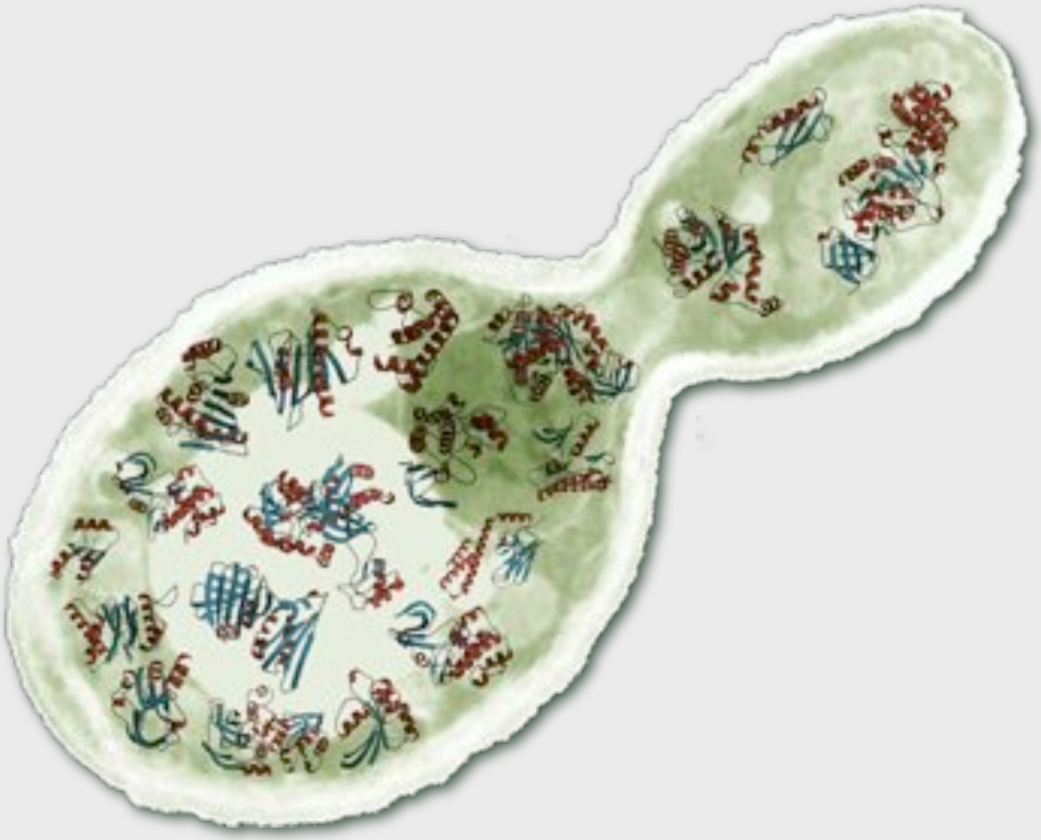
Comparative Protein Structure Prediction



Marc A. Marti-Renom

http://bioinfo.cipf.es/squ/

Structural Genomics Unit
Bioinformatics Department
Prince Felipe Resarch Center (CIPF), Valencia, Spain

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Program

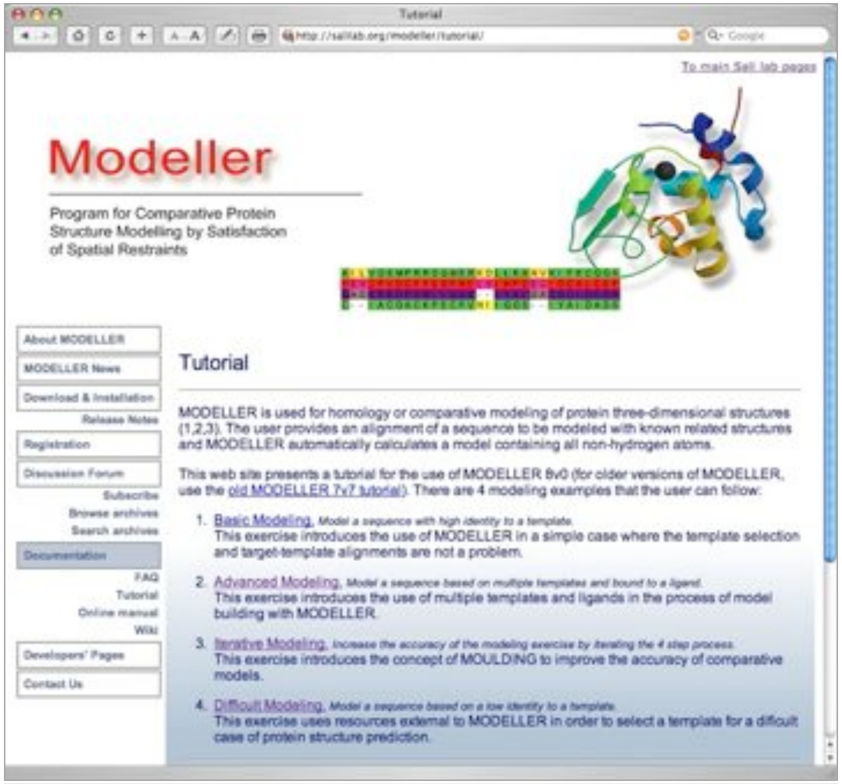
Intro to comparative
protein structure prediction

Template Search

Target – Template Alignment

Model Building

Model Evaluation



Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

Tutorial

MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2,3). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms.

This web site presents a tutorial for the use of MODELLER 8v0 (for older versions of MODELLER, use the old MODELLER 7v7 tutorial). There are 4 modeling examples that the user can follow:

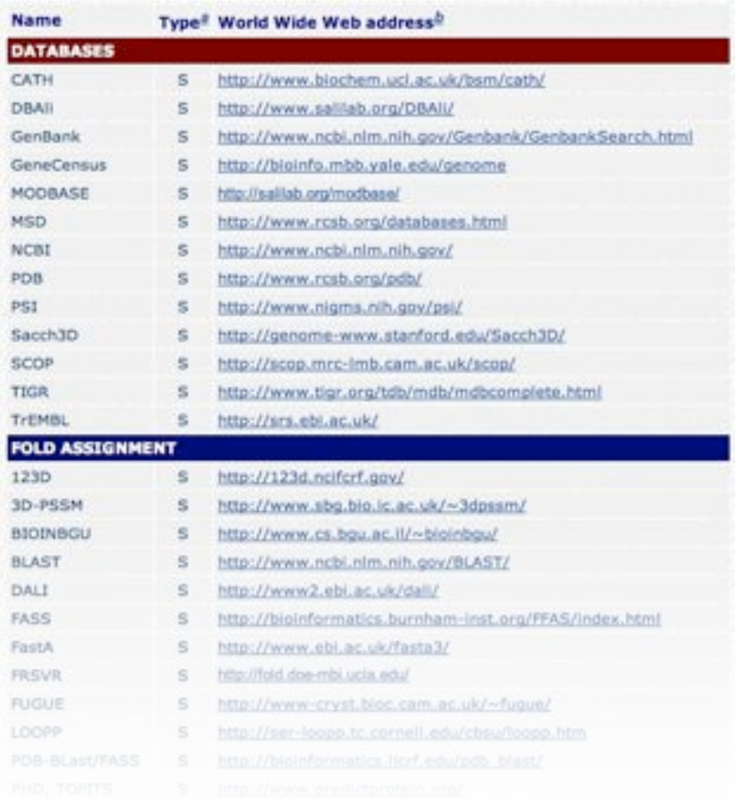
1. Basic Modeling. *Model a sequence with high identity to a template.*
   This exercise introduces the use of MODELLER in a simple case where the template selection and target-template alignments are not a problem.
2. Advanced Modeling. *Model a sequence based on multiple templates and bound to a ligand.*
   This exercise introduces the use of multiple templates and ligands in the process of model building with MODELLER.
3. Iterative Modeling. *Increase the accuracy of the modeling exercise by iterating the 4 step process.*
   This exercise introduces the concept of MOULDING to improve the accuracy of comparative models.
4. Difficult Modeling. *Model a sequence based on a low identity to a template.*
   This exercise uses resources external to MODELLER in order to select a template for a difficult case of protein structure prediction.

http://www.salilab.org/modeller/tutotial/

Objective

TO LEARN HOW-TO MODEL A
3D-STRUCTURE FROM A SEQUENCE
AND A KNOWN STRUCTURE

DISCLAIMER!

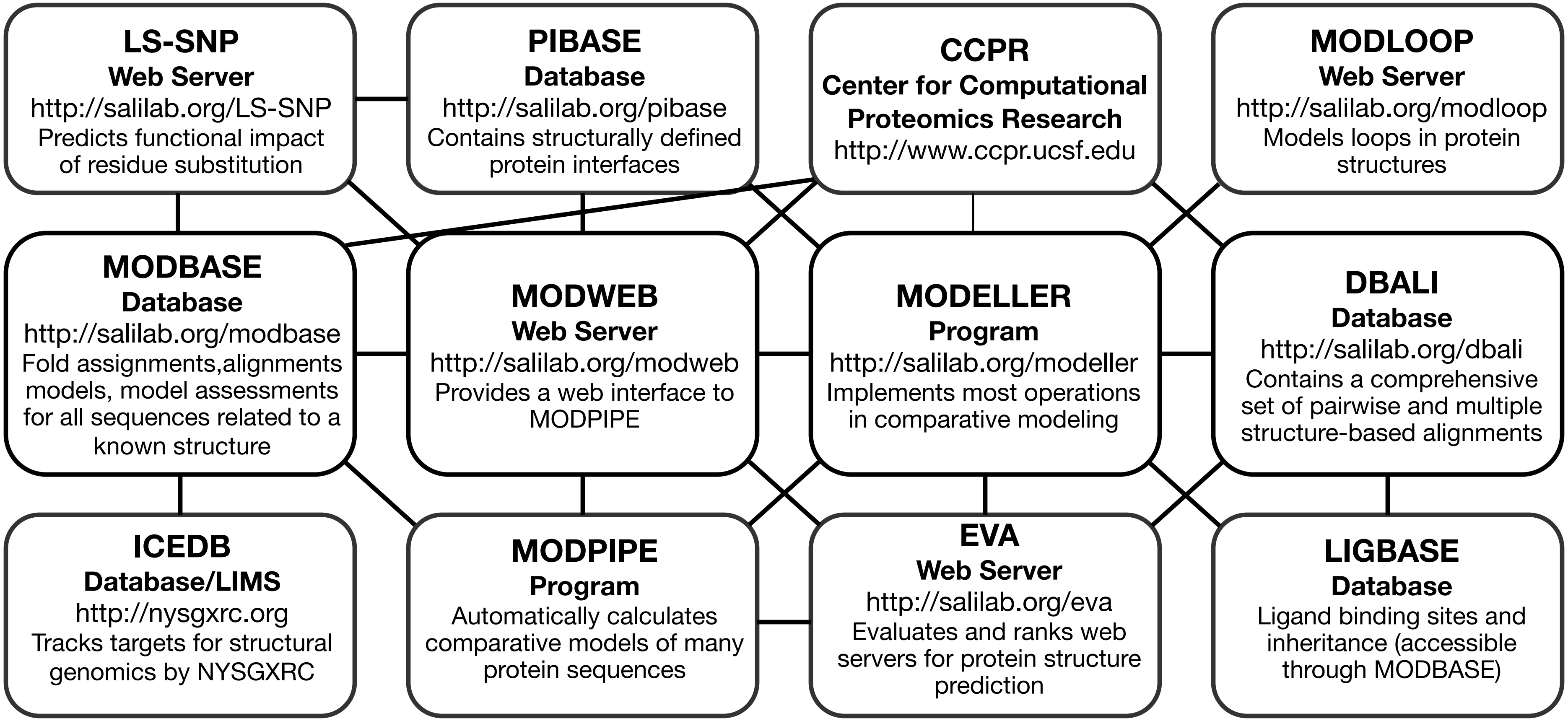


Name	Type[a]	World Wide Web address[b]
DATABASES		
CATH	S	http://www.biochem.ucl.ac.uk/bsm/cath/
DBAli	S	http://www.salilab.org/DBAli/
GenBank	S	http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html
GeneCensus	S	http://bioinfo.mbb.yale.edu/genome
MODBASE	S	http://salilab.org/modbase/
MSD	S	http://www.rcsb.org/databases.html
NCBI	S	http://www.ncbi.nlm.nih.gov/
PDB	S	http://www.rcsb.org/pdb/
PSI	S	http://www.nigms.nih.gov/psi/
Sacch3D	S	http://genome-www.stanford.edu/Sacch3D/
SCOP	S	http://scop.mrc-lmb.cam.ac.uk/scop/
TIGR	S	http://www.tigr.org/tdb/mdb/mdbcomplete.html
TrEMBL	S	http://srs.ebi.ac.uk/
FOLD ASSIGNMENT		
123D	S	http://123d.ncifcrf.gov/
3D-PSSM	S	http://www.sbg.bio.ic.ac.uk/~3dpssm/
BIOINBGU	S	http://www.cs.bgu.ac.il/~bioinbgu/
BLAST	S	http://www.ncbi.nlm.nih.gov/BLAST/
DALI	S	http://www2.ebi.ac.uk/dali/
FASS	S	http://bioinformatics.burnham-inst.org/FFAS/index.html
FastA	S	http://www.ebi.ac.uk/fasta3/
FRSVR	S	http://fold.doe-mbi.ucla.edu/
FUGUE	S	http://www-cryst.bioc.cam.ac.uk/~fugue/
LOOPP	S	http://ser-loopp.tc.cornell.edu/cbsu/loopp.htm
PDB-BLast/FASS	S	http://bioinformatics.ljcrf.edu/pdb_blast/
PHD, TOPITS	S	http://www.predictprotein.org/

http://sgu.bioinfo.cipf.es/home/?page=resources

Programs, servers and databases

http://salilab.org



LS-SNP
Web Server
http://salilab.org/LS-SNP
Predicts functional impact of residue substitution

PIBASE
Database
http://salilab.org/pibase
Contains structurally defined protein interfaces

CCPR
Center for Computational Proteomics Research
http://www.ccpr.ucsf.edu

MODLOOP
Web Server
http://salilab.org/modloop
Models loops in protein structures

MODBASE
Database
http://salilab.org/modbase
Fold assignments,alignments models, model assessments for all sequences related to a known structure

MODWEB
Web Server
http://salilab.org/modweb
Provides a web interface to MODPIPE

MODELLER
Program
http://salilab.org/modeller
Implements most operations in comparative modeling

DBALI
Database
http://salilab.org/dbali
Contains a comprehensive set of pairwise and multiple structure-based alignments

ICEDB
Database/LIMS
http://nysgxrc.org
Tracks targets for structural genomics by NYSGXRC

MODPIPE
Program
Automatically calculates comparative models of many protein sequences

EVA
Web Server
http://salilab.org/eva
Evaluates and ranks web servers for protein structure prediction

LIGBASE
Database
Ligand binding sites and inheritance (accessible through MODBASE)

External Resources
PDB, Uniprot, GENBANK, NR, PIR, INTERPRO, Kinase Resource
UCSC Genome Browser, CHIMERA, Pfam, SCOP, CATH

Nomenclature

Homology: Sharing a common ancestor, may have similar or dissimilar functions

Similarity: Score that quantifies the degree of relationship between two sequences.

Identity: Fraction of identical aminoacids between two aligned sequences (case of similarity).

Target: Sequence corresponding to the protein to be modeled.

Template: 3D structure/s to be used during protein structure prediction.

Model: Predicted 3D structure of the target sequence.

Nomenclature

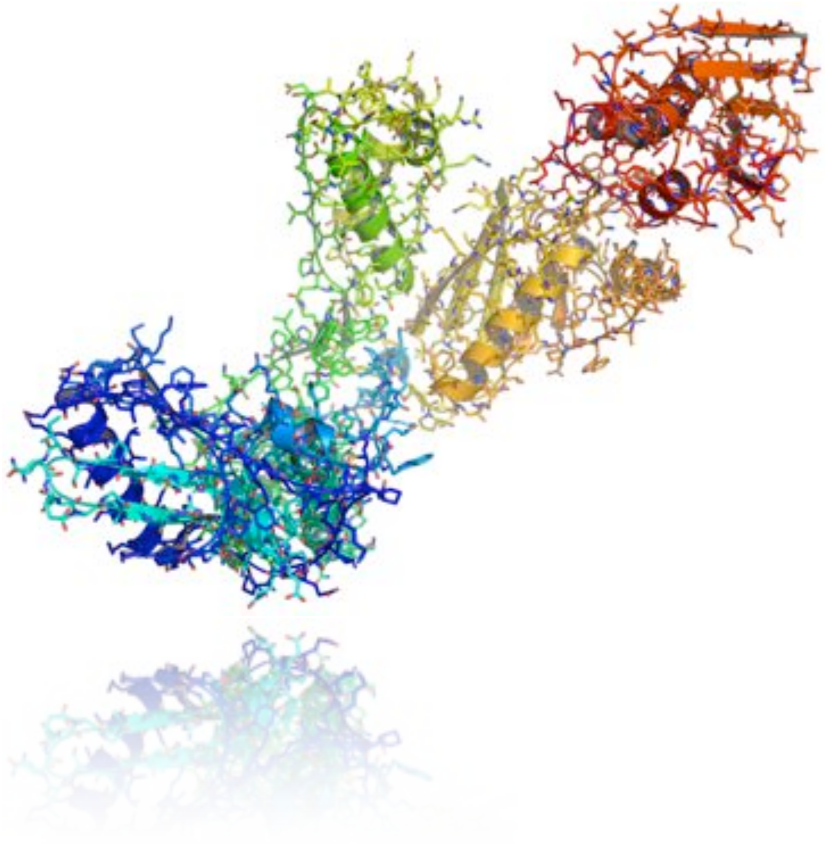
Fold: Three dimensional conformation of a protein sequence (usually at domain level).

