

Databases

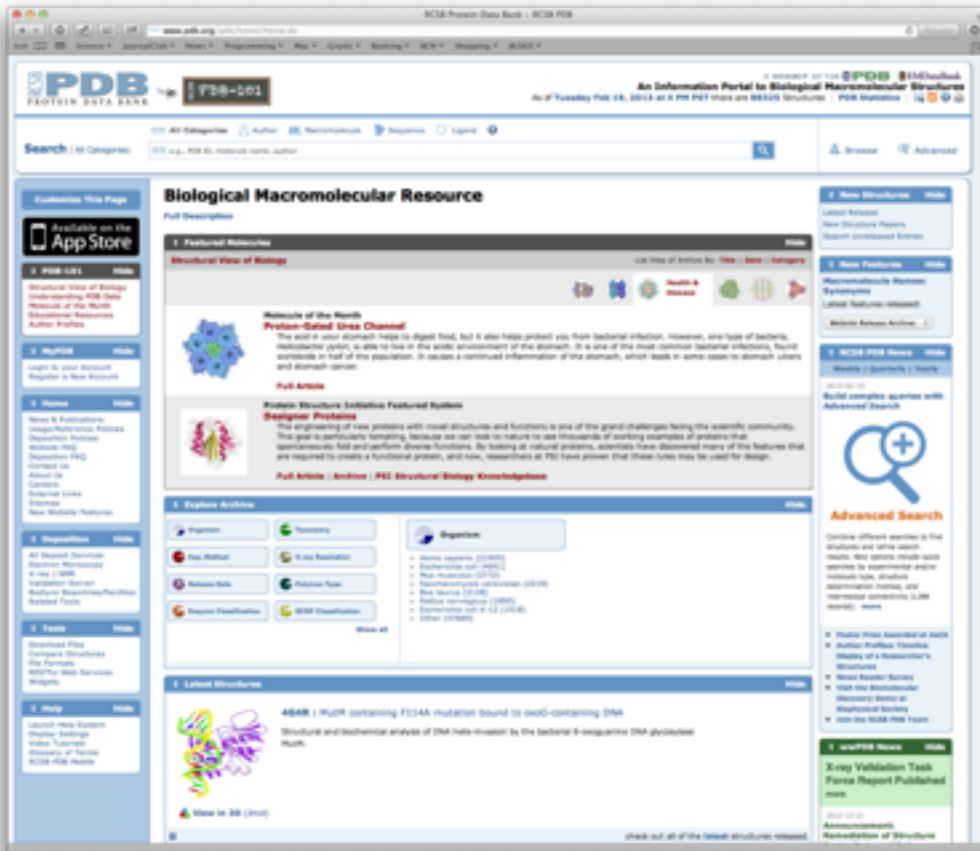
Alignment & structure classification

GOALS

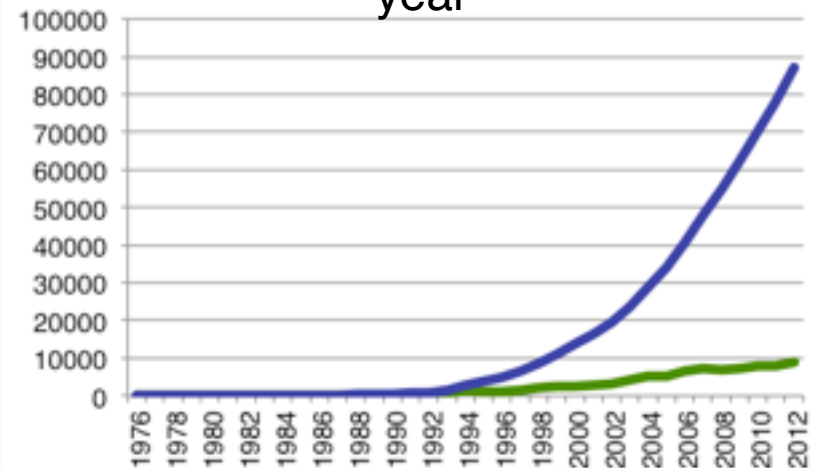
1. Known structures
2. Structure comparison
3. Structure classification
4. Number of folds in nature
5. Sequences VS fold structures

1. Known structures

PDB



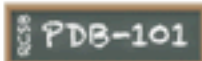
Yearly and total PDB structures per year



— Yearly

— Total

PDB search



A MEMBER OF

An Information Portal to Biological Research

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

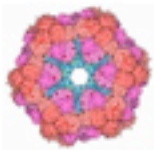
† Learn: Featured Molecules

Hide

Structural View of Biology

List View of Archive By: **Title** | Date | Category

PDB-101



Molecule of the Month

Erythrocrucorin

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 erythrocrucorins.









Biological Energy

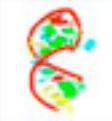




PDB search

Showing 1 - 2 of 2 Results Results : 25 Page: 1 of 1

Filter: View: Reports: Sort:



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

Search

Advanced

Browse

Everything

Author

Macromolecule

Sequence

Ligand

?

e.g., PDB ID, molecule name, author

Search History

Previous Results

↑ 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

↑ Home

Hide

News & Publications

Usage/Reference Policies

Deposition Policies

Website FAQ

Deposition FAQ

Contact Us

About Us

Careers

External Links

Sitemap

New Website Features

↑ Deposition

Hide

All Deposit Services

Advanced Search Interface

Macromolecule Type

?

Search based on whether the structure contains chains of certain molecule types (e.g. protein vs. DNA)

Contains Protein

Yes

Contains DNA

No

Contains RNA

Yes

Contains

No

DNA/RNA Hybrid

Result Count

AND

X-ray Resolution

?

Search by X-ray resolution (mmCIF item _refine.ls_d_res_high)

Between

0

and

3

Result Count

Add Search Criteria

PDB comparison tool

RCSB PDB Protein Comparison Tool

Calculate pairwise sequence or structure alignments.

Compare the following two proteins ⓘ

ID 1:

Cytoplasmic dynein 1 heavy chain 1, seryl t-RNA synthetase chimera
KQQEVIADKQMSVKEDLDK...

ID 2:

Cytoplasmic dynein 1 heavy chain 1, seryl t-RNA synthetase chimera
KQQEVIADKQMSVKEDLDK...

--- Select Comparison Method --- ⌵

--- Select Comparison Method ---

Pairwise Sequence Alignment

blast2seq
Smith-Waterman
Needleman-Wunsch

Pairwise Structure Alignment

jFATCAT - rigid
jFATCAT - flexible
jCE algorithm
jCE Circular Permutation
external server: FATCAT
external server: TM-Align
external server: TopMatch
external server: Dali

Compare

you can use the auto-suggest feature. It supports searching by

)

ptions)

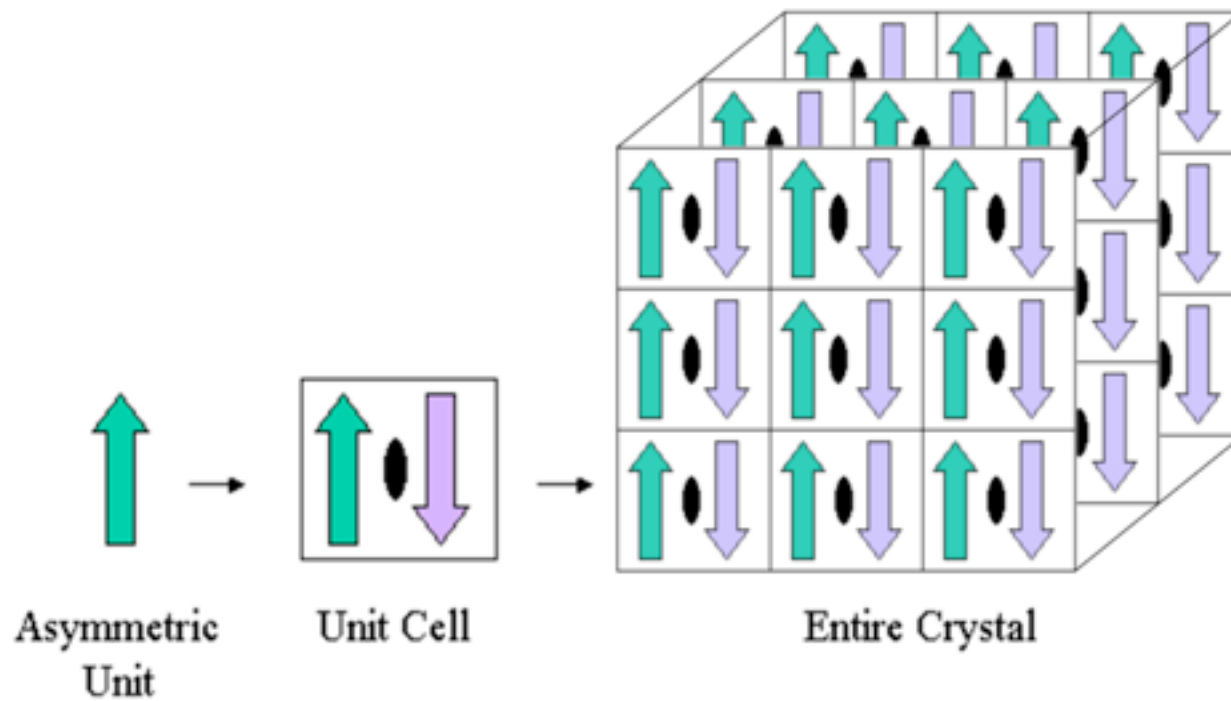
Java Web Start applications, view our [troubleshooting Java Web Start page](#) for more help.

