalin : 1ifc
fatty acid binding prot.
trefoil: 1tld
bovine trypsin
Zn finger: 2adr
Zn finger (ADR)
snake toxin: 1nts
cobra neurotoxin
acid protease: 1g6l
HIV-1 protease
Rossman: 2pth
peptidyl tRNA
hydrolase
GST (C-term): 1ev4
GST A1-1
IL-8 like (OB): 1h9
Sso7d
PLP dep. transferase : 1c5f
methionine γ-lyase
Laminin-like: 1odn
coagulation
factor IX

5. Sequences VS fold structures

Proteins. 2009 Nov 15;77(3):499-508. doi: 10.1002/prot.22458.

Structure is three to ten times more conserved than sequence--a study of structural response in protein cores.

Illergård K, Ardell DH, Elofsson A.

Center for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden.

Structure is three to ten times more conserved than sequence—A study of structural response in protein cores

Kristoffer Illergård,[1] David H. Ardell,[2,3] and Arne Elofsson[1]*

[1] Center for Biomembrane Research and Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden

[2] Department of Natural Sciences, School of Natural Sciences, University of California, Merced, California 95344

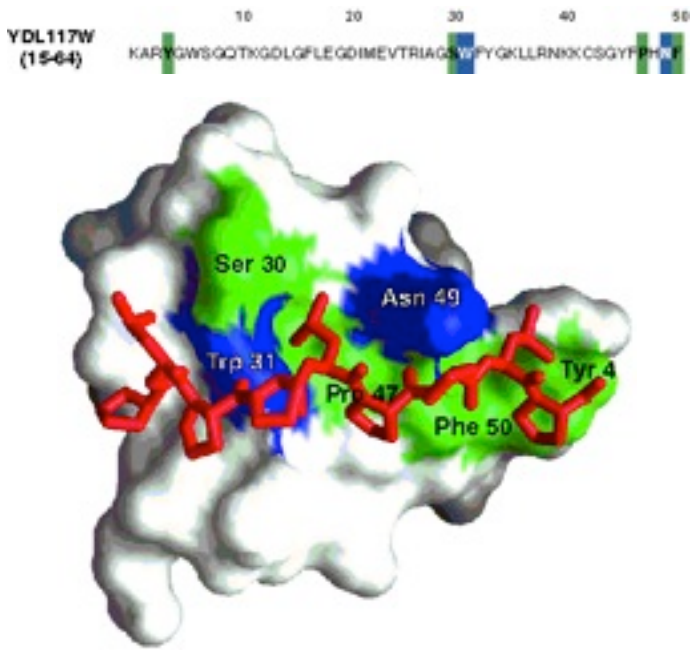
[3] Linnaeus Centre for Bioinformatics, Uppsala University, SE-751 24 Uppsala, Sweden

INTRODUCTION

Evolutionary changes of individual protein domain primary structures that become fixed in populations are mainly replacements of single amino acid residues and short insertions or deletions. Since most three-dimensional structures of proteins are determined by their sequences[1] and solvent interactions, higher-order structure will also change in response to these changes. The extent of higher-order structural perturbation in response to sequence evolution will depend on the type and location of sequence changes. Some single mutations will completely disrupt structure, while others that conserve the physicochemical properties of the sequence will barely affect structure at all.[2]

Why is it useful to know the structure of a protein, not only its sequence?

- The biochemical function (activity) of a protein is defined by its interactions with other molecules.
- The biological function is in large part a consequence of these interactions.
- The 3D structure is more informative than sequence because interactions are determined by residues that are close in space but are frequently distant in sequence.



In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence**.

The net result is that **patterns in space are frequently more recognizable than patterns in sequence**.