Domain: Structurally globular part of a protein, which may independently fold.

Secondary Structure: Regular sub-domain structures composed by alpha-helices, beta-sheets and coils (or loops).

Backbone: Protein structure skeleton composed by the carbon, nitrogen and oxygen atoms.

Side-Chain: Specific atoms identifying each of the 20 residues types.



General References

Protein Structure Prediction:

Marti-Renom et al. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.
Baker & Sali. Science 294, 93-96, 2001.

Comparative Modeling:

Marti-Renom et al. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.
Madhusudhan et al. The Proteomics Protocols Handbook. Ed. Walker. Humana Press Inc., Totowa, NJ. 831-860, 2005.

MODELLER:

Sali & Blundell. J. Mol. Biol. 234, 779-815, 1993.

Structural Genomics:

Sali. Nat. Struct. Biol. 5, 1029, 1998.
Burley et al. Nat. Genet. 23, 151, 1999.
Sali & Kuriyan. TIBS 22, M20, 1999.
Sanchez et al. Nat. Str. Biol. 7, 986, 2000.
Baker & Sali. Science 294, 93-96, 2001.

protein prediction .vs. protein determination

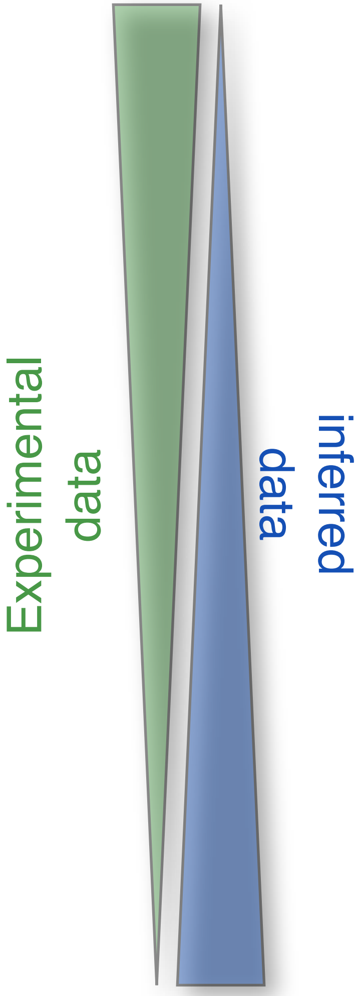
X-Ray

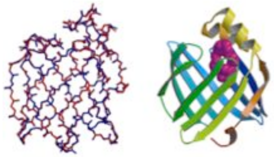
NMR

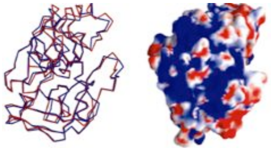
Comparative Modeling

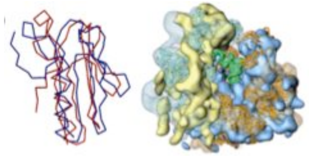
Threading

Ab-initio



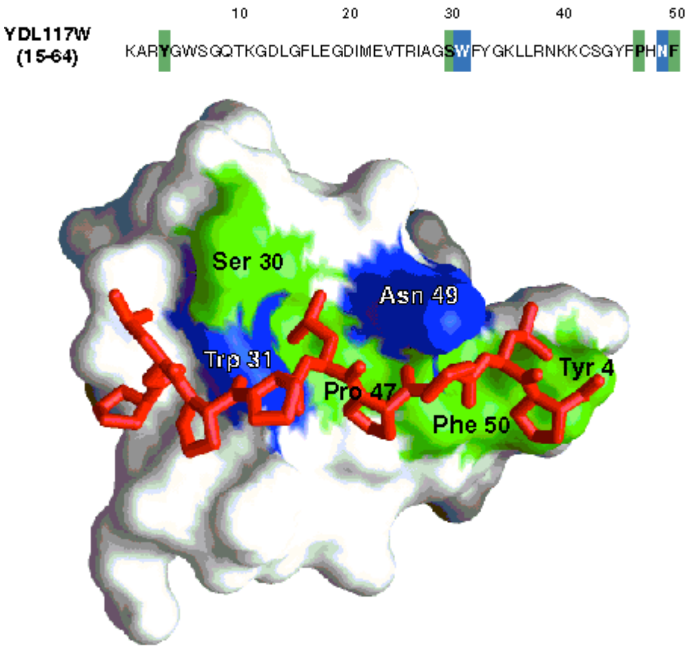






Why is it useful to know the **structure** of a protein, not only its sequence?

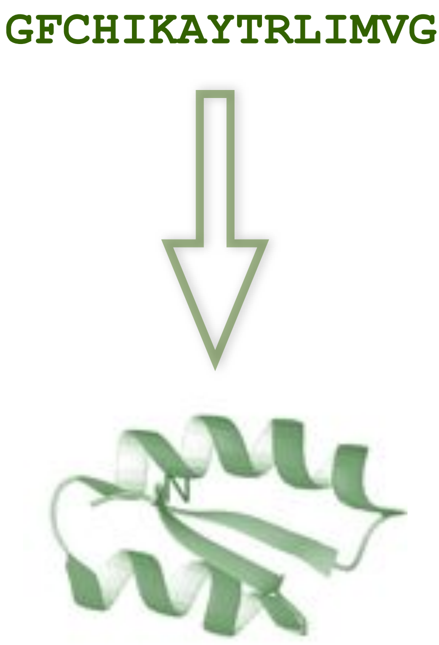
- The biochemical function (activity) of a protein is defined by its interactions with other molecules.
- The biological function is in large part a consequence of these interactions.
- The 3D structure is more informative than sequence because interactions are determined by residues that are close in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



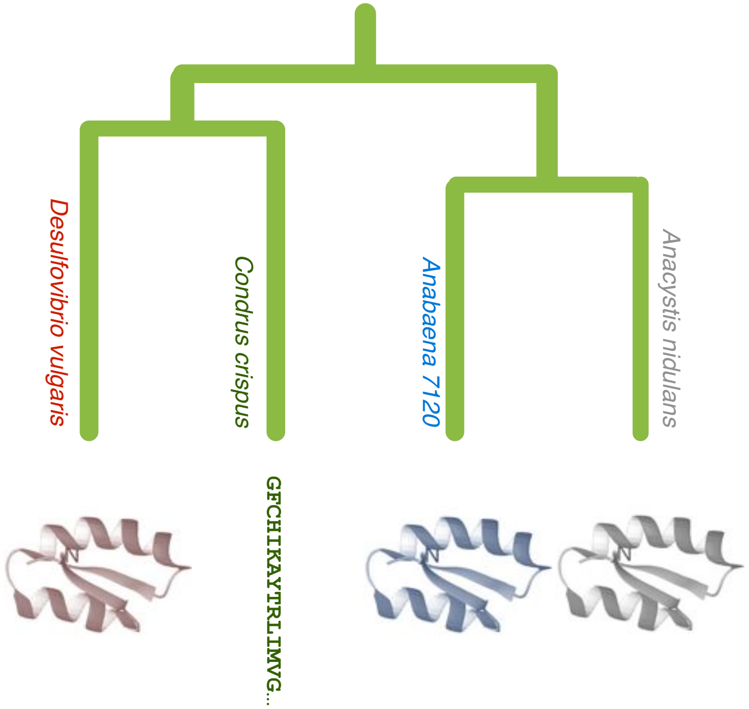
In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence**.

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

Principles of protein structure



GFCHIKAYTRLIMVG...



Folding (physics)

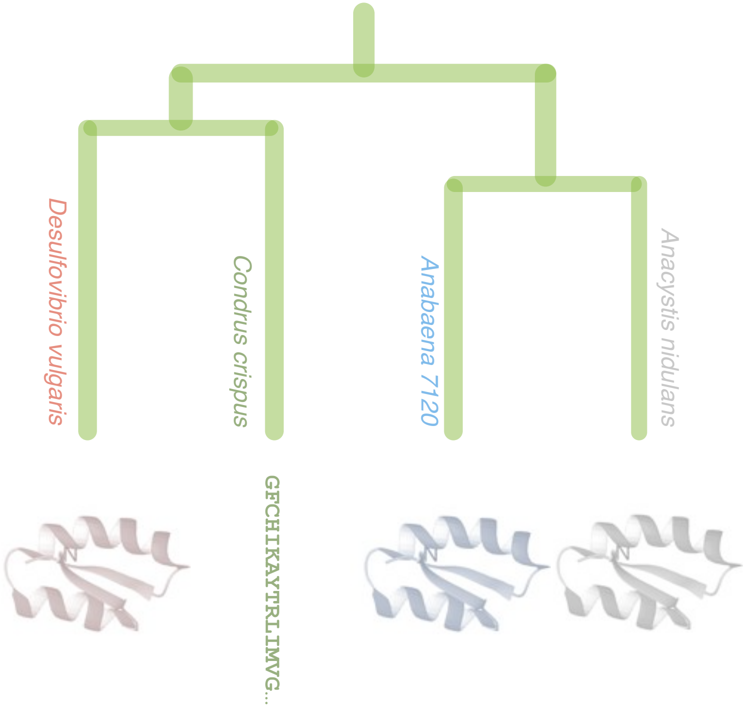
Ab initio prediction

Evolution (rules)

Threading

Comparative Modeling

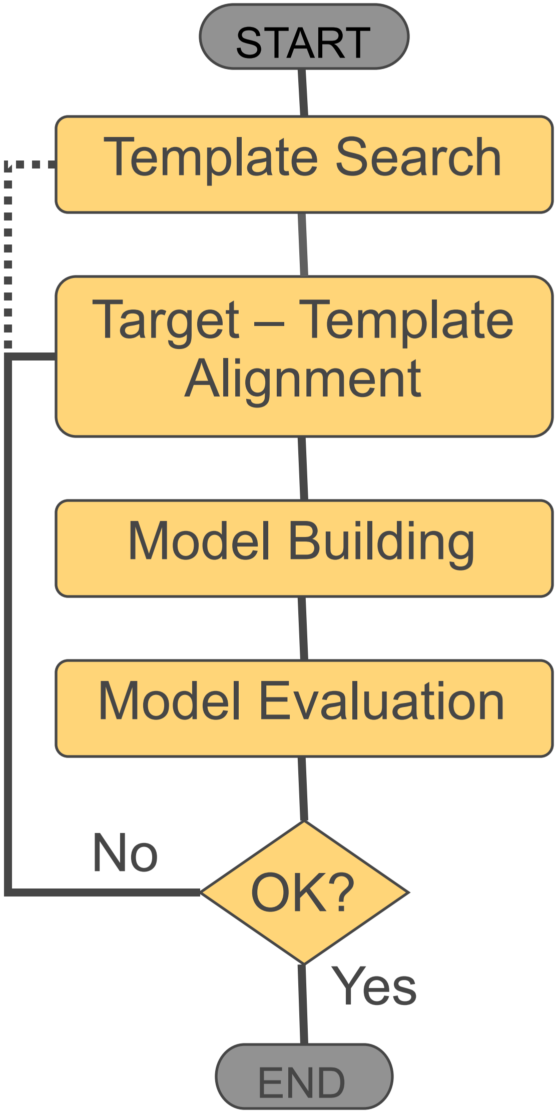
D. Baker & A. Sali. Science 294, 93, 2001.



MODELLER

1. N. Eswar, et al. *Comparative Protein Structure Modeling With MODELLER. Current Protocols in Bioinformatics, John Wiley & Sons, Inc., Supplement 15, 5.6.1-5.6.30, 2008.*
2. M.A. Marti-Renom, et al.. *Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.*
3. A. Sali & T.L. Blundell. *Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815, 1993.*
4. A. Fiser, R.K. Do, & A. Sali. *Modeling of loops in protein structures, Protein Science 9. 1753-1773, 2000.*

Steps in Comparative Protein Structure Modeling

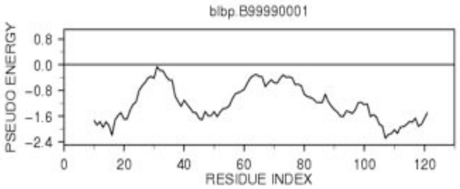


TARGET

ASILPKRLFGNCEQTSDEG
LKIERTPLVPHISAQNVCLKI
DDVPERLIPERASFQWMN
DK

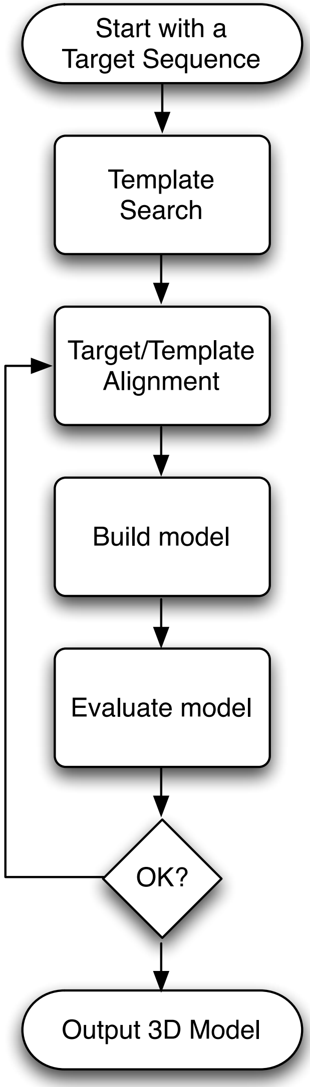
TEMPLATE

```
ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE
```



A. Šali, Curr. Opin. Biotech. 6, 437, 1995.
R. Sánchez & A. Šali, Curr. Opin. Str. Biol. 7, 206, 1997.
M. Marti et al. Ann. Rev. Biophys. Biomolec. Struct., 29, 291, 2000.

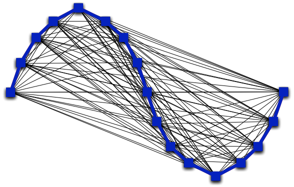
Comparative modeling by satisfaction of spatial restraints
MODELLER



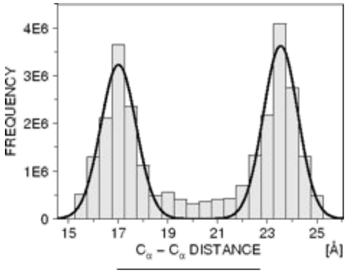
Given an alignment...

```
MSVIPKR--GNCEQTSE
ASILPKRLFGNCEQTSD
```

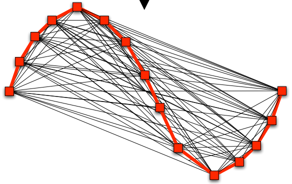
extract spatial features from the template(s) and statistics from known structures



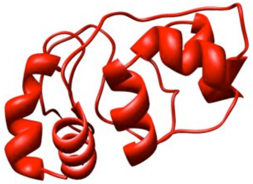
+



apply these features as restraints on your target sequence



optimize to find the best solution for the restraints to produce your 3D model



A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.