PDB format

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

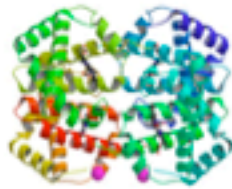
```
HEADER      MOTOR PROTEIN/STRUCTURAL PROTEIN          25-JUN-12   3J1U
TITLE       LOW AFFINITY DYNEIN MICROTUBULE BINDING DOMAIN - TUBULIN COMPLEX
COMPND      MOL_ID: 1;
COMPND      2 MOLECULE: CYTOPLASMIC DYNEIN 1 HEAVY CHAIN 1, SERYL T-RNA SYNTHETASE
COMPND      3 CHIMERA;
COMPND      4 CHAIN: 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, DHCL, DNCH1, DNCHCL, 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
```

Assymmetric Unit VS Biological



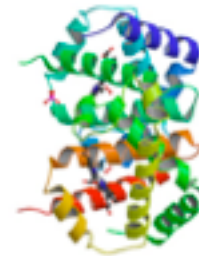
Assymmetric Unit VS Biological

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

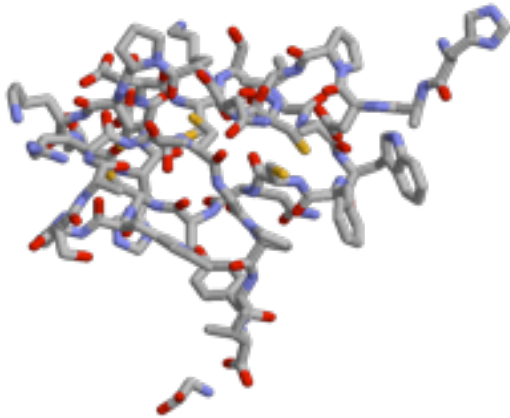
General steps in a bioinformatics procedure:

Representation

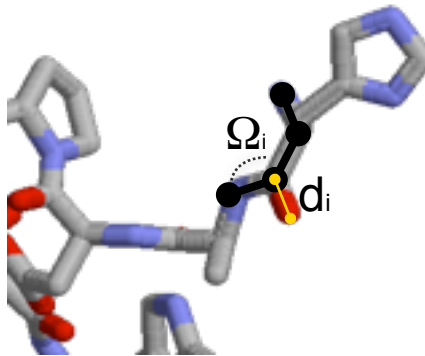
Scoring

Optimizer

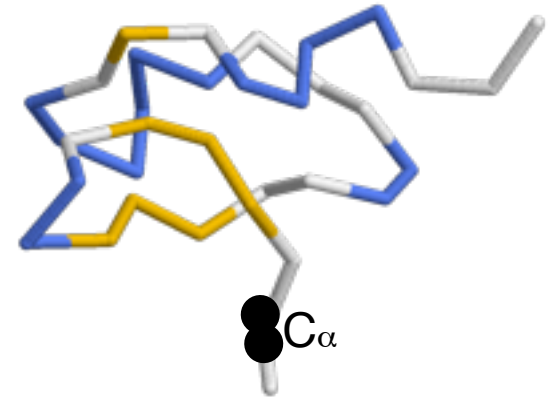
Structures



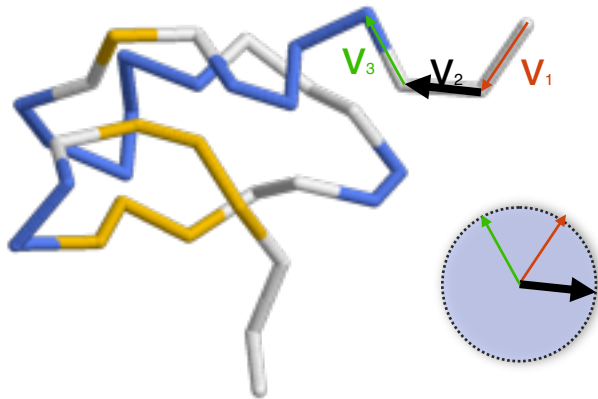
All atoms and coordinates



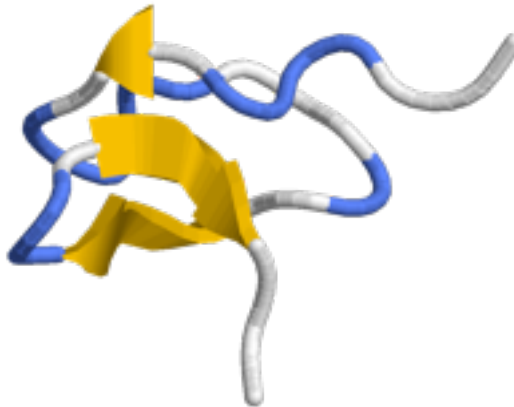
Dihedral space or distance space



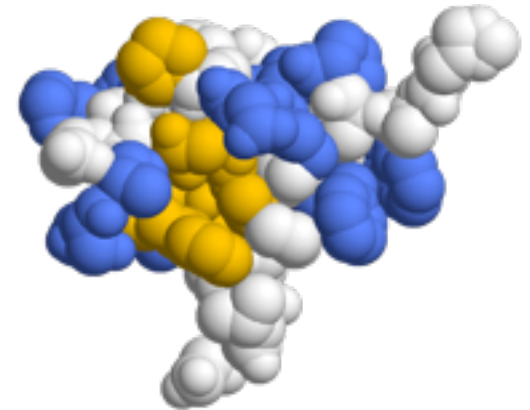
Reduced atom representation



Vector representation



Secondary Structure



Accessible surface (and others)

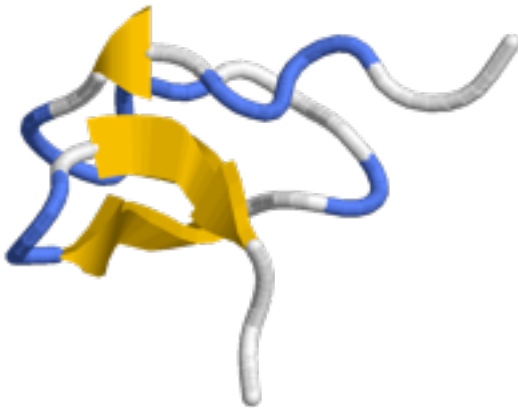
Raw scores

	C	S	T	P	A	G	N	D	E	Q	R	K	L	M	I	V	F	Y	W
C	*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
S	-1	*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
T	-1	-1	*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
P	-1	-1	-1	*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
A	-1	-1	-1	-1	*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
G	-1	-1	-1	-1	-1	*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
N	-1	-1	-1	-1	-1	-1	*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
D	-1	-1	-1	-1	-1	-1	-1	*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
E	-1	-1	-1	-1	-1	-1	-1	-1	*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Q	-1	-1	-1	-1	-1	-1	-1	-1	-1	*	-1	-1	-1	-1	-1	-1	-1	-1	-1
R	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	*	-1	-1	-1	-1	-1	-1	-1	-1
K	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	*	-1	-1	-1	-1	-1	-1	-1
L	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	*	-1	-1	-1	-1	-1	-1
M	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	*	-1	-1	-1	-1	-1
I	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	*	-1	-1	-1	-1
V	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	*	-1	-1	-1
F	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	*	-1	-1
Y	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	*	-1
W	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	*

Aminoacid substitutions

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

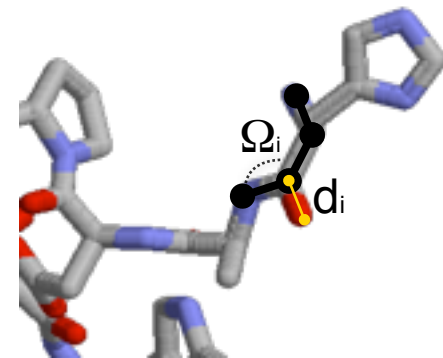
Root Mean Square Deviation



Secondary Structure (H,B,C)



Accessible surface (B,A [%])



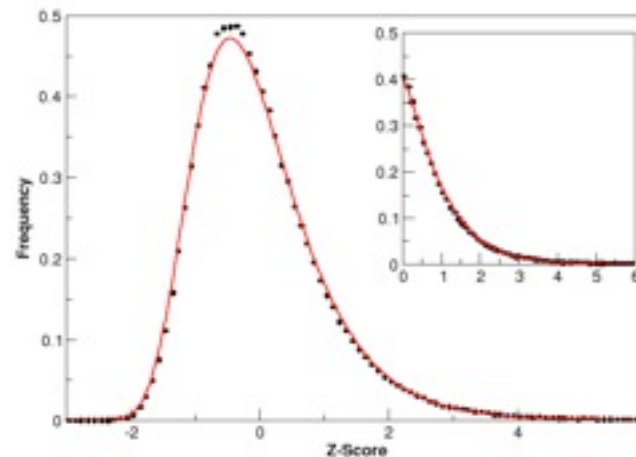
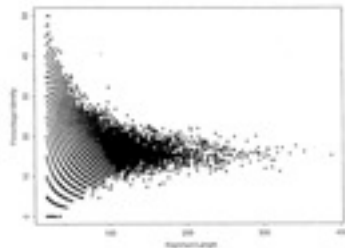
Angles or distances

Scoring

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).

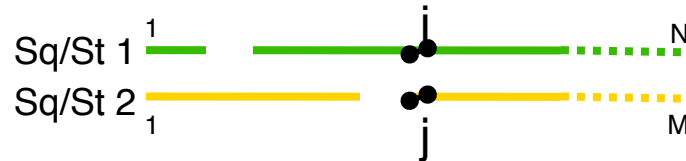
Analytic

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

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

Karlin and Altschul, 1990 PNAS 87, pp2264

Global dynamic programming alignment



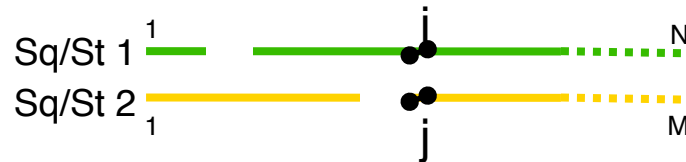
	1	2	3	...	N
1	*	*	*	*	*
2	*	*	*	*	*
3	*	*	*		
...					
M					*

$$D_{i,j} = \min \begin{cases} D_{i,j-1} + \text{Score}_{(\Delta, r_j)} \\ D_{i-1,j-1} + \text{Score}_{(r_i, r_j)} \\ D_{i-1,j} + \text{Score}_{(r_i, \Delta)} \end{cases}$$

Best alignment score

Backtracking to get the best alignment

Local dynamic programming alignment



	1	2	3	...	N
1	*	*	*	*	*
2	*	*	*	*	*
3	*	*	*	*	*
...	*	*	*	*	*
M	*	*	*	*	*

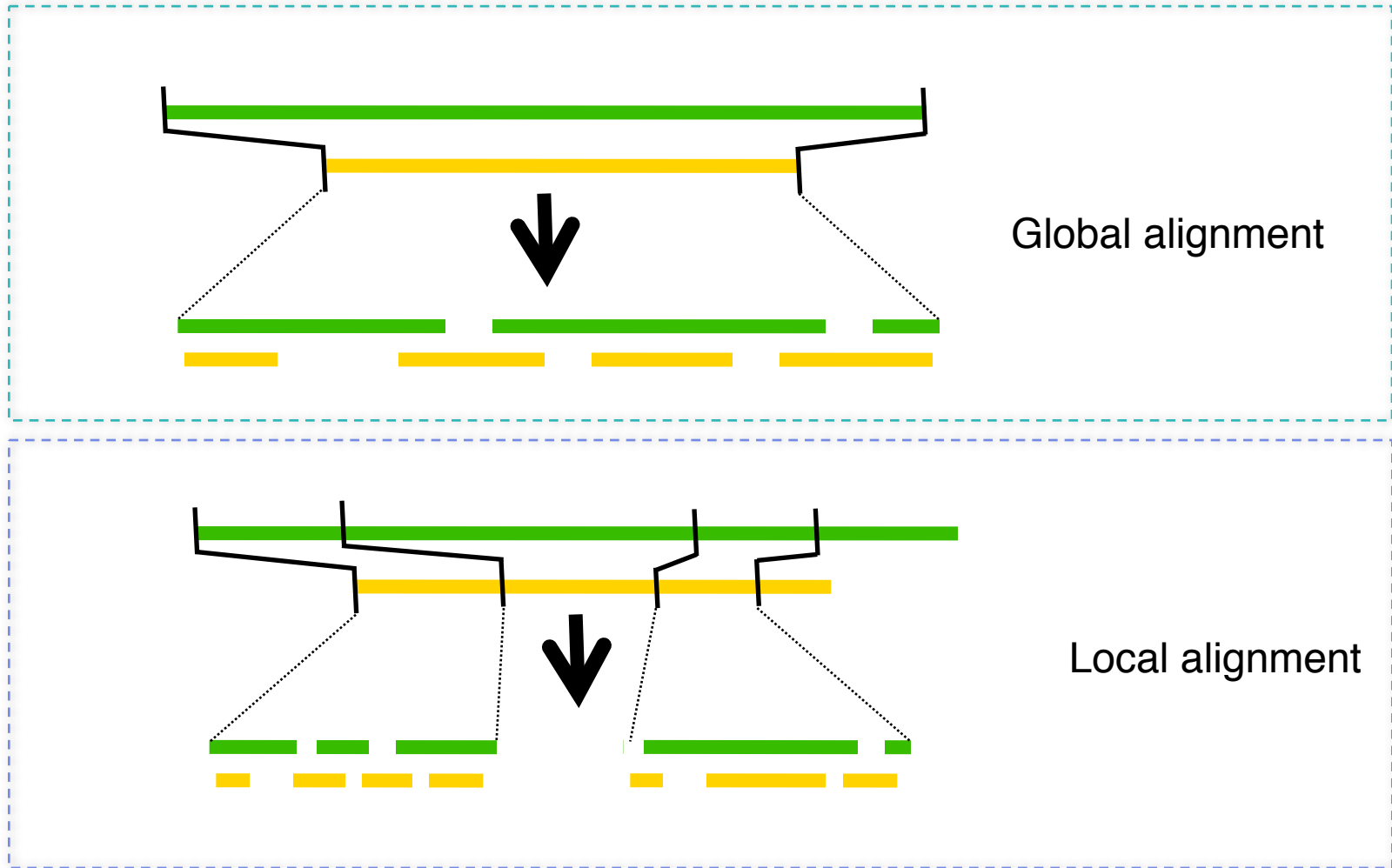
Best local alignment

Best score

$$D_{i,j} = \min \begin{cases} D_{i,j-1} + \text{Score}_{(\Delta, r_j)} \\ D_{i-1,j-1} + \text{Score}_{(r_i, r_j)} \\ D_{i-1,j} + \text{Score}_{(r_i, \Delta)} \\ 0 \end{cases}$$

Backtracking to get the best alignment

Global .vs. local alignment



Multiple alignment

Pairwise alignments

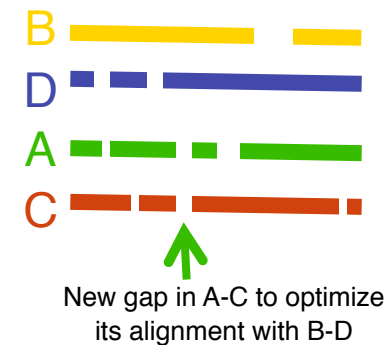
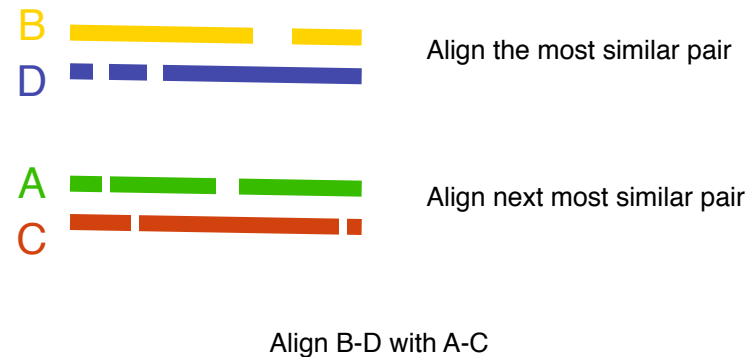
Example – 4 sequences A, B, C, D.