Template Selection

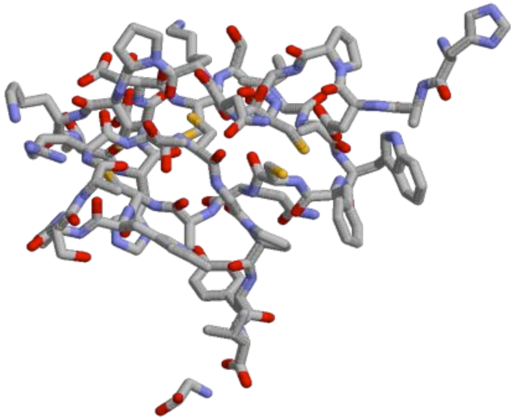
“Structural Space”

Structure-Structure alignments

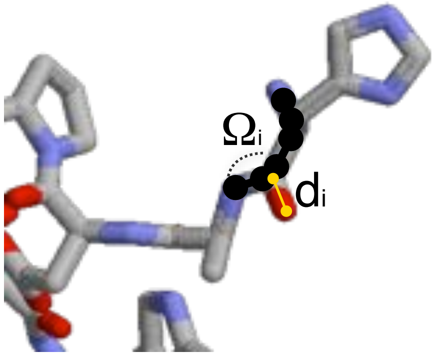
As any other bioinformatics problem…

- Representation
- Scoring
- Optimizer

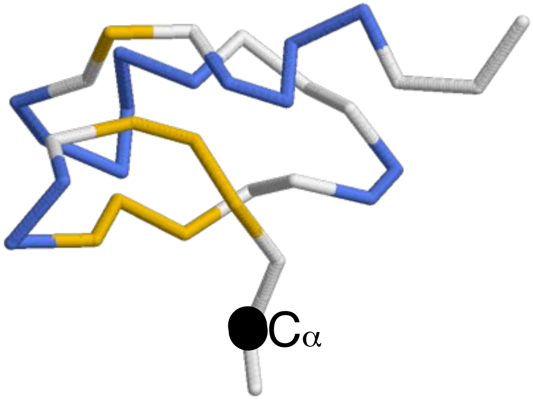
Structures



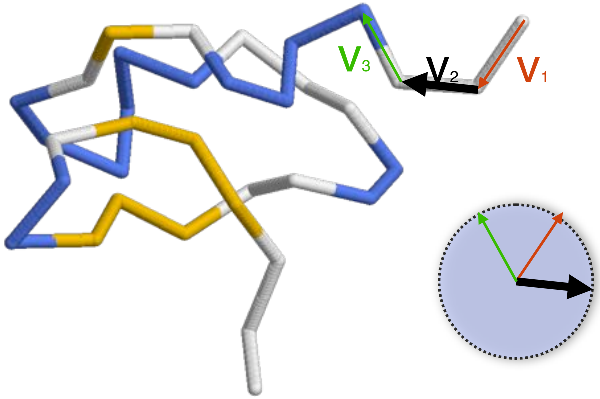
All atoms and coordinates



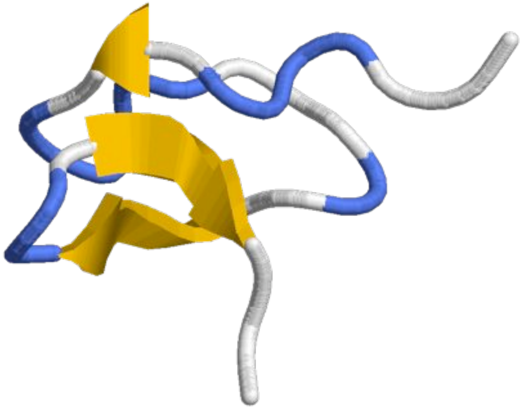
Dihedral space or distance space



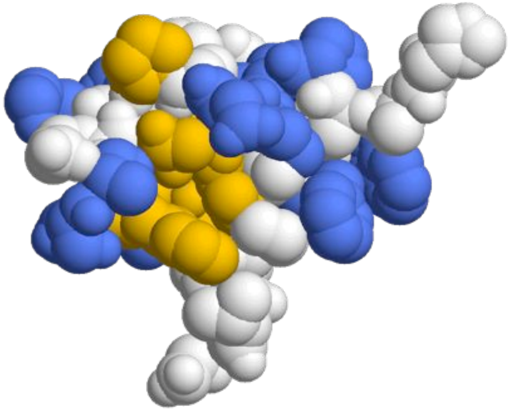
Reduced atom representation



Vector representation



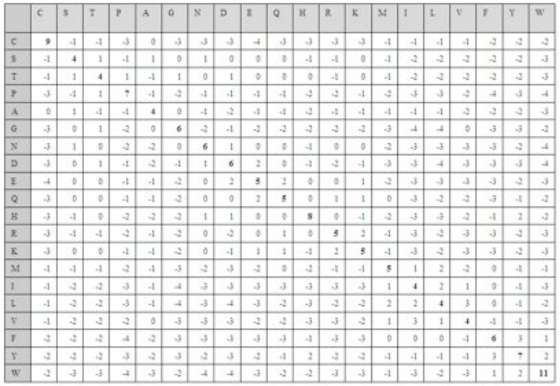
Secondary Structure



Accessible surface (and others)

Scoring

Raw scores

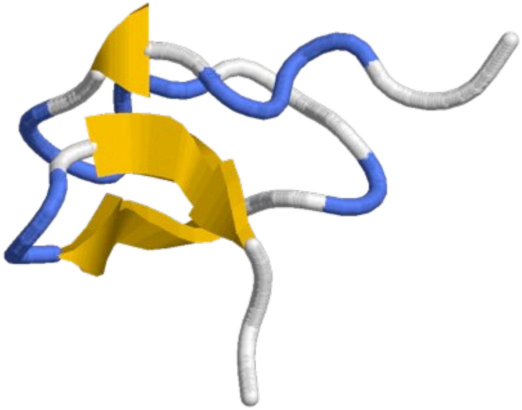


	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	4	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	6	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	8	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

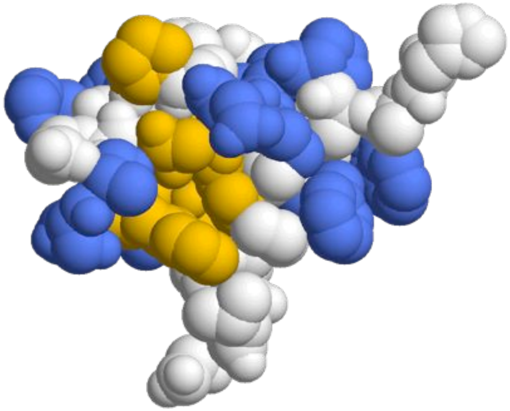
Aminoacid substitutions

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

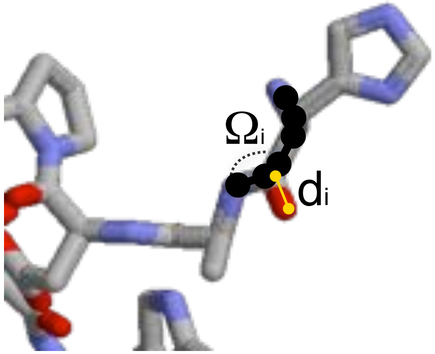
Root Mean Square Deviation



Secondary Structure (H,B,C)



Accessible surface (B,A [%])

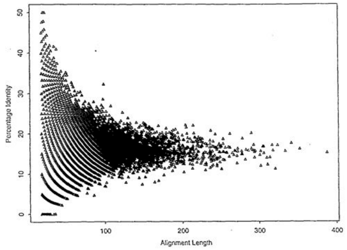


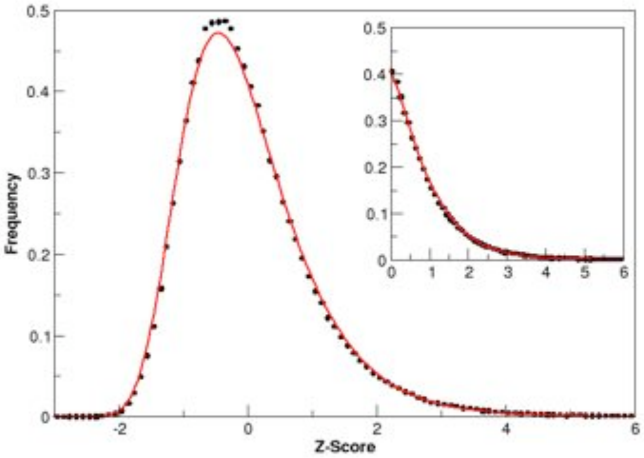
Angles or distances

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

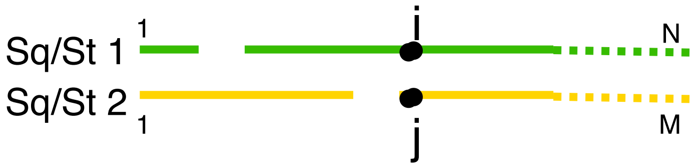
$$P(s) = e^{-\lambda\,(s-\mu)}$$

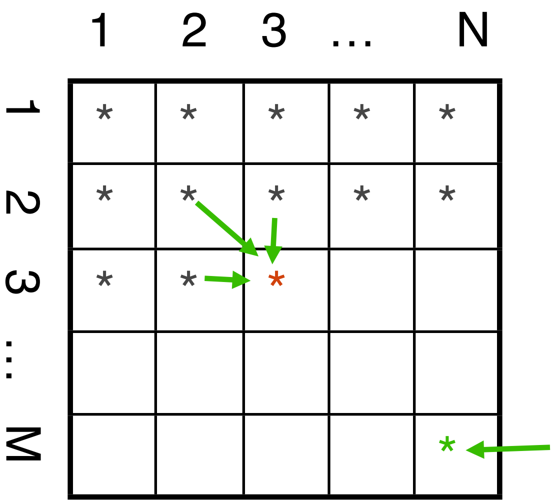
$$P(s \geq x) = 1 - \exp\left(e^{-\lambda\,(s-\mu)}\right)$$

Karlin and Altschul, 1990 PNAS 87, pp2264

Optimizer

Global dynamic programming alignment





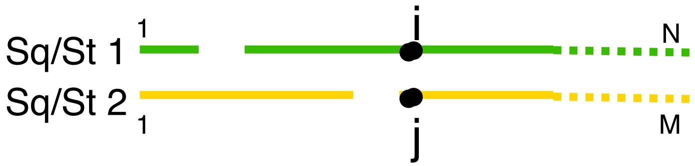
Best alignment score

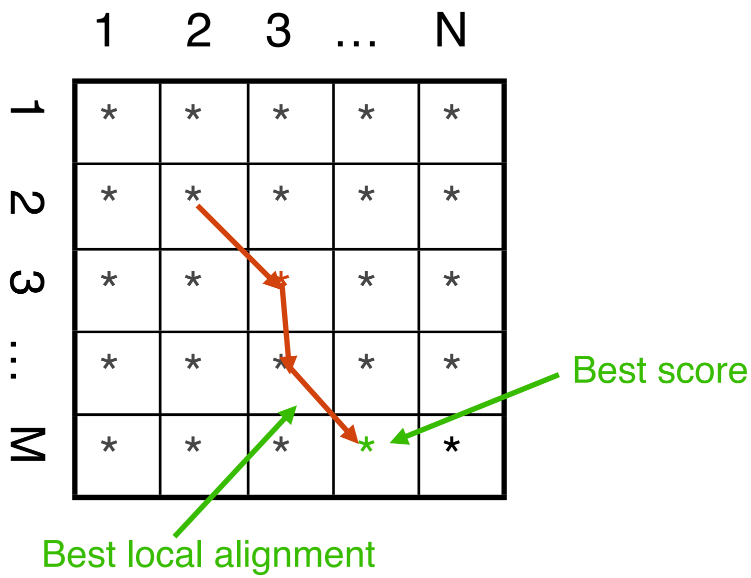
$$D_{i,j}=\min\begin{cases} D_{i,j-1}+Score_{(\Delta,rj)} \\ D_{i-1,j-1}+Score_{(ri,rj)} \\ D_{i-1,j}+Score_{(ri,\Delta)} \end{cases}$$

Backtracking to get the best alignment

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

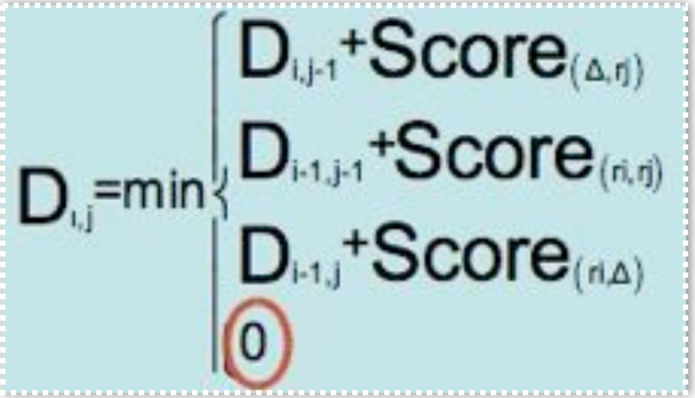
Local dynamic programming alignment





Best score

Best local alignment

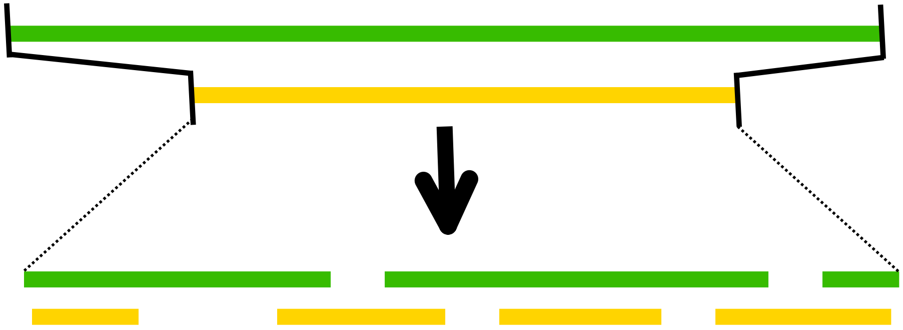


$$D_{i,j}=\min\begin{cases} D_{i,j-1}+\mathrm{Score}_{(\Delta,r_j)} \\ D_{i-1,j-1}+\mathrm{Score}_{(r_i,r_j)} \\ D_{i-1,j}+\mathrm{Score}_{(r_i,\Delta)} \\ \boxed{0} \end{cases}$$

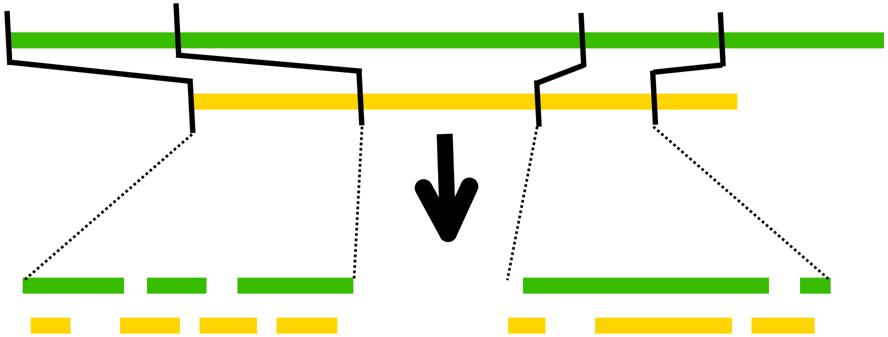
Backtracking to get the best alignment

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

Global .vs. local alignment



Global alignment



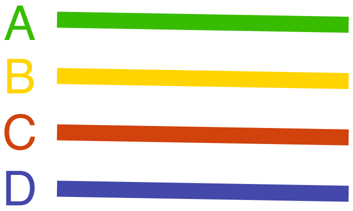
Local alignment

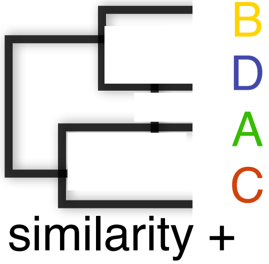
Optimizer

Multiple alignment

Pairwise alignments

Example – 4 sequences A, B, C, D.