6 pairwise comparisons
then cluster analysis

Multiple alignments

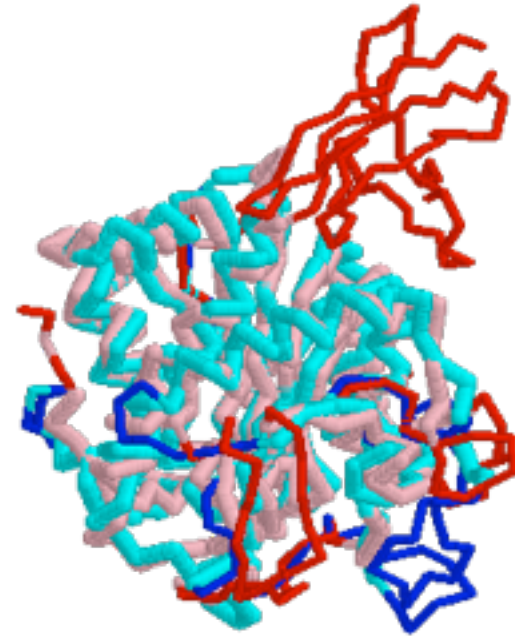
Following the tree from step 1



Coverage .vs. Accuracy



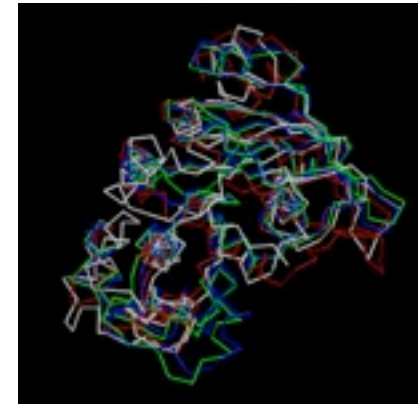
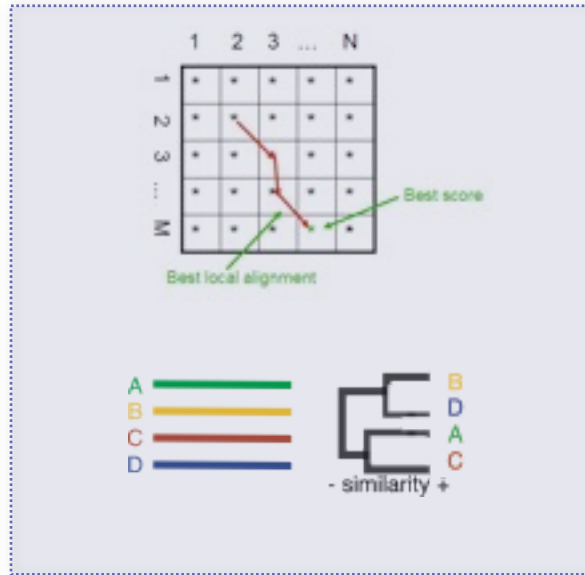
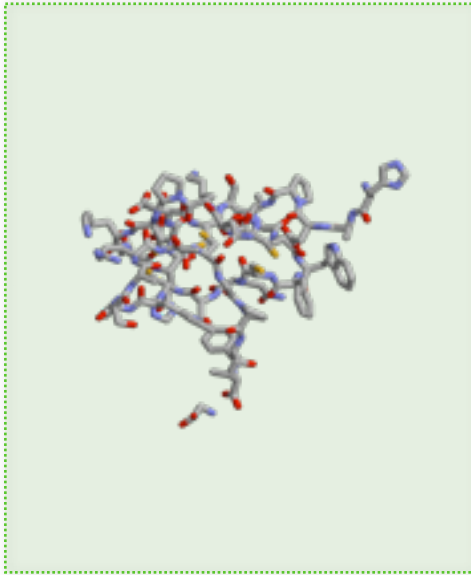
Coverage ~90% $C\alpha$



Coverage ~75% $C\alpha$

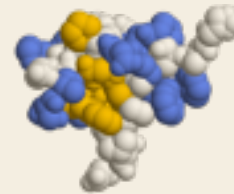
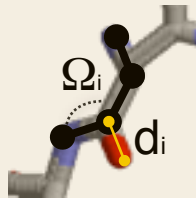
Same RMSD $\sim 2.5\text{\AA}$

Structural alignment by properties conservation (SALIGN-MODELLER)



- ✓ Uses all available structural information
- ✓ Provides the optimal alignment

Computationally expensive



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

$R_{i,j}$

$D_{,i(3),j(3)}$

$S_{i,j}$

$B_{i,j}$

$I_{i,j}$

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>



The screenshot shows the SALIGN Webserver interface. At the top, there is a logo and the text "SALIGN Webserver". Below this is a navigation bar with links: "Salign Lab Home", "ModWeb", "ModBase", "ModEval", "PCSS", "FoXS", "IMP", "MultiFit", and "ModPipe". Another navigation bar contains links: "SALIGN Home", "SALIGN Help", "SALIGN Examples", and "SALIGN Contact". The main heading reads "SALIGN: A multiple protein sequence/structure alignment server." Below this, the "Developers:" section lists: Hannes Bruberg, M.S. Madhusudan, Ursula Pieper, Ben Webb, Elina Tjoe, and Andrej Sali. A paragraph describes SALIGN as a general alignment module of the modeling program MODELLER, noting that alignments are computed using dynamic programming and that benchmarks are available. The "General information" section includes an "Email address" field. The "Input alignment information" section explains that users can upload their own sequences/structures or choose from the PDB. It features a text area for pasting sequences, an "Upload" button, and a file upload section with a "Choose File" button (showing "No file chosen") and another "Upload" button.

SALIGN Webserver

• Salign Lab Home • ModWeb • ModBase • ModEval • PCSS • FoXS • IMP • MultiFit • ModPipe •

[SALIGN Home](#) [SALIGN Help](#) [SALIGN Examples](#) [SALIGN Contact](#)

SALIGN: A multiple protein sequence/structure alignment server.

Developers:

Hannes Bruberg
M.S. Madhusudan
Ursula Pieper
Ben Webb
Elina Tjoe
Andrej Sali

SALIGN is a general alignment module of the modeling program MODELLER. The alignments are computed using dynamic programming, making use of several features of the protein sequences and structures. SALIGN benchmarks from published papers are [also available](#).

General information

Email address 

Input alignment information

Users can either upload their own sequences/structures to align or choose structures from the PDB.

Paste one sequence at a time, without header 



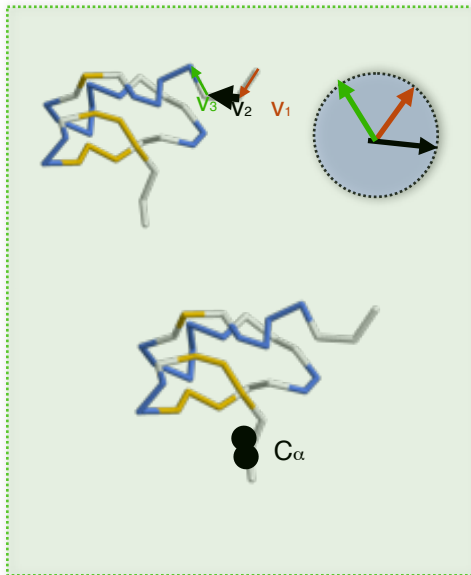


Upload sequence/PDB file(s) 

 No file chosen



Vector Alignment Search Tool (VAST)

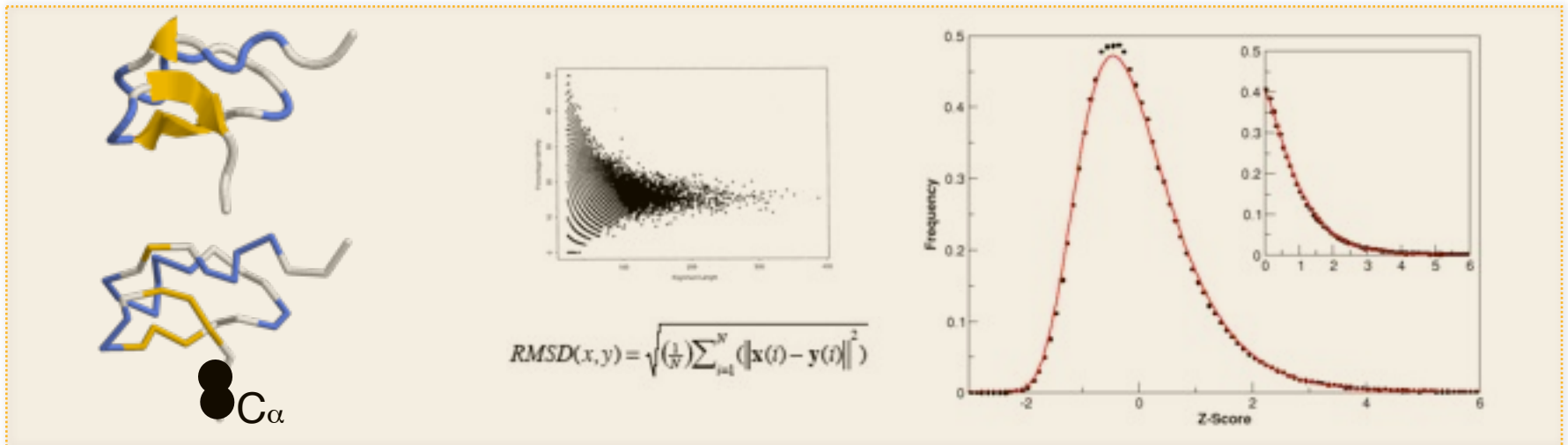


Graph theory search
of similar SSE
Refining by Monte Carlo
at all atom resolution



✓ Good scoring system with significance

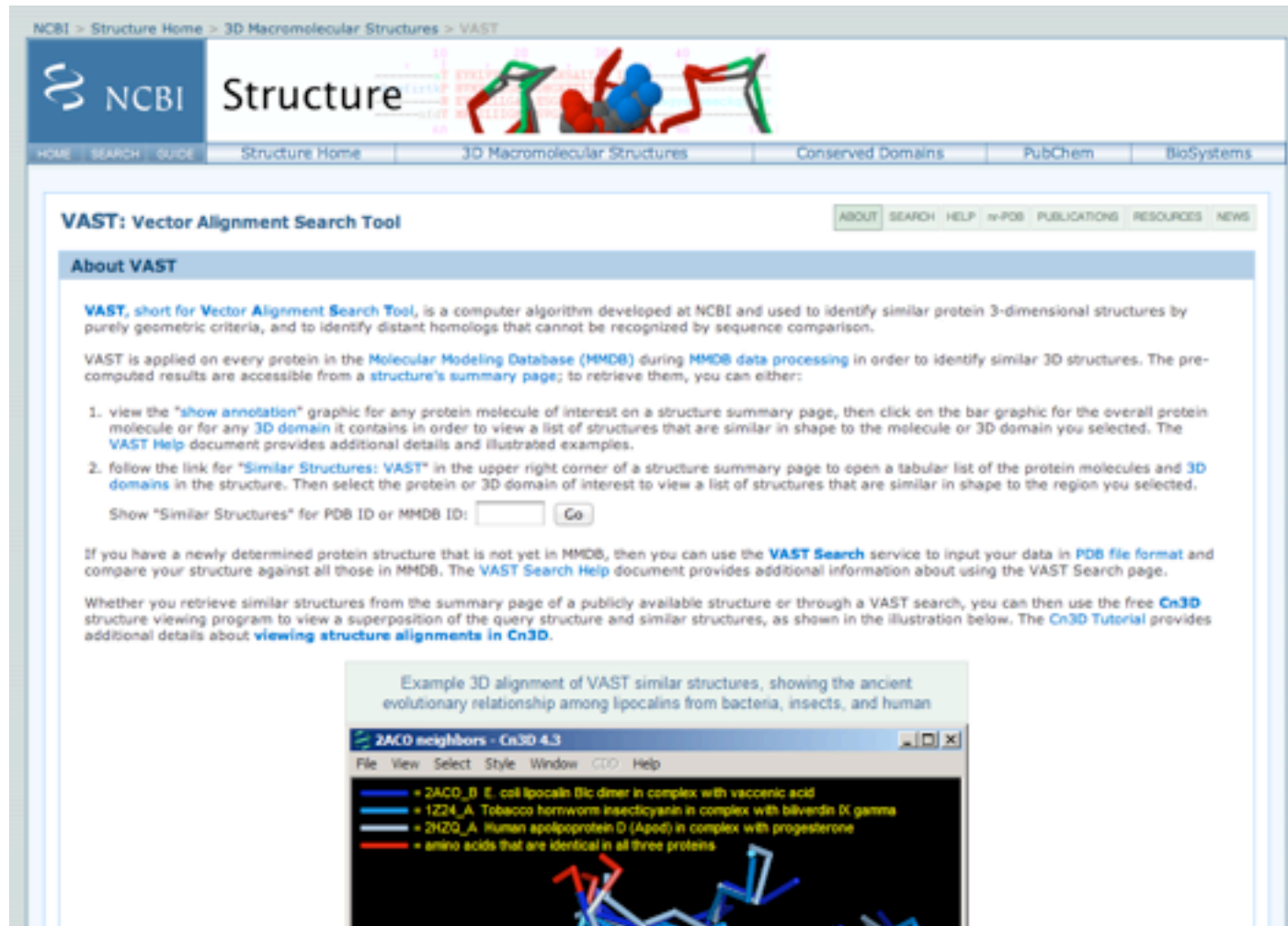
Reduces the protein representation



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>



NCBI > Structure Home > 3D Macromolecular Structures > VAST

NCBI Structure

HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains PubChem BioSystems

VAST: Vector Alignment Search Tool

ABOUT SEARCH HELP IN-POB PUBLICATIONS RESOURCES NEWS

About VAST

VAST, short for **Vector Alignment Search Tool**, is a computer algorithm developed at NCBI and used to identify similar protein 3-dimensional structures by purely geometric criteria, and to identify distant homologs that cannot be recognized by sequence comparison.

VAST is applied on every protein in the [Molecular Modeling Database \(MMDB\)](#) during [MMDB data processing](#) in order to identify similar 3D structures. The pre-computed results are accessible from a [structure's summary page](#); to retrieve them, you can either:

1. view the "show annotation" graphic for any protein molecule of interest on a structure summary page, then click on the bar graphic for the overall protein molecule or for any 3D domain it contains in order to view a list of structures that are similar in shape to the molecule or 3D domain you selected. The [VAST Help](#) document provides additional details and illustrated examples.
2. follow the link for "Similar Structures: VAST" in the upper right corner of a structure summary page to open a tabular list of the protein molecules and 3D domains in the structure. Then select the protein or 3D domain of interest to view a list of structures that are similar in shape to the region you selected.

Show "Similar Structures" for PDB ID or MMDB ID:

If you have a newly determined protein structure that is not yet in MMDB, then you can use the [VAST Search](#) service to input your data in [PDB file format](#) and compare your structure against all those in MMDB. The [VAST Search Help](#) document provides additional information about using the VAST Search page.

Whether you retrieve similar structures from the summary page of a publicly available structure or through a VAST search, you can then use the free [Cn3D](#) structure viewing program to view a superposition of the query structure and similar structures, as shown in the illustration below. The [Cn3D Tutorial](#) provides additional details about [viewing structure alignments in Cn3D](#).

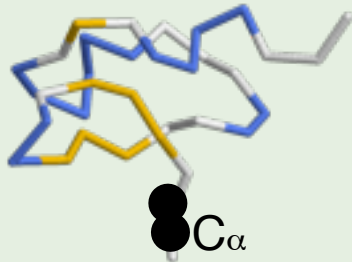
Example 3D alignment of VAST similar structures, showing the ancient evolutionary relationship among lipocalins from bacteria, insects, and human

2ACD neighbors - Cn3D 4.3

File View Select Style Window CDD Help

- 2ACD_B E. coli lipocalin B1c dimer in complex with valeric acid
- 1Z24_A Tobacco hornworm insecticyanin in complex with biliverdin IX gamma
- 2H2Q_A Human apolipoprotein D (ApoD) in complex with progesterone
- amino acids that are identical in all three proteins

Incremental combinatorial extension (CE)



Exhaustive combination
of fragments

Longest combination of
AFPs

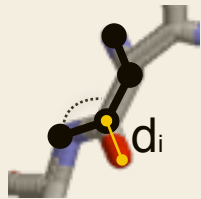
Heuristic similar to
PSI-BLAST



✓ FAST!

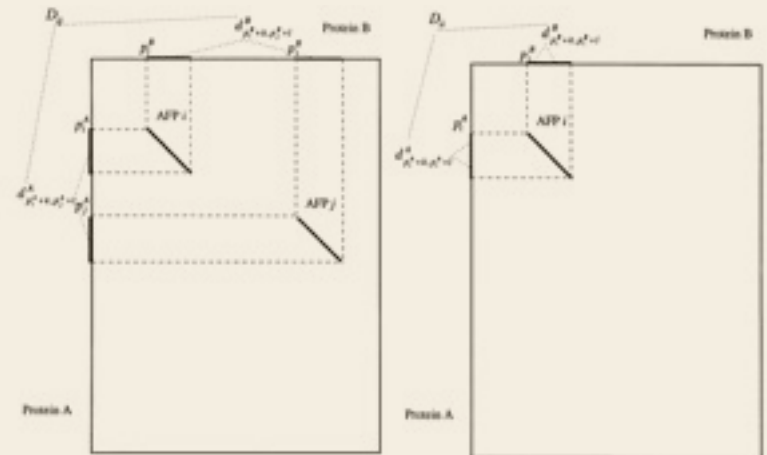
✓ Good quality of local alignments

Complicated scoring and heuristics



8 residues peptides

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



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

Incremental combinatorial extension (CE)

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



Combinatorial Extension (CE)
A method for comparing and aligning protein structures

Help

Calculate
[Two Chains](#)
[Symmetry](#)
[DB Search](#)
[All vs All](#)

Download
[Download jCE/jFatCat](#)
[Documentation](#)

History
[Release History](#)
[CE History](#)

Links
[Multiple Structure Alignment](#)

Results
[Subdomains](#) (PDF )

CruiseControl
[CruiseControl](#)



Combinatorial Extension (CE)

A method for comparing and aligning protein structures

This page is intended as a pointer to get you to the most recent information on CE and to enable you to perform the calculations you need. CE is now an integral part of the [RCSB Protein Data Bank](#) (PDB) and continues to be developed in the [Bourne laboratory](#) as needed.

Key Pointers

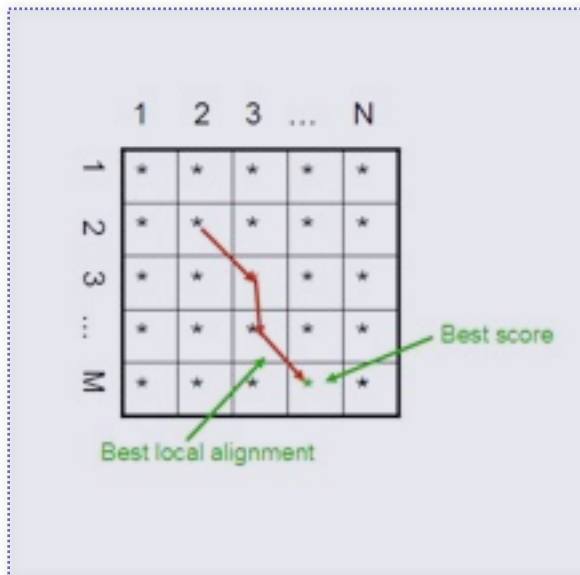
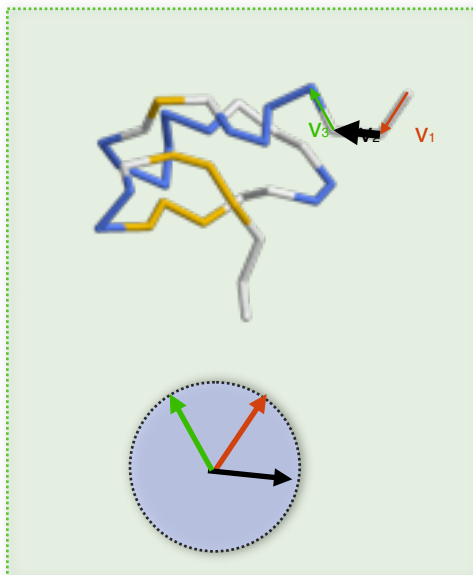
- Access to CE from the RCSB PDB <http://www.rcsb.org/pdb/workbench/workbench.do>
- Standalone server <http://source.rcsb.org/jfatcatserver/>
- Access to the CE code in Java (jCE) and the original source <http://source.rcsb.org/jfatcatserver/download.jsp>
- [Legacy](#) CE - web site

What follows is a brief description of the history of CE and some additional references and pointers.