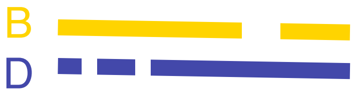




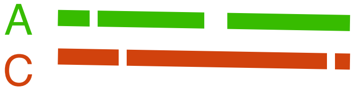
6 pairwise comparisons
then cluster analysis

Multiple alignments

Following the tree from step 1

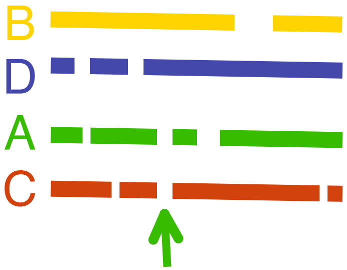


Align the most similar pair



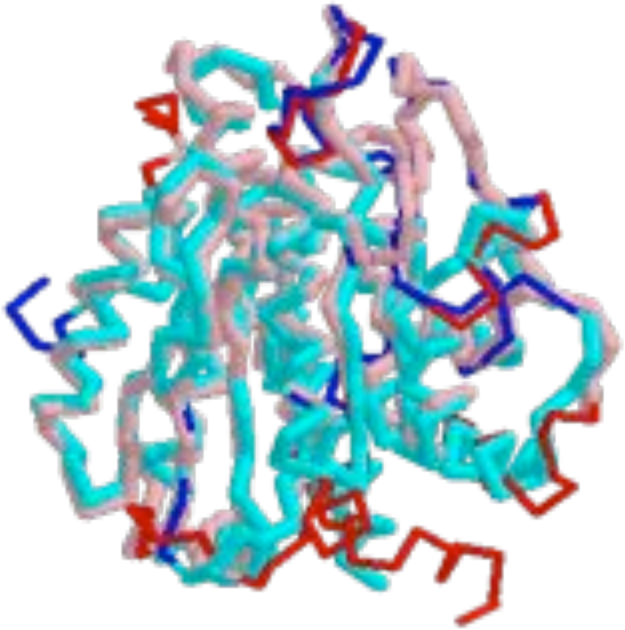
Align next most similar pair

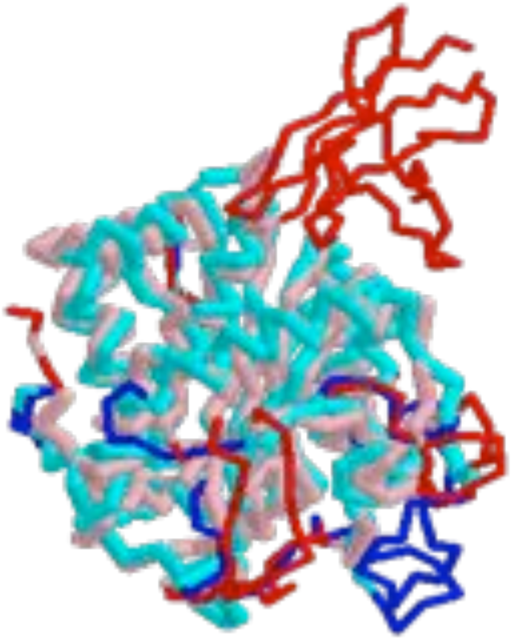
Align B-D with A-C



New gap in A-C to optimize
its alignment with B-D

Coverage .vs. Accuracy



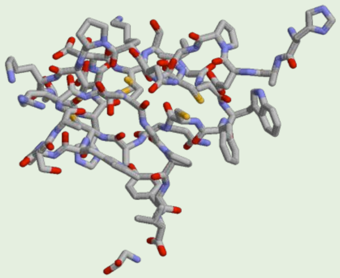


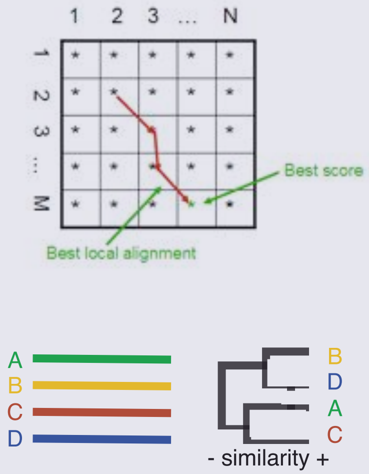
Same RMSD ~ 2.5Å

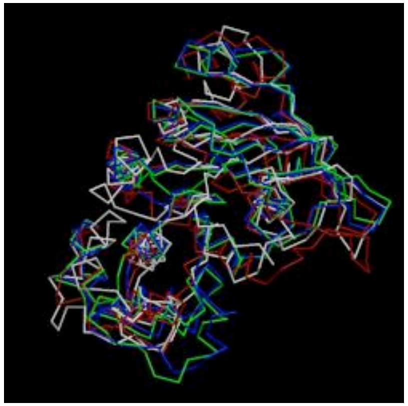
Coverage ~90% Cα

Coverage ~75% Cα

Structural alignment by properties conservation (SALIGN-MODELLER)



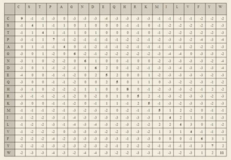


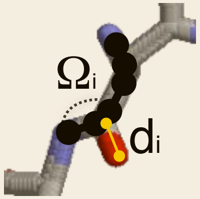


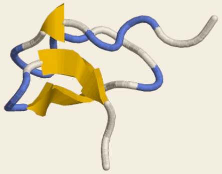
✓ Uses all available structural information

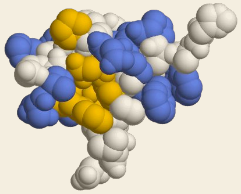
✓ Provides the optimal alignment

Computationally expensive









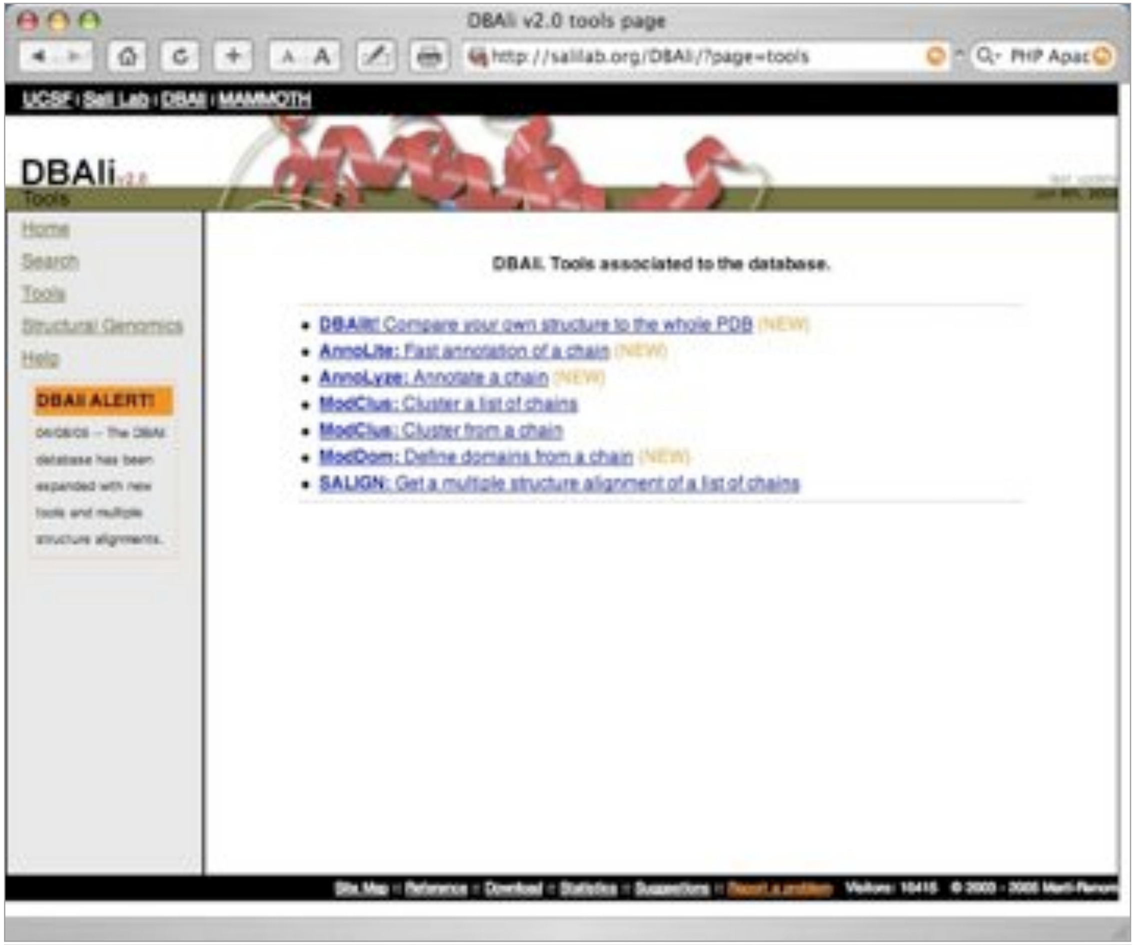
$$RMSD(x,y)=\sqrt{\left(\tfrac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i)-\mathbf{y}(i)\right\|^{2}\right)}$$

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

Madhusudhan et al. in preparation

Structural alignment by properties conservation (SALIGN-MODELLER)

http://salilab.org/DBAli


DBAli v2.0 tools page
http://salilab.org/DBAli/?page=tools
PHP Apac
UCSF | Sali Lab | DBAli | MAMMOTH
DBAli v2.0
Tools
Home
Search
Tools
Structural Genomics
Help
DBAli ALERT!
Site Map | Reference | Download | Statistics | Suggestions | Report a problem
Visitors: 10415
© 2003 - 2008 Marti-Renom

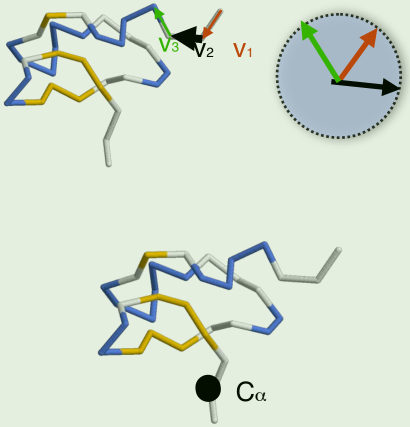
06/08/08 -- The DBAli database has been expanded with new tools and multiple structure alignments.

DBAli. Tools associated to the database.

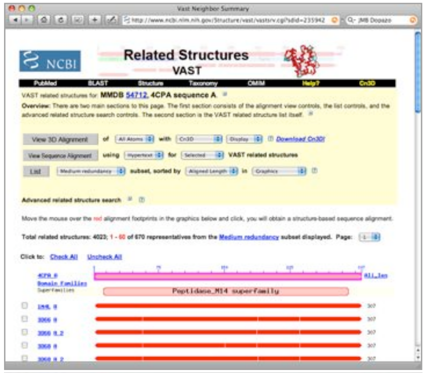
- DBAlit! Compare your own structure to the whole PDB (NEW)
- AnnoLite: Fast annotation of a chain (NEW)
- AnnoLyze: Annotate a chain (NEW)
- ModClus: Cluster a list of chains
- ModClus: Cluster from a chain
- ModDom: Define domains from a chain (NEW)
- SALIGN: Get a multiple structure alignment of a list of chains

Madhusudhan, in preparation

Vector Alignment Search Tool (VAST)



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


Vast Neighbor Summary
http://www.ncbi.nlm.nih.gov/Structure/vast/vastsrv.cgi?sdid=235942
JMB Dopazo
NCBI
Related Structures
VAST
PubMed BLAST Structure Taxonomy OMIM Help? Cn3D

VAST related structures for: **MMDB 54712, 4CPA sequence A**

Overview: There are two main sections to this page. The first section consists of the alignment view controls, the list controls, and the advanced related structure search controls. The second section is the VAST related structure list itself.

View 3D Alignment of All Atoms with Cn3D Display Download Cn3D!

View Sequence Alignment using Hypertext for Selected VAST related structures

List Medium redundancy subset, sorted by Aligned Length in Graphics

Advanced related structure search

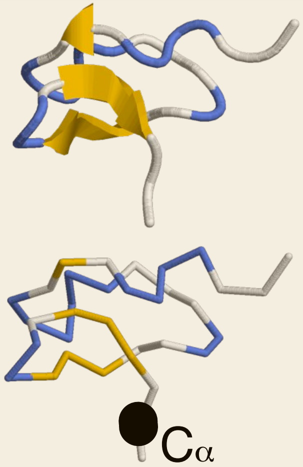
Move the mouse over the red alignment footprints in the graphics below and click, you will obtain a structure-based sequence alignment.

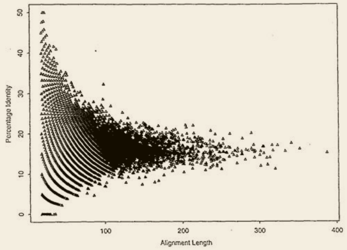
Total related structures: 4023; 1 - 60 of 670 representatives from the Medium redundancy subset displayed. Page: 1

Click to: Check All Uncheck All

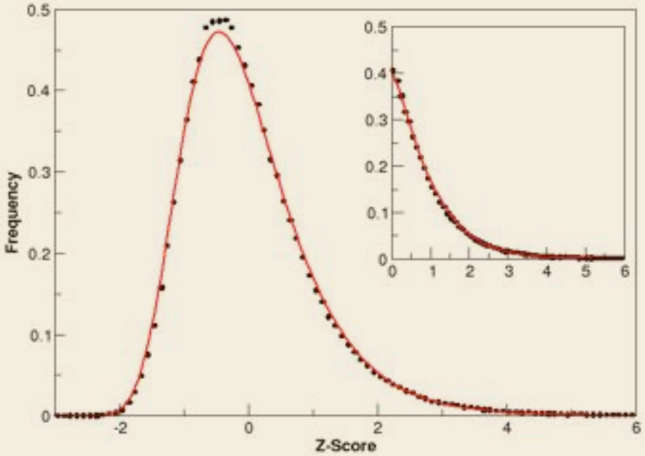
✓ Good scoring system with significance

Reduces the protein representation





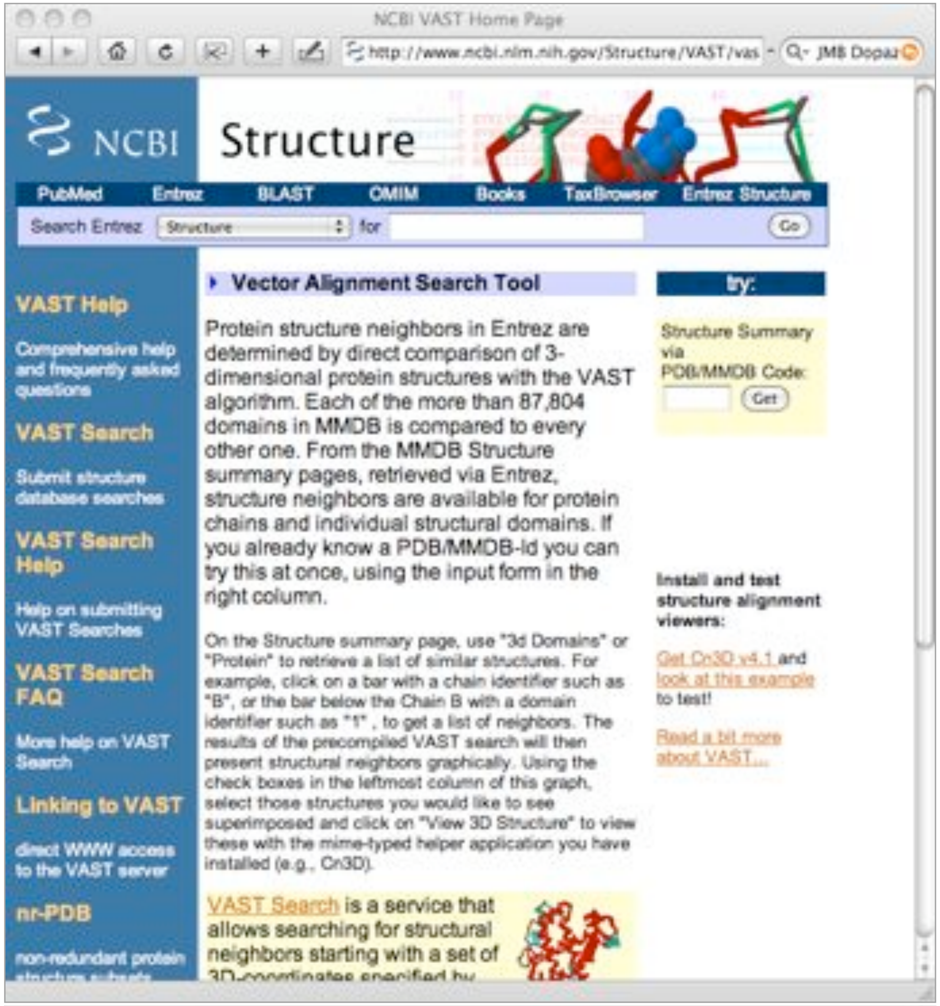
$$RMSD(x,y) = \sqrt{\left(\tfrac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i) - \mathbf{y}(i)\right\|^2\right)}$$



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



Vector Alignment Search Tool

Protein structure neighbors in Entrez are determined by direct comparison of 3-dimensional protein structures with the VAST algorithm. Each of the more than 87,804 domains in MMDB is compared to every other one. From the MMDB Structure summary pages, retrieved via Entrez, structure neighbors are available for protein chains and individual structural domains. If you already know a PDB/MMDB-Id you can try this at once, using the input form in the right column.

On the Structure summary page, use "3d Domains" or "Protein" to retrieve a list of similar structures. For example, click on a bar with a chain identifier such as "B", or the bar below the Chain B with a domain identifier such as "1" , to get a list of neighbors. The results of the precompiled VAST search will then present structural neighbors graphically. Using the check boxes in the leftmost column of this graph, select those structures you would like to see superimposed and click on "View 3D Structure" to view these with the mime-typed helper application you have installed (e.g., Cn3D).

VAST Search is a service that allows searching for structural neighbors starting with a set of 3D-coordinates specified by

VAST Help

Comprehensive help and frequently asked questions

VAST Search

Submit structure database searches

VAST Search Help

Help on submitting VAST Searches

VAST Search FAQ

More help on VAST Search

Linking to VAST

direct WWW access to the VAST server

nr-PDB

non-redundant protein structure subsets

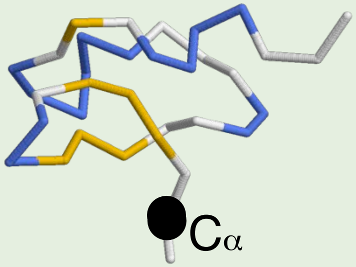
Structure Summary via PDB/MMDB Code:

Install and test structure alignment viewers:

Get Cn3D v4.1 and look at this example to test!

Read a bit more about VAST...