Chronology

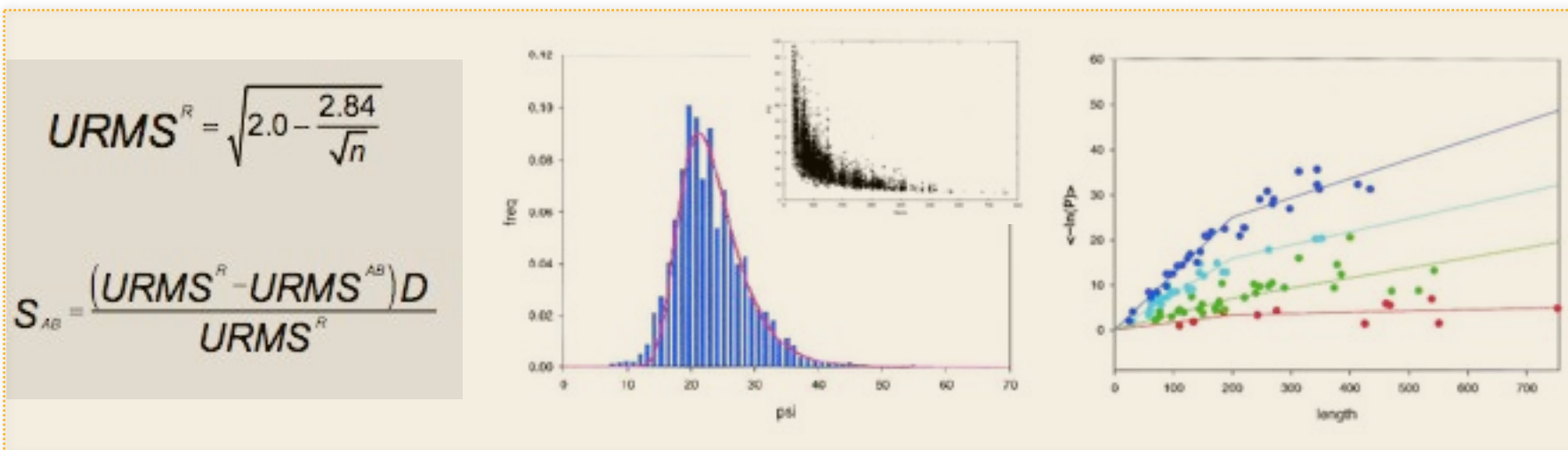
- 1998 - CE method released and original paper published [1]
- 2000 - CE used to map existing protein fold space [2]
- 2001 - Pairwise alignment database made available [3]

Matching molecular models obtained from theory (MAMMOTH)



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

Reduces the protein representation



Matching molecular models obtained from theory (MAMMOTH)

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

Bioinformatics Unit

Centro de Biología Molecular Severo Ochoa



- Home
- Publications
- Research
- Services
- Software

Currently we are under testing, so this service may not be functioning properly. Sorry for the inconvenience



MAMMOTH

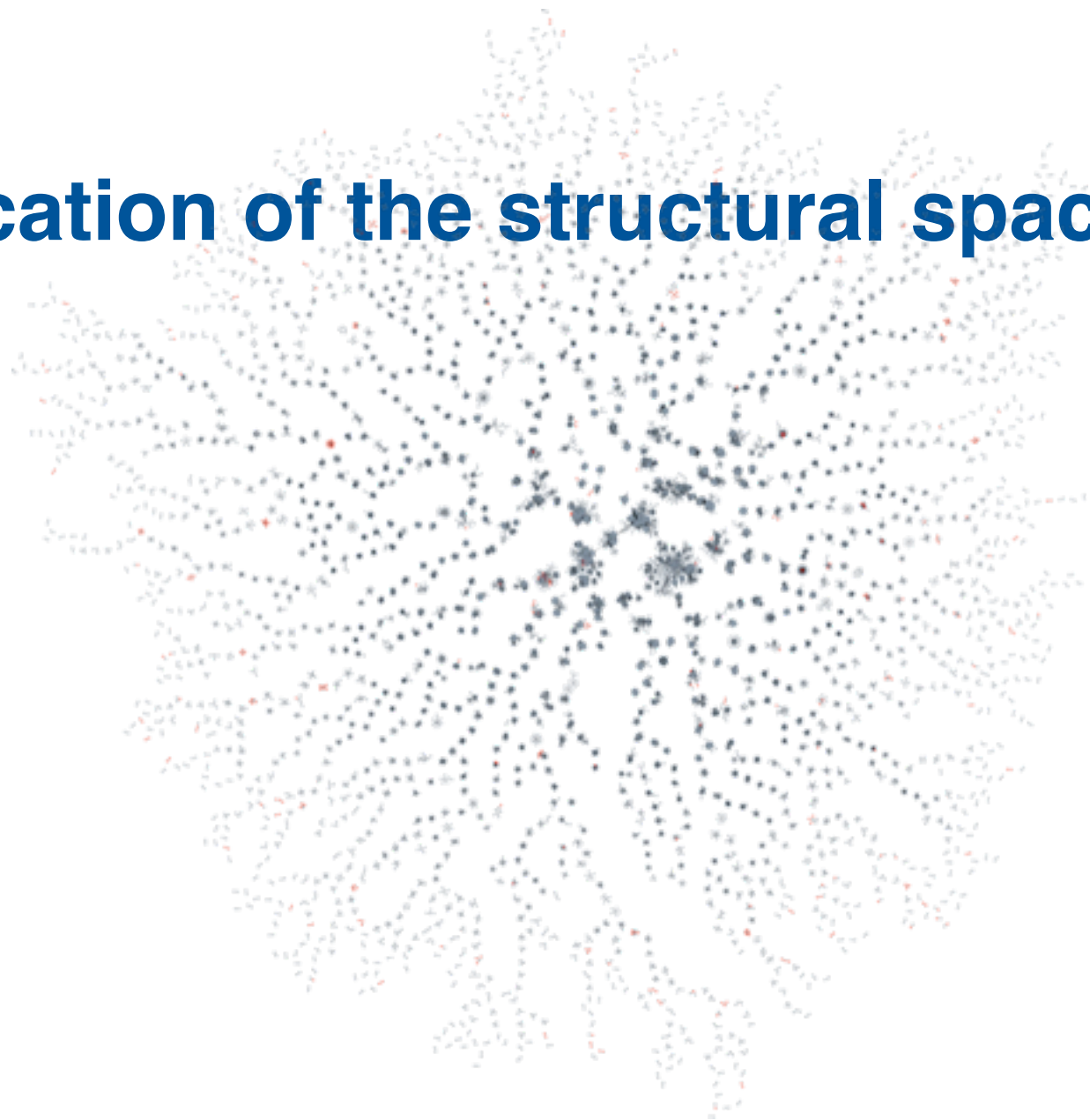
- MAMMOTH (Matching Molecular Models Obtained from Theory) es un método de alineamiento estructural de proteínas independiente de su secuencia. Esto permite la comparación de la estructura experimental de una proteína con un modelo arbitrario de baja resolución. También permite la comparación de dos estructuras experimentales, así como la búsqueda de estructuras similares en una base de datos.
- Versión: 1.0
- Uso gratuito para fines educativos y de investigación.
- Referencia: Ortiz AR, Strauss CE, Olmea O (2002) *Protein Sci.* 11:2606-21.

Alinea tus propias proteínas.

- Sube el archivo de coordenadas (PDB) de tu **primera proteína**:
- Sube el archivo de coordenadas (PDB) de tu **segunda proteína**:
- Tu **correo electrónico** para el envío de los resultados:

3. Structure classification

Classification of the structural space



SCOP_{1.75} database

<http://scop.berkeley.edu/>

Murzin A. G., et al. (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

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

[Browse](#) [Stats & History](#) [ASTRAL Subsets](#) [Downloads](#) [Related Resources](#) [References](#) [Help](#) [About](#)

Welcome to the new SCOP+ASTRAL website!

This release is part of a series of planned releases based on SCOP 1.75 before the advent of a major reclassification, **SCOP 2.0**.

This website now provides integrated access to data previously found in the [SCOP](#) and [ASTRAL](#) databases. For prior releases of SCOP and ASTRAL, click on the [Stats & History](#) tab above. For more info, click on the [About](#) tab above.

Search SCOP (example):

Classes in SCOP 1.75B:

1.  [a: All alpha proteins](#) [46456] (284 folds)
2.  [b: All beta proteins](#) [48724] (174 folds)
3.  [c: Alpha and beta proteins \(a/b\)](#) [51349] (147 folds)
4.  [d: Alpha and beta proteins \(a+b\)](#) [53931] (376 folds)
5.  [e: Multi-domain proteins \(alpha and beta\)](#) [56572] (66 folds)
6.  [f: Membrane and cell surface proteins and peptides](#) [56835] (57 folds)
7.  [g: Small proteins](#) [56992] (90 folds)
8.  [h: Coiled coil proteins](#) [57942] (7 folds)
9.  [i: Low resolution protein structures](#) [58117] (25 folds)
10.  [j: Peptides](#) [58231] (120 folds)
11.  [k: Designed proteins](#) [58788] (44 folds)



Copyright © 1994-2013 The SCOP and ASTRAL authors
scop@mrc-lmb.cam.ac.uk and astral@compbio.berkeley.edu

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 [\(example\)](#):

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) 
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)
 1.  Species [Chromatium vinosum](#) [[TaxId:1049](#)] [46684] (1 PDB entry)

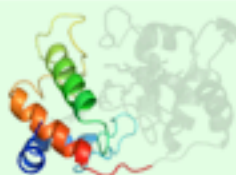
More info for Family a.3.1.4: Two-domain cytochrome c