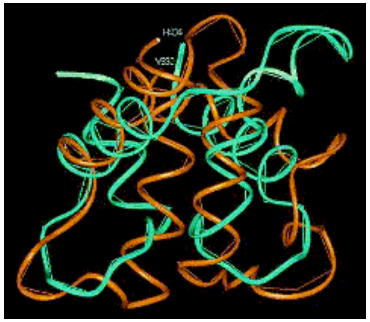
Incremental combinatorial extension (CE)



Exhaustive combination of fragments

Longest combination of AFPs

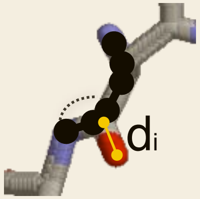
Heuristic similar to PSI-BLAST



✓ FAST!

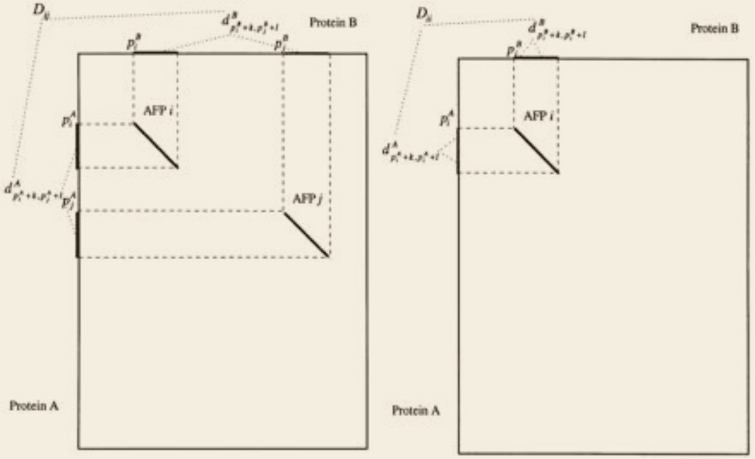
✓ Good quality of local alignments

Complicated scoring and heuristics



8 residues peptides

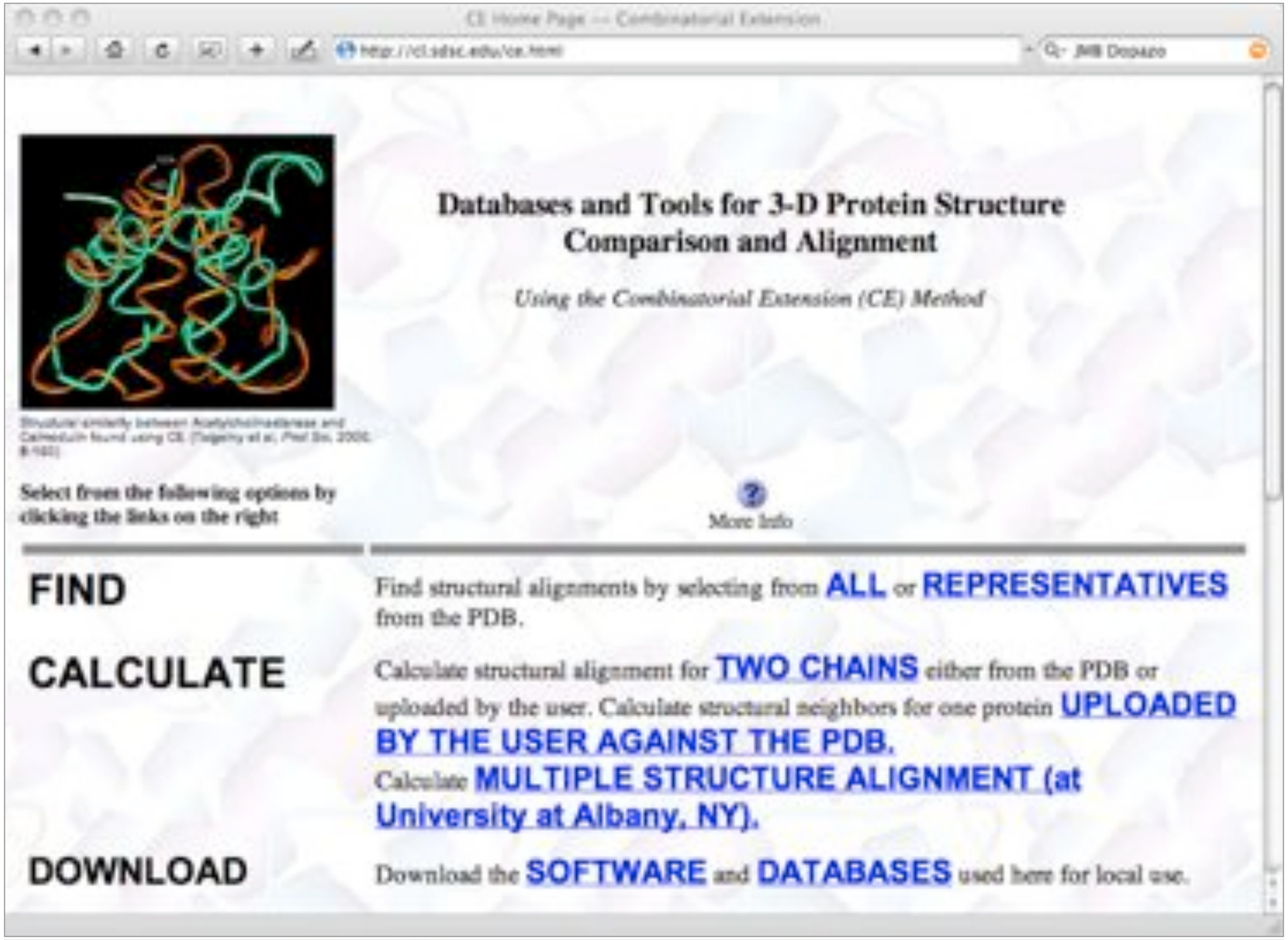
$$RMSD(x,y) = \sqrt{\left(\tfrac{1}{N}\right)\sum_{i=1}^{N}\left(\|\mathbf{x}(i) - \mathbf{y}(i)\|^2\right)}$$



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

Incremental combinatorial extension (CE)

http://cl.sdsc.edu/ce.html


CE Home Page -- Combinatorial Extension
http://cl.sdsc.edu/ce.html
JMB Dopazo
More Info

Structural similarity between Acetylcholinesterase and Calmodulin found using CE. (Tsigelny et al, Prot Sci. 2000, 9:180)

Databases and Tools for 3-D Protein Structure Comparison and Alignment

Using the Combinatorial Extension (CE) Method

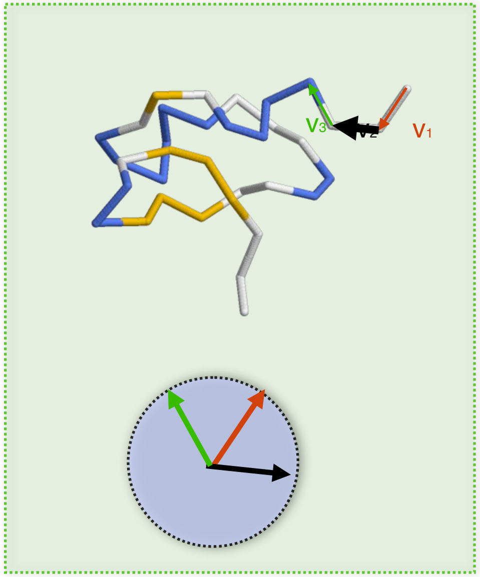
Select from the following options by clicking the links on the right

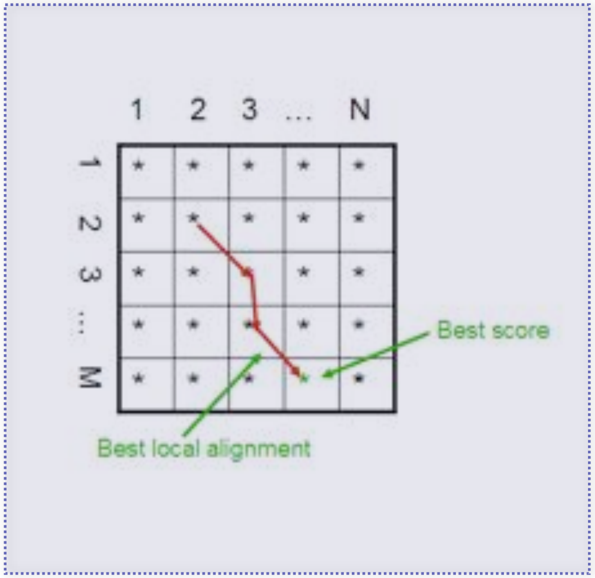
FIND Find structural alignments by selecting from **ALL** or **REPRESENTATIVES** from the PDB.

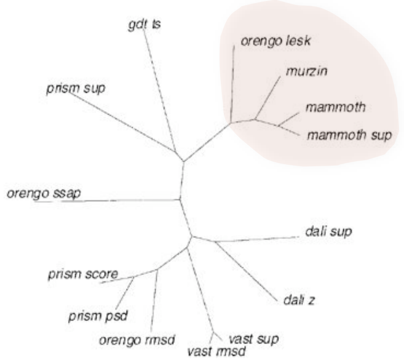
CALCULATE Calculate structural alignment for **TWO CHAINS** either from the PDB or uploaded by the user. Calculate structural neighbors for one protein **UPLOADED BY THE USER AGAINST THE PDB.** Calculate **MULTIPLE STRUCTURE ALIGNMENT (at University at Albany, NY).**

DOWNLOAD Download the **SOFTWARE** and **DATABASES** used here for local use.

Matching molecular models obtained from theory (MAMMOTH)





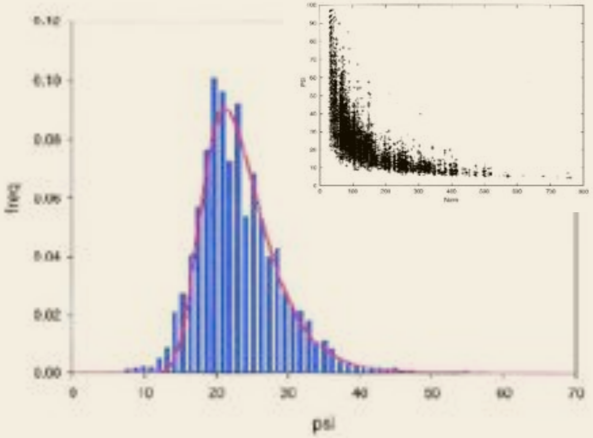


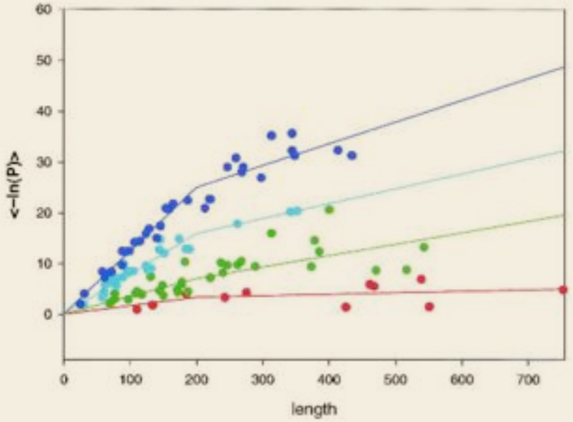
- ✓ VERY FAST!
- ✓ Good scoring system with significance

Reduces the protein representation

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

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

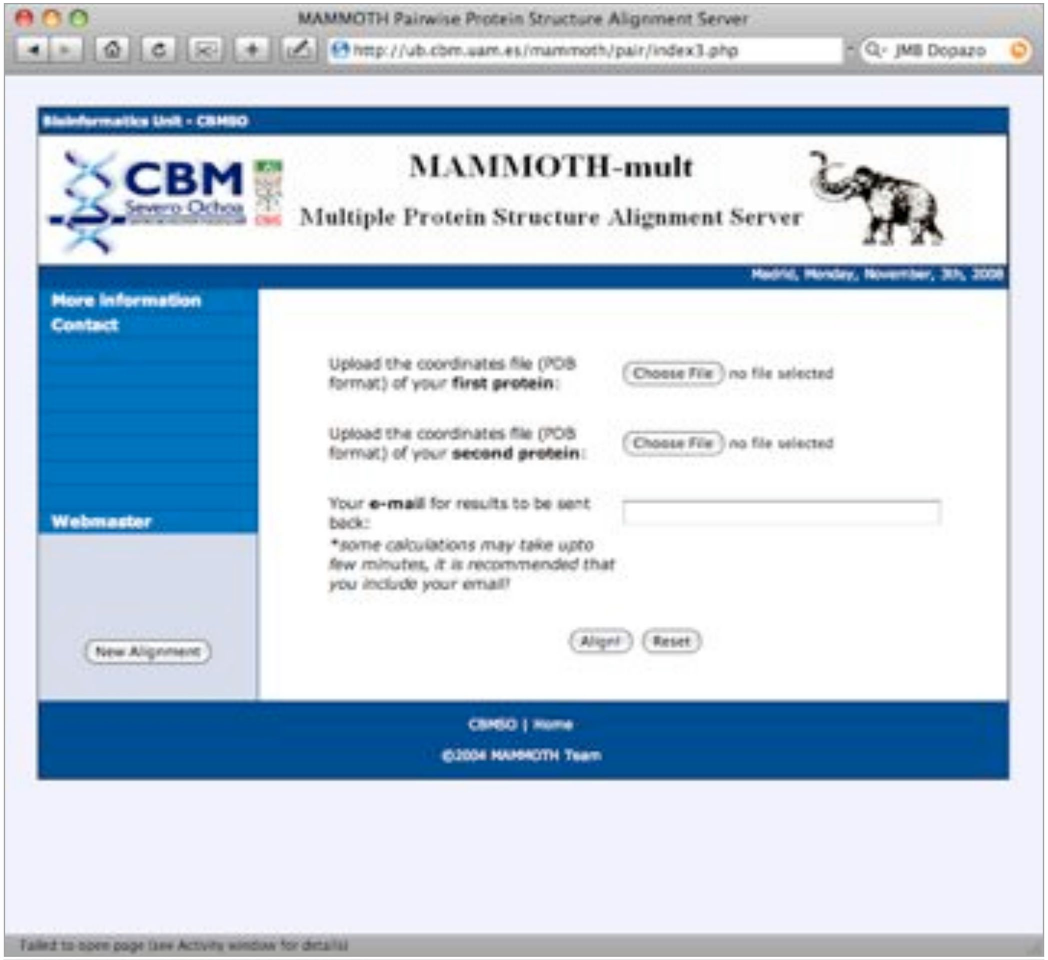




Ortiz AR, (2002) Protein Sci. 11 pp2606

Matching molecular models obtained from theory (MAMMOTH)

http://ub.cbm.uam.es/mammoth/pair/index3.php



MAMMOTH Pairwise Protein Structure Alignment Server

http://ub.cbm.uam.es/mammoth/pair/index3.php

JMB Dopazo

Bioinformatics Unit - CBMSO

CBM Severo Ochoa

MAMMOTH-mult

Multiple Protein Structure Alignment Server

Madrid, Monday, November, 3th, 2008

More information

Contact

Webmaster

Upload the coordinates file (PDB format) of your **first protein**: Choose File no file selected

Upload the coordinates file (PDB format) of your **second protein**: Choose File no file selected

Your **e-mail** for results to be sent back:

**some calculations may take upto few minutes, it is recommended that you include your email!*

Align! Reset

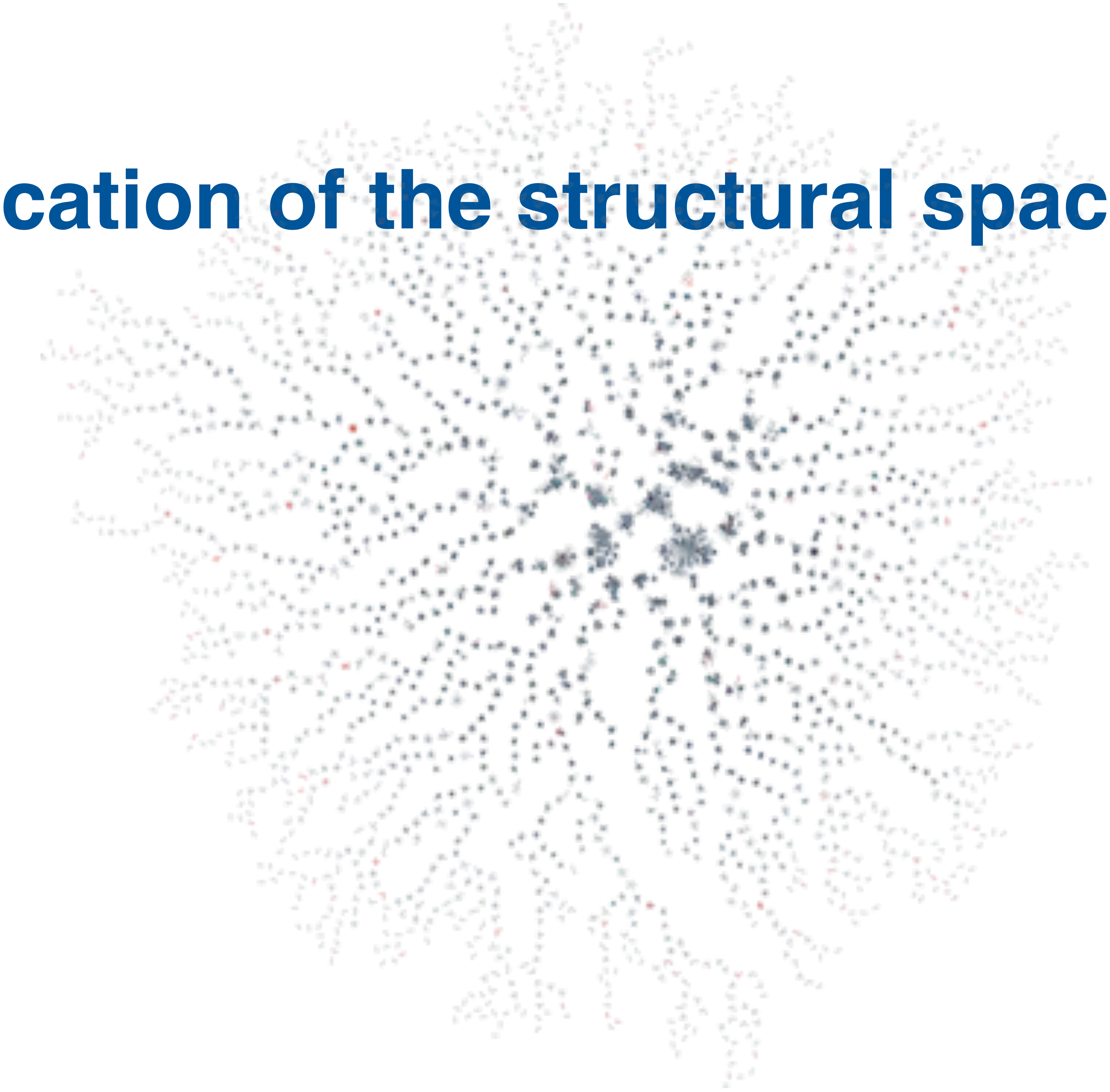
New Alignment

CBMSO | Home

©2004 MAMMOTH Team

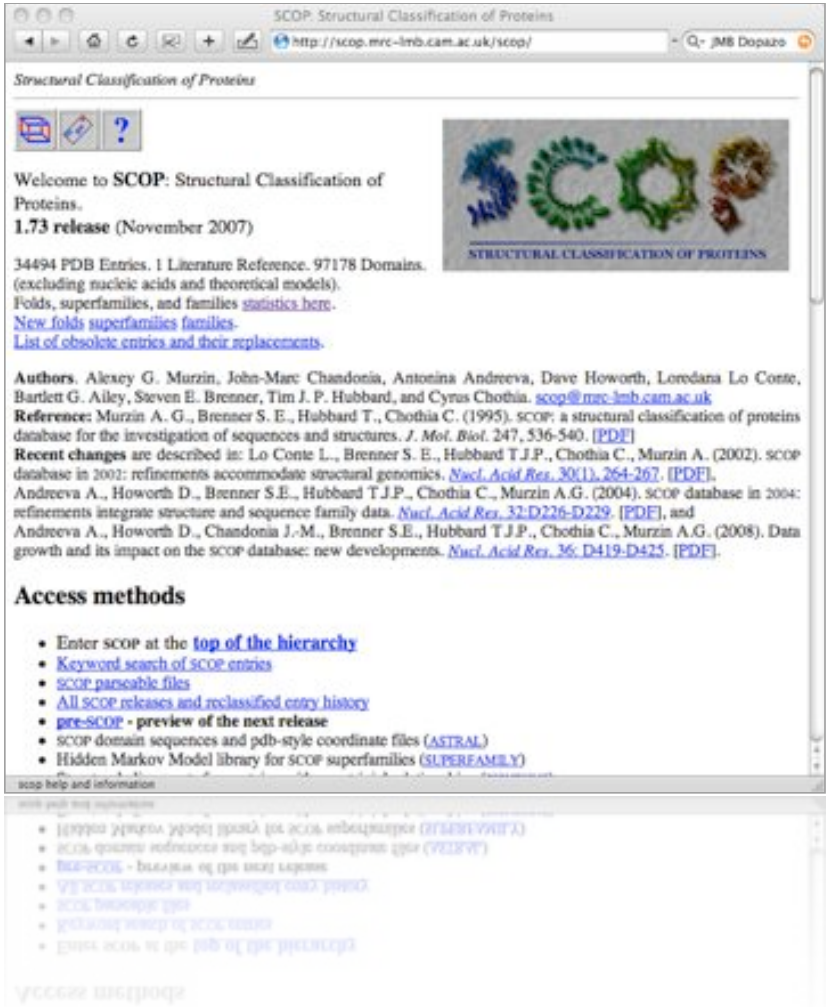
Failed to open page (see Activity window for details)

Classification of the structural space



SCOP$_{1.73}$ database

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


SCOP: Structural Classification of Proteins
http://scop.mrc-lmb.cam.ac.uk/scop/
JMB Dopazo
STRUCTURAL CLASSIFICATION OF PROTEINS
scop help and information

Structural Classification of Proteins

Welcome to **SCOP**: Structural Classification of Proteins.
1.73 release (November 2007)

34494 PDB Entries. 1 Literature Reference. 97178 Domains. (excluding nucleic acids and theoretical models).
Folds, superfamilies, and families statistics here.
New folds superfamilies families.
List of obsolete entries and their replacements.

Authors. Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. scop@mrc-lmb.cam.ac.uk
Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [PDF]
Recent changes are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res. 30(1), 264-267*. [PDF].
Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res. 32:D226-D229*. [PDF], and
Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucl. Acid Res. 36: D419-D425*. [PDF].

Access methods

- Enter SCOP at the **top of the hierarchy**
- Keyword search of SCOP entries
- SCOP parseable files
- All SCOP releases and reclassified entry history
- pre-SCOP - **preview of the next release**
- SCOP domain sequences and pdb-style coordinate files (ASTRAL)
- Hidden Markov Model library for SCOP superfamilies (SUPERFAMILY)

- ✓ 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

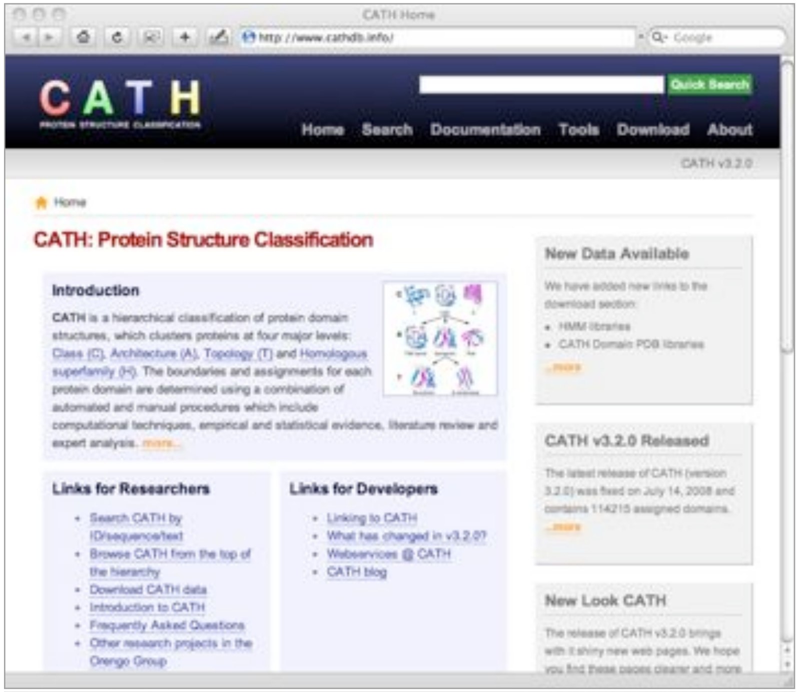
Manually classification
Not 100% up-to-date
Domain boundaries definition

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	259	459	772
All beta proteins	165	331	679
Alpha and beta proteins (a/b)	141	232	736
Alpha and beta proteins (a+b)	334	488	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins	85	122	202
Total	1086	1777	3464

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

CATH$_{3.2}$ database

http://www.cathdb.info



CATH: Protein Structure Classification

Introduction

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). The boundaries and assignments for each protein domain are determined using a combination of automated and manual procedures which include computational techniques, empirical and statistical evidence, literature review and expert analysis. more...

Links for Researchers

- Search CATH by ID/sequence/text
- Browse CATH from the top of the hierarchy
- Download CATH data
- Introduction to CATH
- Frequently Asked Questions
- Other research projects in the Orengo Group

Links for Developers

- Linking to CATH
- What has changed in v3.2.0?
- Webservices @ CATH
- CATH blog

New Data Available

We have added new links to the download section:

- HMM libraries
- CATH Domain PDB libraries

...more

CATH v3.2.0 Released

The latest release of CATH (version 3.2.0) was fixed on July 14, 2008 and contains 114215 assigned domains.

...more

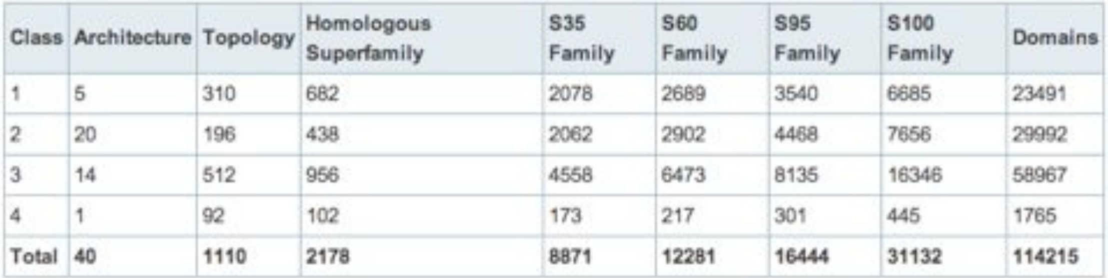
New Look CATH

The release of CATH v3.2.0 brings with it shiny new web pages. We hope you find these pages clearer and more

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

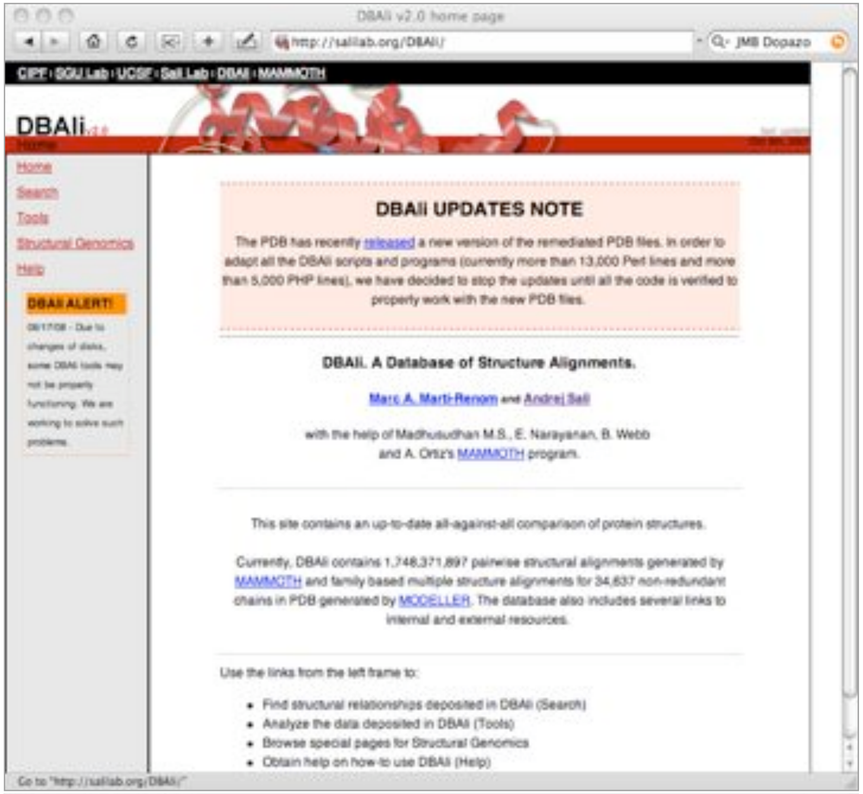


Class	Architecture	Topology	Homologous Superfamily	S35 Family	S60 Family	S95 Family	S100 Family	Domains
1	5	310	682	2078	2689	3540	6685	23491
2	20	196	438	2062	2902	4468	7656	29992
3	14	512	956	4558	6473	8135	16346	58967
4	1	92	102	173	217	301	445	1765
Total	40	1110	2178	8871	12281	16444	31132	114215

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

DBAli$_{v2.0}$ database

http://salilab.org/DBAli/



DBAli ALERT!

08/17/08 - Due to changes of disks, some DBAli tools may not be properly functioning. We are working to solve such problems.

DBAli UPDATES NOTE

The PDB has recently released a new version of the remediated PDB files. In order to adapt all the DBAli scripts and programs (currently more than 13,000 Perl lines and more than 5,000 PHP lines), we have decided to stop the updates until all the code is verified to properly work with the new PDB files.

DBAli. A Database of Structure Alignments.

Marc A. Marti-Renom and Andrej Sali

with the help of Madhusudhan M.S., E. Narayanan, B. Webb
and A. Ortiz's MAMMOTH program.

This site contains an up-to-date all-against-all comparison of protein structures.

Currently, DBAli contains 1,748,371,897 pairwise structural alignments generated by MAMMOTH and family based multiple structure alignments for 34,637 non-redundant chains in PDB generated by MODELLER. The database also includes several links to internal and external resources.

Use the links from the left frame to:

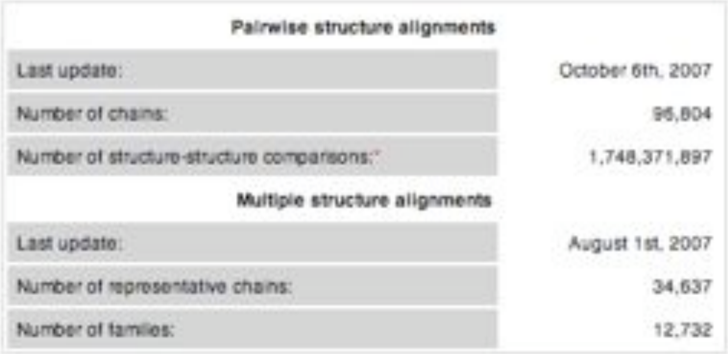
- Find structural relationships deposited in DBAli (Search)
- Analyze the data deposited in DBAli (Tools)
- Browse special pages for Structural Genomics
- Obtain help on how-to use DBAli (Help)

Go to "http://salilab.org/DBAli/"

Uses MAMMOTH for superimposition

- ✓ **Fully-automatic**
- ✓ **Data is kept up-to-date with PDB releases**
- ✓ **Tools for "on the fly" classification of families**
- ✓ **Up-to-date multiple structure alignments**
- ✓ **Easy to navigate**
- ✓ **Provides some tools for structure comparison**