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]}  
agdaeaagqgkvavcgachgvdgngspapnfpklaggqeryllkqlqdikagstpgapegvg  
rkvlemtgmldplsdqdlediaayfssqkgsv
```

SCOP Domain Coordinates for d1etpa1:

Click to download the [PDB-style file with coordinates for d1etpa1](#).

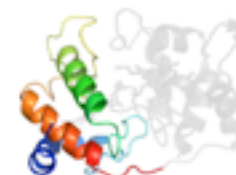
(The format of our PDB-style files is described [here](#).)

Timeline for d1etpa1:



[d1etpa1](#)

([click for larger image](#))



[d1etpa1 in context of chain](#)

Domains from same chain:

[d1etpa2](#)

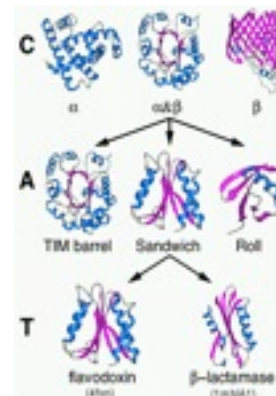
CATH_{3.5} database

<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.

The screenshot shows the CATH / Gene3D website. The header includes the CATH logo and a search bar. The main content area features the text "CATH / Gene3D" and "16 million protein domains classified into 2,626 superfamilies". Below this are four buttons: "Get Started", "Search", "Download", and "Take the Tour". The "What's New?" section mentions a recent website overhaul. The "Searching CATH" section lists search options: "Search by ID / keyword", "Search by FASTA sequence", and "Search by PDB structure". The "Example pages" section lists links to "PDB '1dan'", "Domain '1oukA01'", "Relatives of '1oukA01'", "Superfamily 'HUPs'", and "Functional Family".

Browse - tree

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

H Nop10-like SnoRNP

[View Superfamily](#)

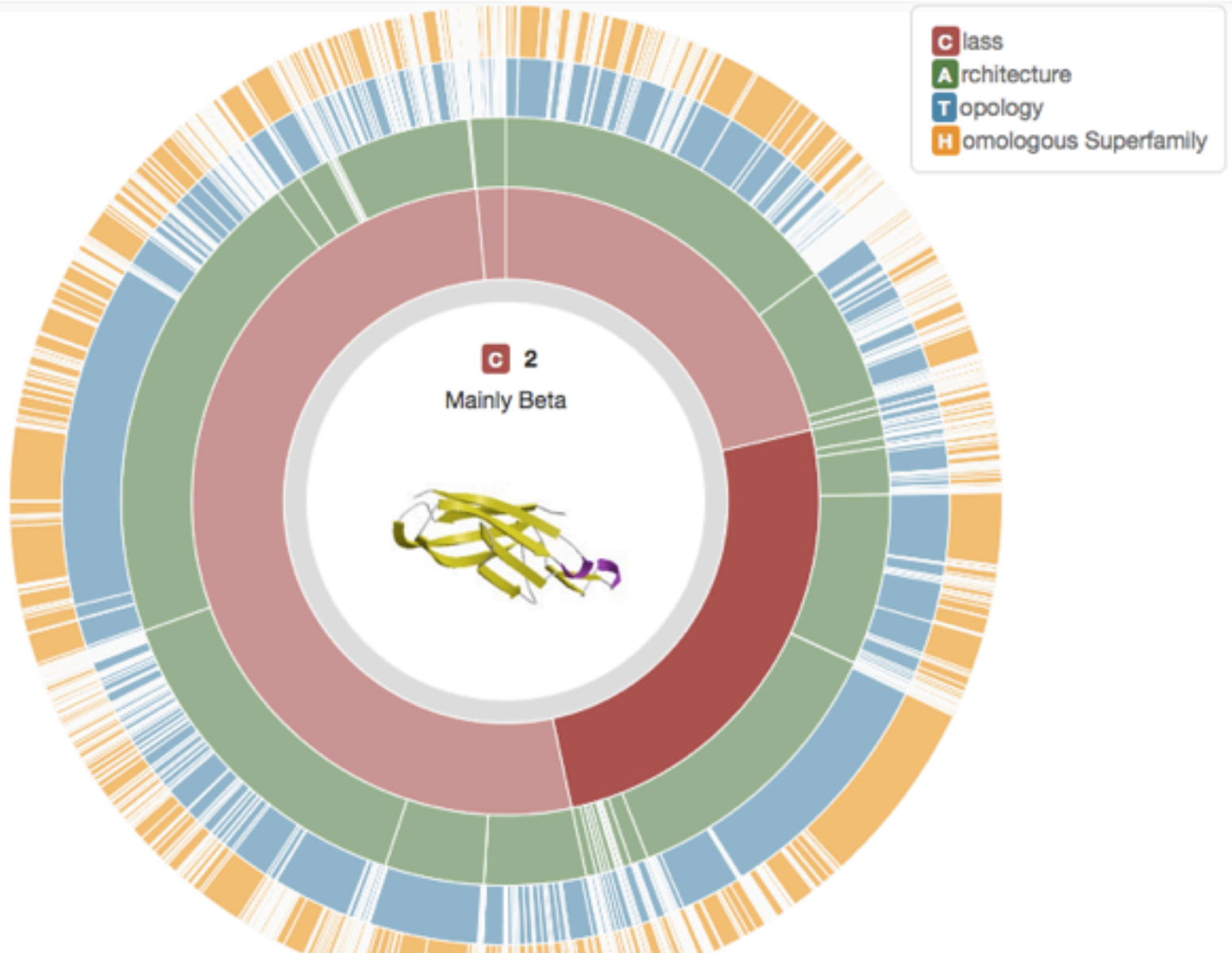
CATH ID	2.20.28.40
Non-redundant Sequences (<35% seq id)	1
Domains	3
Example Domain	2apo800 [PDB]



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	Rubryerythrin, 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
▷ T 2.20.140	Not yet named	1 Superfamilies, 1 Domains

Browse - sunburst



SUPERFAMILY LINKS

Summary

[Superfamily Superposition](#)[Classification / Domains](#)[Alignments](#)[Structural Neighbourhood](#)[Functional Annotations](#)[Taxonomy](#)[Multi-Domain Organisation](#)

Functional Families

Overview of the Structural Clusters (SC) and Functional Families (FF) within this CATH Superfamily



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

Structural Diversity

Structural domains within this superfamily



Domain Organisation

View multi-domain architectures via ArchSchema (Laskowski/EBI)



Enzyme Function

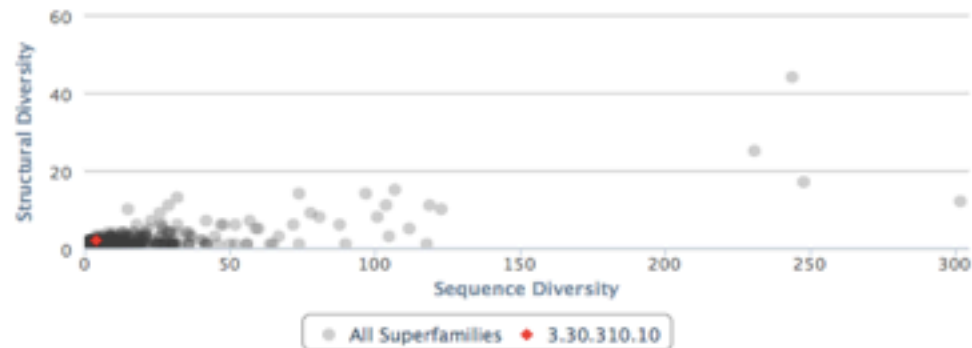
Evolution of Enzyme Function via FunTree (Furnham/EBI)





Sequence/Structure Diversity

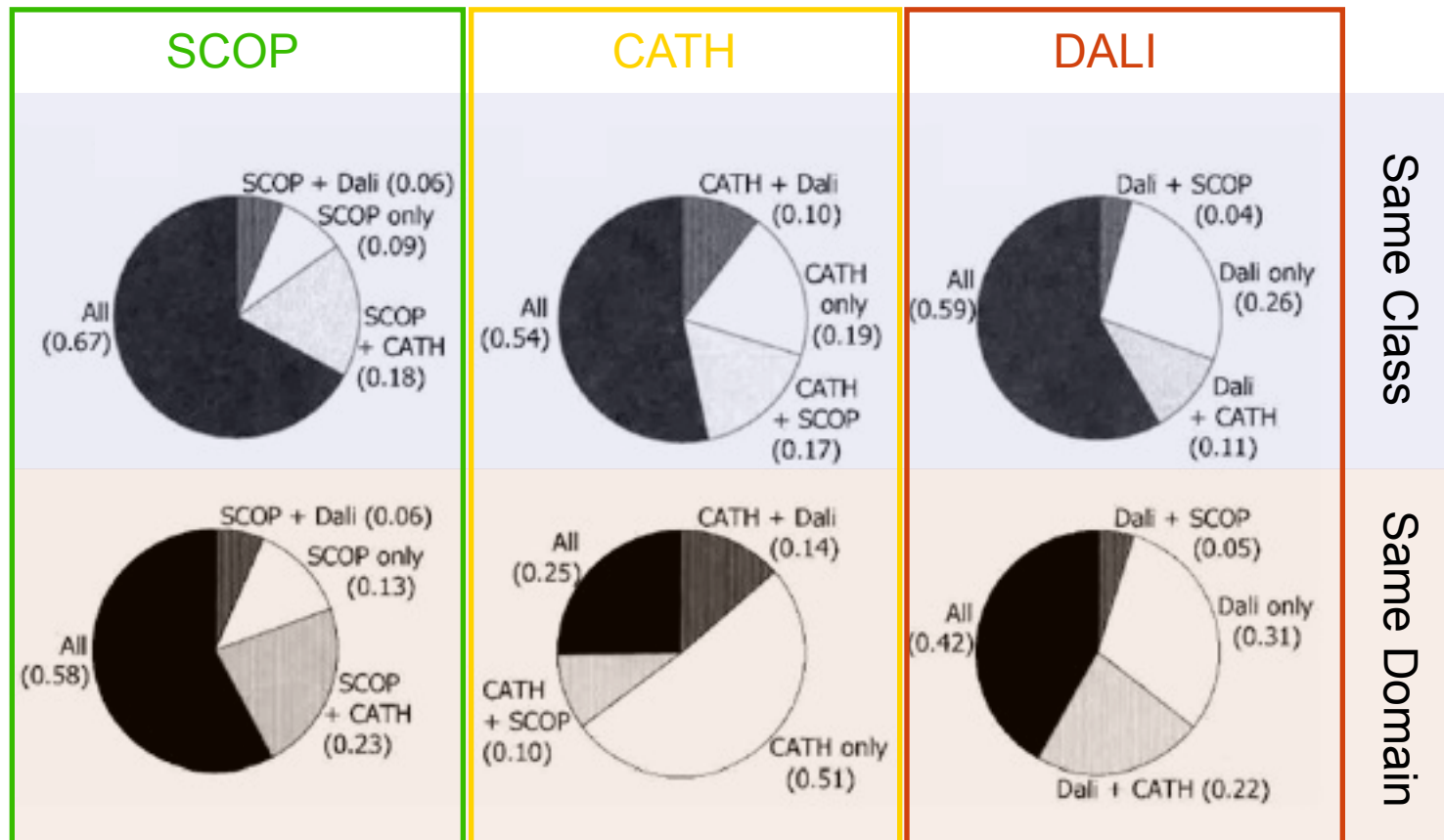
Overview of the sequence / structure diversity of this superfamily compared to other superfamilies in CATH. Click on the chart to view the data in more detail.



Classification of the structural space

Not an easy task!

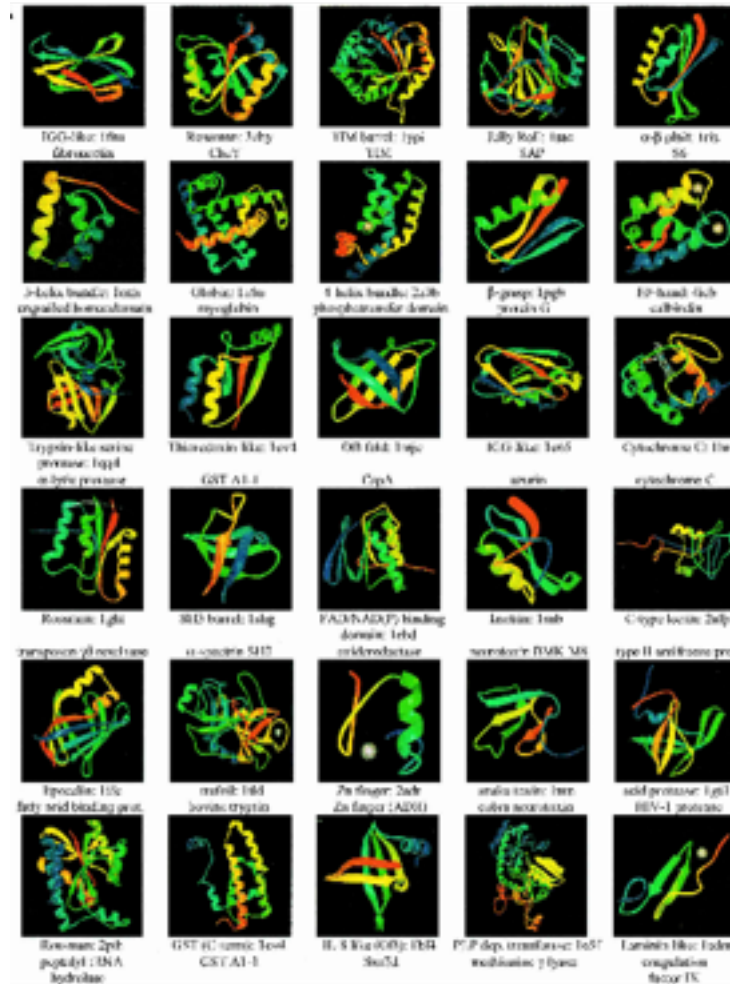
Domain definition AND domain classification

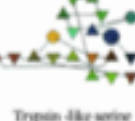
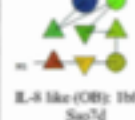


Day, et al. (2003) Protein Sciences, 12 pp2150

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

4. Number of folds in nature



				
EGG-like: 1fba fibronectin	Rossmann: 3chv CheY	TBM barrel: 1yp1 TBM	Jelly roll: 1uc SAP	a-b pleix: 1ris S6
				
3-helix bundle: 1emb Engrailed Homeodomain	Globin: 1afn myoglobin	4-helix bundle: 2afh phosphotransfer domain	b-group: 1pgh Protein G	EF-hand: 6ch calbindin
				
Trypsin-like serine protease: 1q44 trypsin	Thioredoxin-like: 1ev4 GST A1-1	OB fold: 1mpe CapA	EGG-like: 1e65 azurin	Cytochrome C: 1bnc cytochrome C
				
Rossmann: 1ght transposon yb resolvase	SH3 barrel: 1ahg α -spectrin SH3	FAD/NAD(P) binding domain: 1ebl oxidoreductase	knottin: 1abn neurotoxin BMR MB	C-type lectin: 2afp type II asialofetate prot.
				
Lipocalin: 1lfc fatty acid binding prot.	trefoil: 1fd bovine trypsin	Zn finger: 2zfr Zn finger (ADR)	snake toxin: 1tm cobra neurotoxin	acid protease: 1g6l HIV-1 protease
				
Rossmann: 2pth peptidyl tRNA hydrolase	GST (C-term): 1ev4 GST A1-1	IL-8 like (OB): 1b64 Saa7d	PLP dep. transaminase: 1e5f methionine γ -lyase	Laminin-like: 1odn coagulation factor IX

5. Sequences VS fold structures

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*}

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

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

³Linnaeus 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 sequences¹ 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

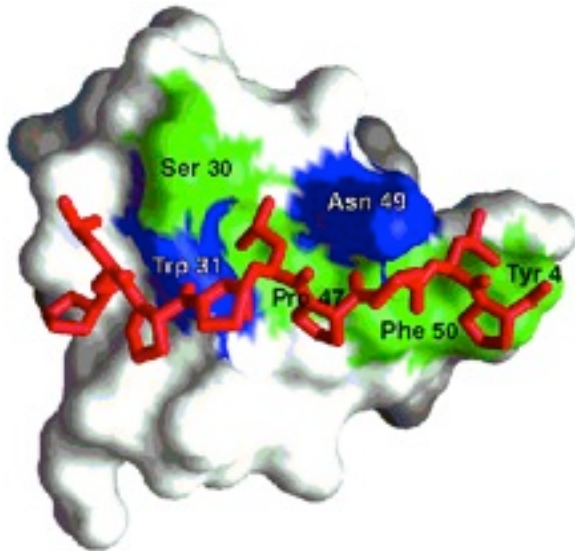
Why is it useful to know the **structure** of a protein, not only its sequence?

- ◆ 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.

YDL117W
(15-64)

10 20 30 40 50

K A R T G W S G Q T X G D L G F L E G D I M E V T R I A G S Y P Y G K L L R N K X C S G Y P P H L F



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.**