Does not provide a stable classification



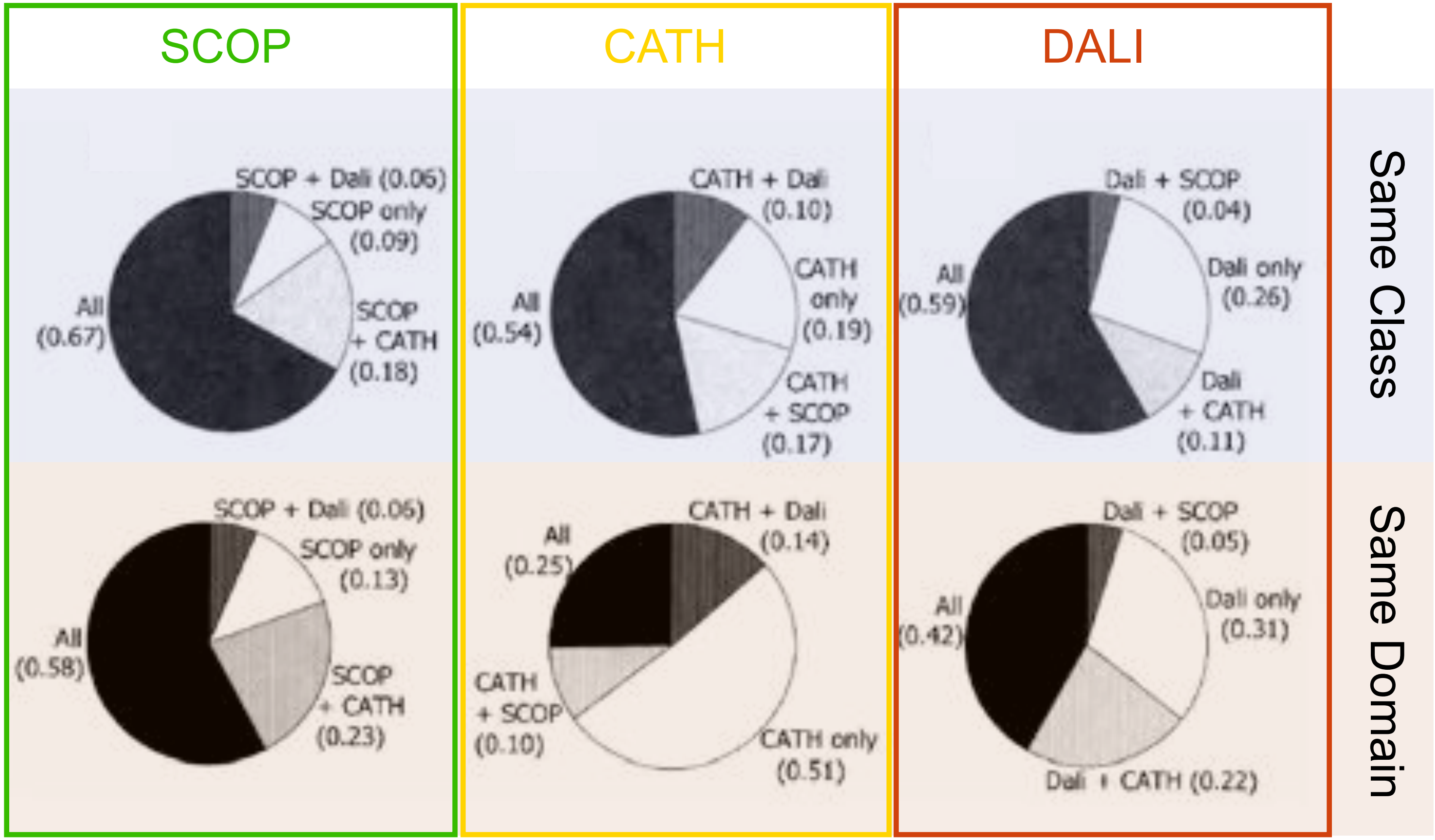
Pairwise structure alignments	
Last update:	October 6th, 2007
Number of chains:	96,804
Number of structure-structure comparisons:*	1,748,371,897
Multiple structure alignments	
Last update:	August 1st, 2007
Number of representative chains:	34,637
Number of families:	12,732

*Marti-Renom et al. 2001. Bioinformatics. **17**, 746*
Marti-Renom et al. 2007. BMC BMC Bioinformatics (2007) 8 (Suppl 4) S4
Marti-Renom et al. 2007. Nucleic Acid Research (2007) 35 W393-W397

Classification of the structural space

Not an easy task!

Domain definition AND domain classification

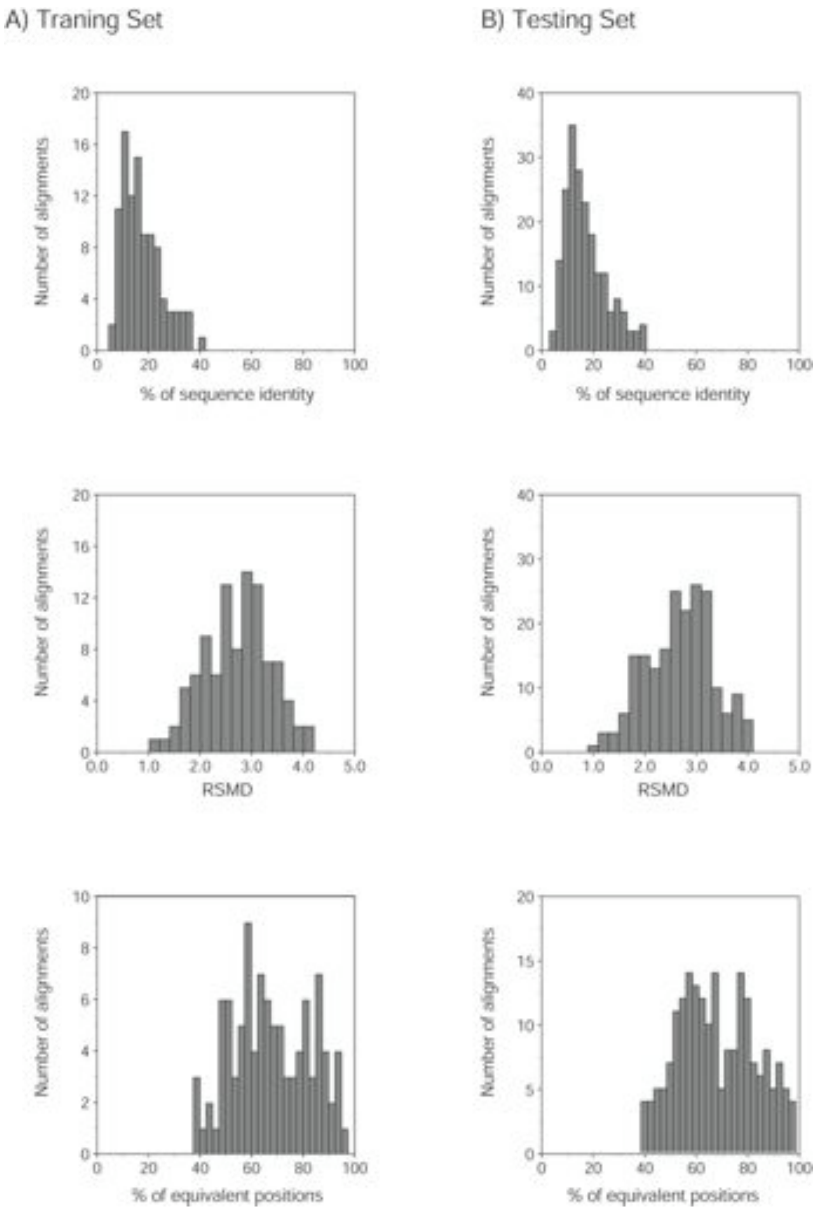


Day, et al. (2003) Protein Sciences, ***12*** *pp2150*

template search and template-target alignment (pp_scan)

Marti-Renom, et al. (2004) Prot. Sci. 13 pp1071
Narayanan, et al. in prepration

PP_SCAN or profile-profile alignments



Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local heuristic method

Seq.-Str.

SEA: Local structure prediction method

Prof.-Seq.

SAM: HMM method

PSI-BLAST: Local search method that uses multiple sequence information for one of the sequences.

LOBSTER: HHM + Phylogeny Method

Prof.-Prof.

CLUSTALW: DP multiple sequence method.

COMPASS: DP profile-profile method

PP_SCAN: DP pairwise method that uses multiple sequence information for both sequences.

PP_SCAN protocols

Profile generation

- PSI-Blast (PBP)
- Henikoff & Henikoff (HH)
- Henikoff & Henikoff + Similarity (HS)
- Henikoff & Henikoff substitution matrix (MAT)

Profile comparison

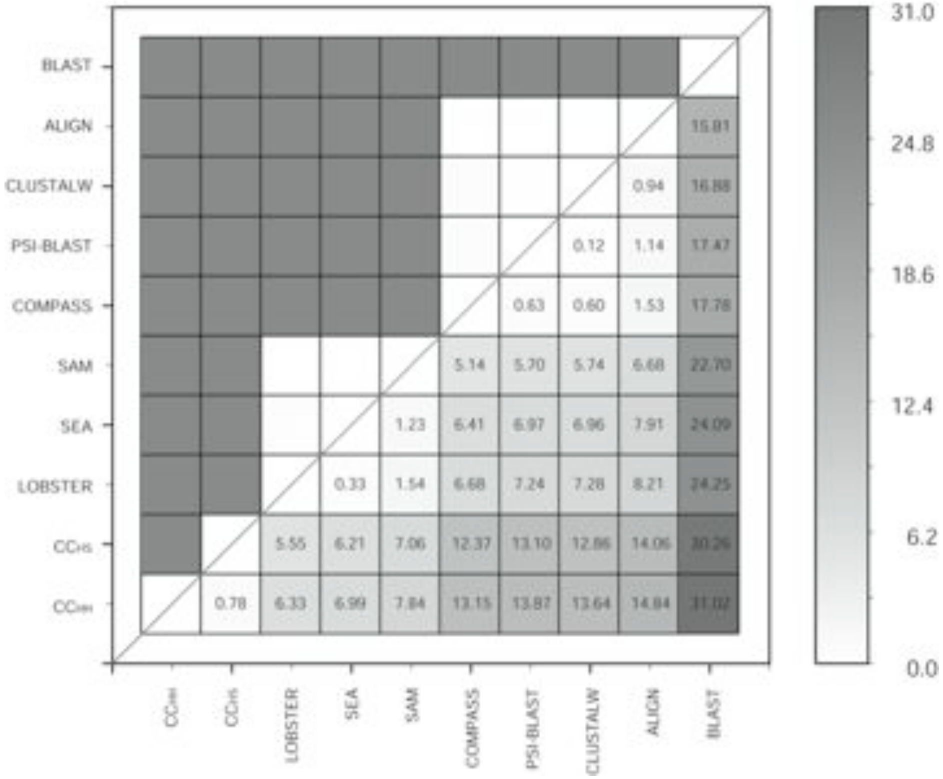
- Correlation coefficient (CC)
- Euclidean distance (ED)
- Dot product (DP)
- Jensen-Shannon distance (JS)
- Average value (Ave)

PP_SCAN protocols accuracy

SALIGN protocol	CE overlap [%]	Shift score
CCPBP	55 ± 23	0.61 ± 0.24
CCHH	**56 ± 23**	**0.61 ± 0.24**
CCHS	**56 ± 24**	**0.62 ± 0.23**
CCMAT	51 ± 25	0.55 ± 0.27
EDPBP	54 ± 24	0.60 ± 0.25
EDHH	54 ± 24	0.59 ± 0.26
EDHS	55 ± 24	0.59 ± 0.26
DPPBP	55 ± 23	0.61 ± 0.24
DPHH	56 ± 23	0.60 ± 0.25
DPHS	55 ± 24	0.61 ± 0.24
JSHH	53 ± 24	0.60 ± 0.24
JSHS	54 ± 24	0.60 ± 0.24
AveMAT	49 ± 26	0.52 ± 0.29
TOP	62 ± 20	0.67 ± 0.20

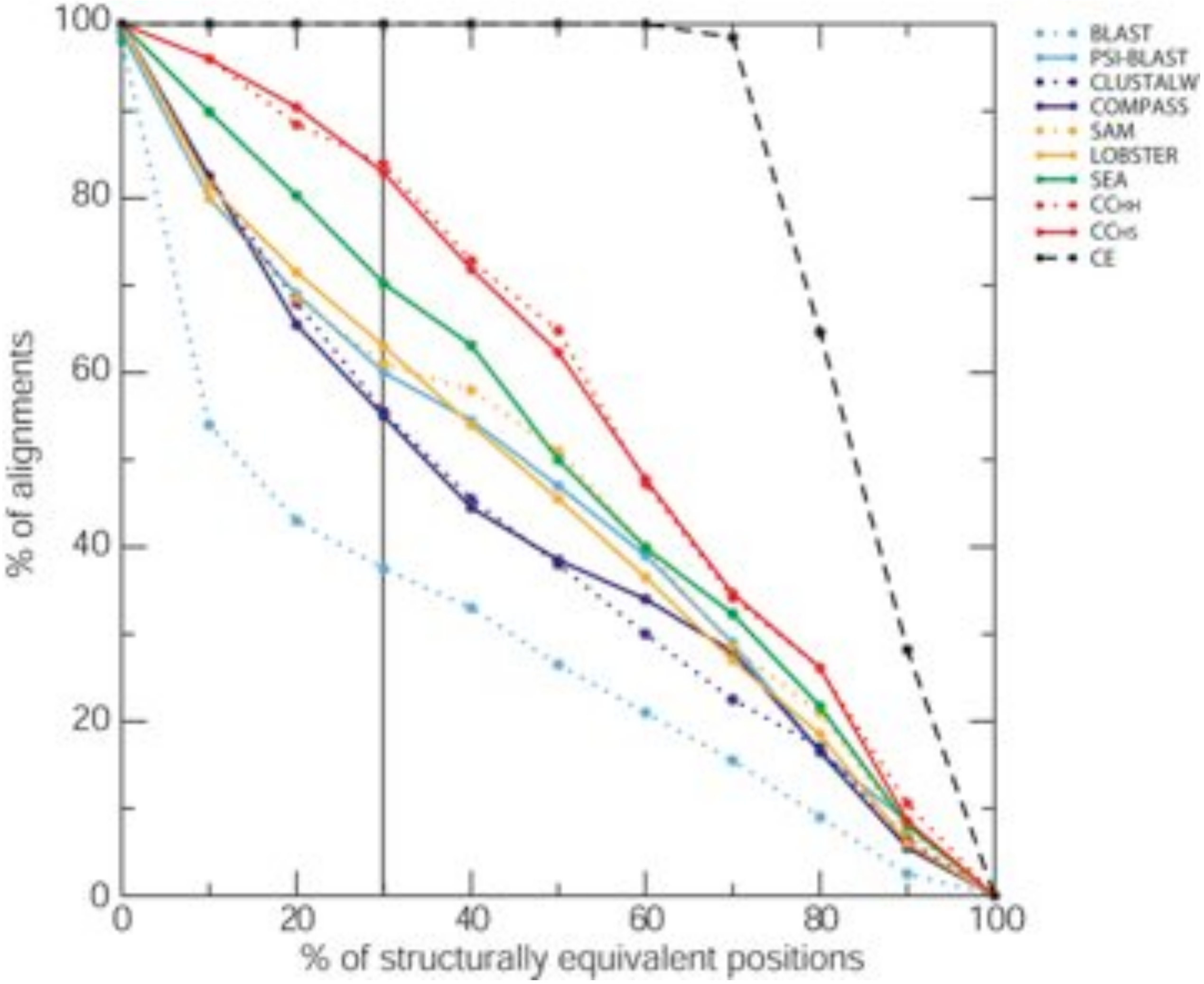
PP_SCAN accuracy

Method	CE overlap	Shift score
CE	100 ± 0	1.00 ± 0.00
BLAST	26 ± 29	0.32 ± 0.33
PSI-BLAST	43 ± 31	0.48 ± 0.35
SAM	48 ± 26	0.50 ± 0.34
LOBSTER	50 ± 27	0.51 ± 0.32
SEA	49 ± 27	0.53 ± 0.29
ALIGN	42 ± 25	0.44 ± 0.28
CLUSTALW	43 ± 27	0.44 ± 0.31
COMPASS	43 ± 32	0.49 ± 0.35
CCHH	**56 ± 23**	**0.61 ± 0.24**
CCHS	**56 ± 24**	**0.62 ± 0.24**
TOP	62 ± 20	0.67 ± 0.20



	CCHH	CCHS	LOBSTER	SEA	SAM	COMPASS	PSI-BLAST	CLUSTALW	ALIGN	BLAST
BLAST										
ALIGN										15.81
CLUSTALW									0.94	16.88
PSI-BLAST								0.12	1.14	17.47
COMPASS							0.63	0.60	1.53	17.78
SAM						5.14	5.70	5.74	6.68	22.70
SEA					1.23	6.41	6.97	6.96	7.91	24.09
LOBSTER				0.33	1.54	6.68	7.24	7.28	8.21	24.25
CCHS			5.55	6.21	7.06	12.37	13.10	12.86	14.06	30.26
CCHH		0.78	6.33	6.99	7.84	13.15	13.87	13.64	14.84	31.02

PP_SCAN success



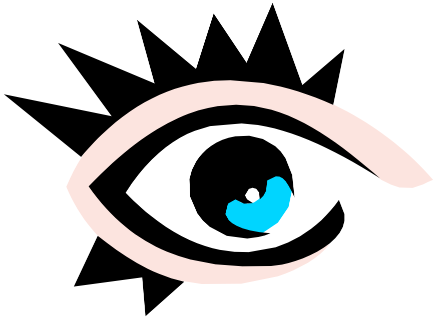
Alignment accuracy (CE overlap)

200 pairwise DBAli alignments

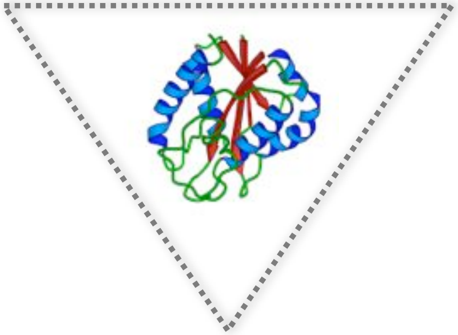
PSI-BLAST (sequence-profile alignment)	43%
SEA (local structure alignment)	49%
PP_SCAN (profile-profile alignment)	56%

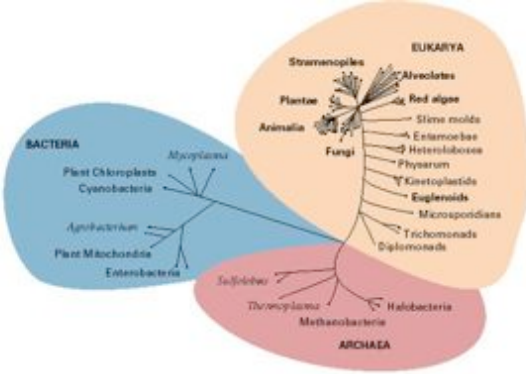
model building and model assessment

Information about a protein can come from three distinct sources

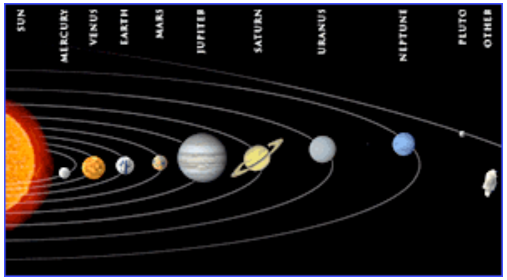


Experimental observations





Statistical rules



Laws of physics

Classes of methods for comparative protein structure modeling

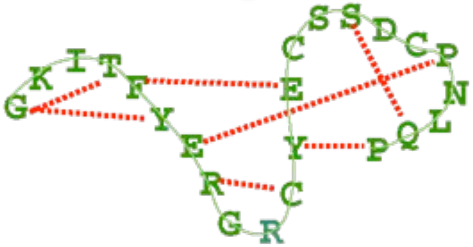
- Model building by assembly of rigid bodies
  core, loops, sidechains.
- Model building by segment matching.
- Model building by satisfaction of spatial restraints.

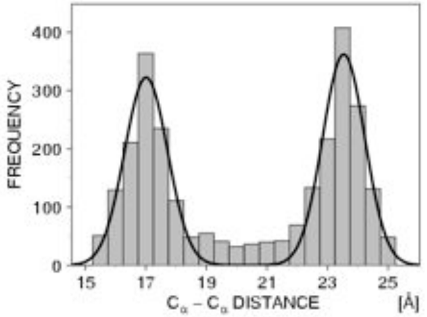
Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

Comparative modeling by satisfaction of spatial restraints
MODELLER

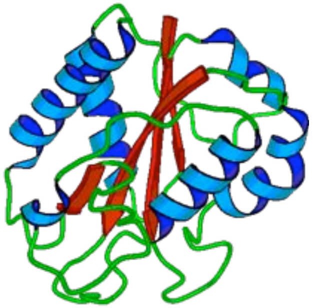
```
3D   GKITFYERGFQGHCYESDC-NLQP…
SEQ  GKITFYERG---RCYESDCPNLQP…
```

1. Extract spatial restraints





2. Satisfy spatial restraints



$$F(R) = \prod_i p_i\,(f_i\,/I)$$

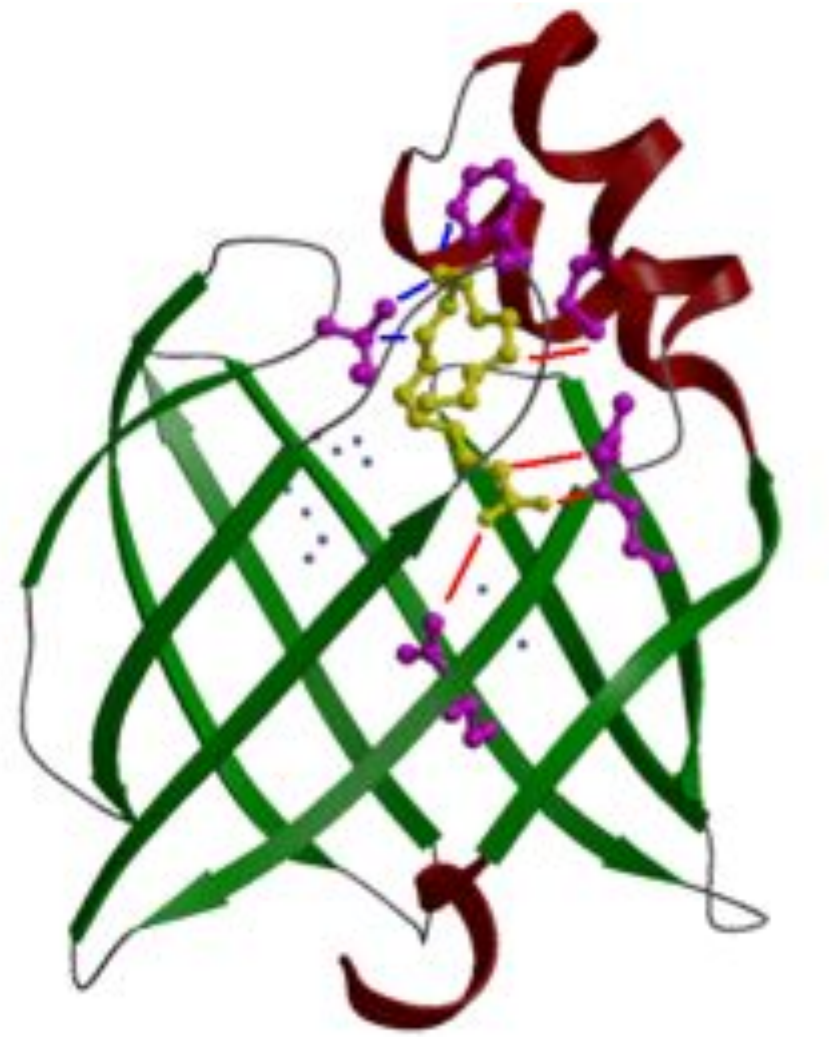
A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.

Multiple Templates

Local similarity
extracted from
closest template

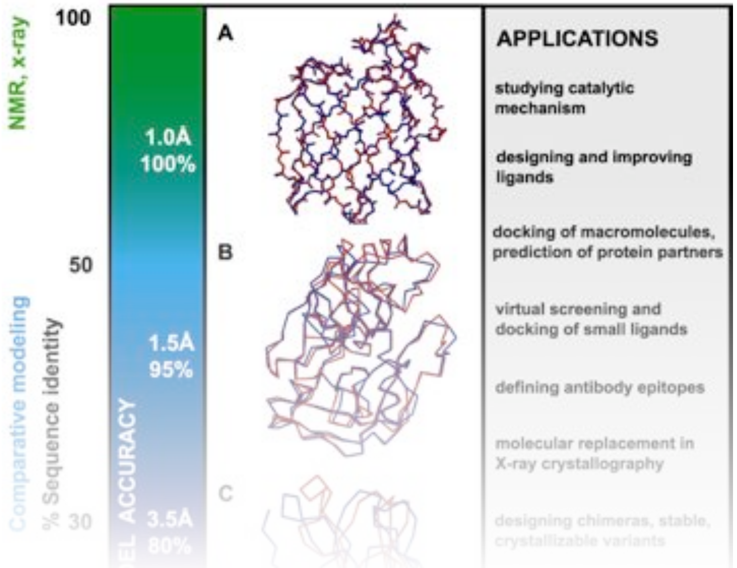
```
Templates   KSINPIHGDNCEQTSDEGLKIERTPL--------QWLKSSICDMRGLIPE
            ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
Target      MSVIPKRLYGNCEQTSEEAIRIEDSPIVRWISAQLVCLKIDEIPERLVGE
```

Modeling ligands and using external restraints



🟦 Homology derived restraint

🟥 External Restraint

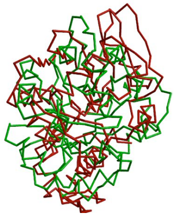


APPLICATIONS

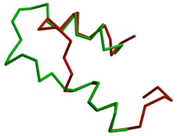
- studying catalytic mechanism
- designing and improving ligands
- docking of macromolecules, prediction of protein partners
- virtual screening and docking of small ligands
- defining antibody epitopes
- molecular replacement in X-ray crystallography
- designing chimeras, stable, crystallizable variants

Accuracy and applicability of comparative models

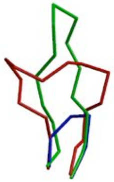
Comparative modeling by satisfaction of spatial restraints Types of errors and their impact



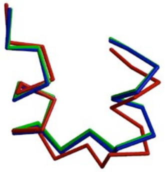
Wrong fold



Miss alignments

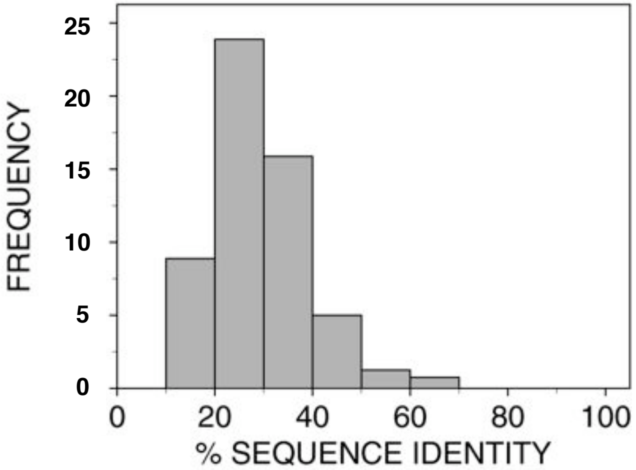


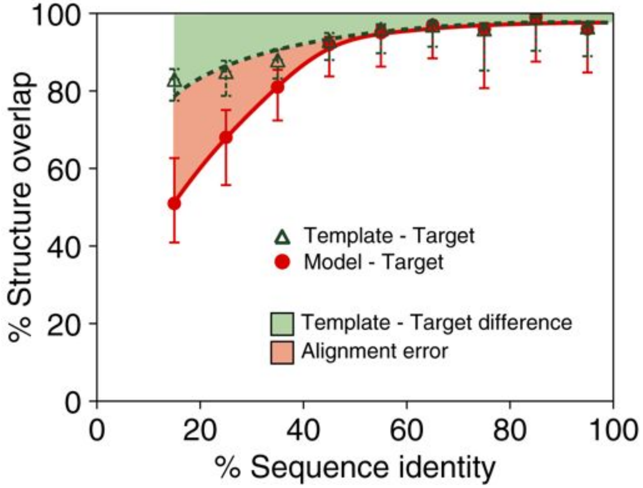
Loop regions



Rigid body distortions

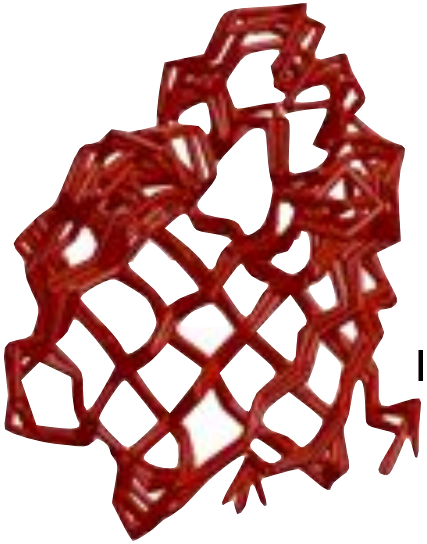
Side-chain packing





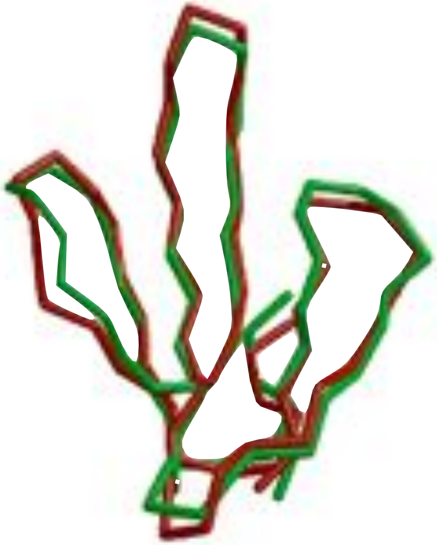
Marti-Renom etal. Ann Rev Biophys Biomol Struct (2000) 29, 291

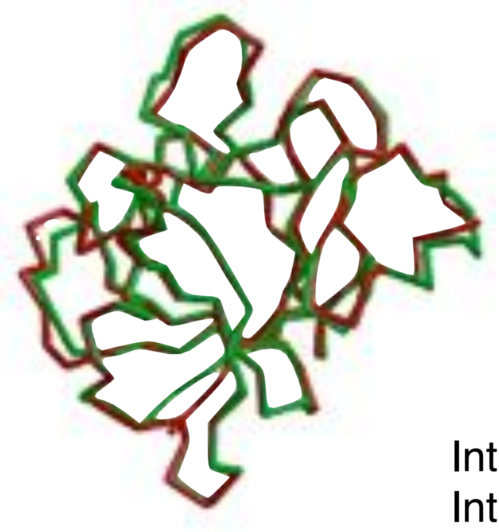
“Biological” significance of modeling errors



NMR
Ileal lipid-binding protein
1eal

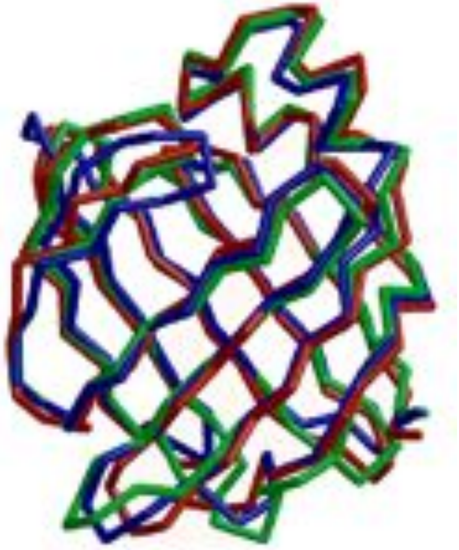
NMR – X-RAY
Erabutoxin 3ebx
Erabutoxin 1era





X-RAY
Interleukin 1β 41bi (2.9Å)
Interleukin 1β 2mib (2.8Å)

CRABPII 1opbB
FABP 1ftpA
ALBP 1lib
40% seq. id.

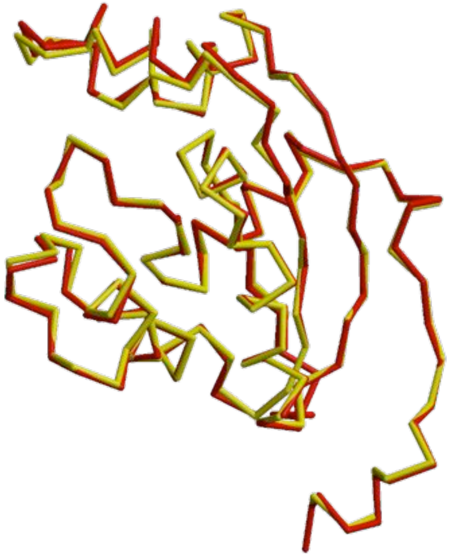


Model Accuracy

HIGH ACCURACY

NM23 Seq id 77%

$C\alpha$ equiv 147/148
RMSD 0.41Å

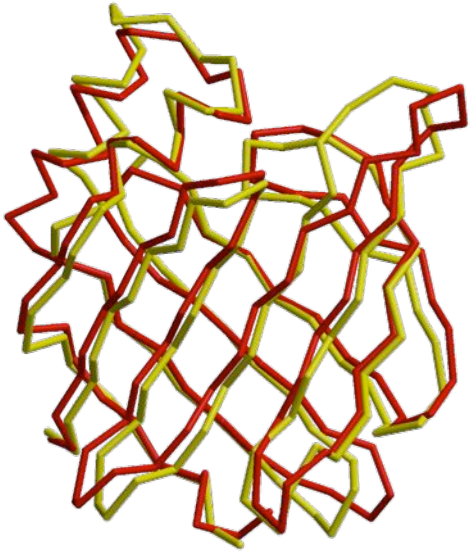


Sidechains
Core backbone
Loops

MEDIUM ACCURACY

CRABP Seq id 41%

$C\alpha$ equiv 122/137
RMSD 1.34Å

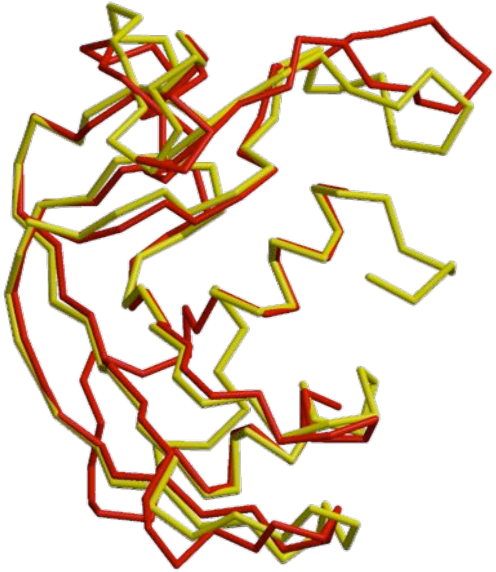


Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

EDN Seq id 33%

$C\alpha$ equiv 90/134
RMSD 1.17Å

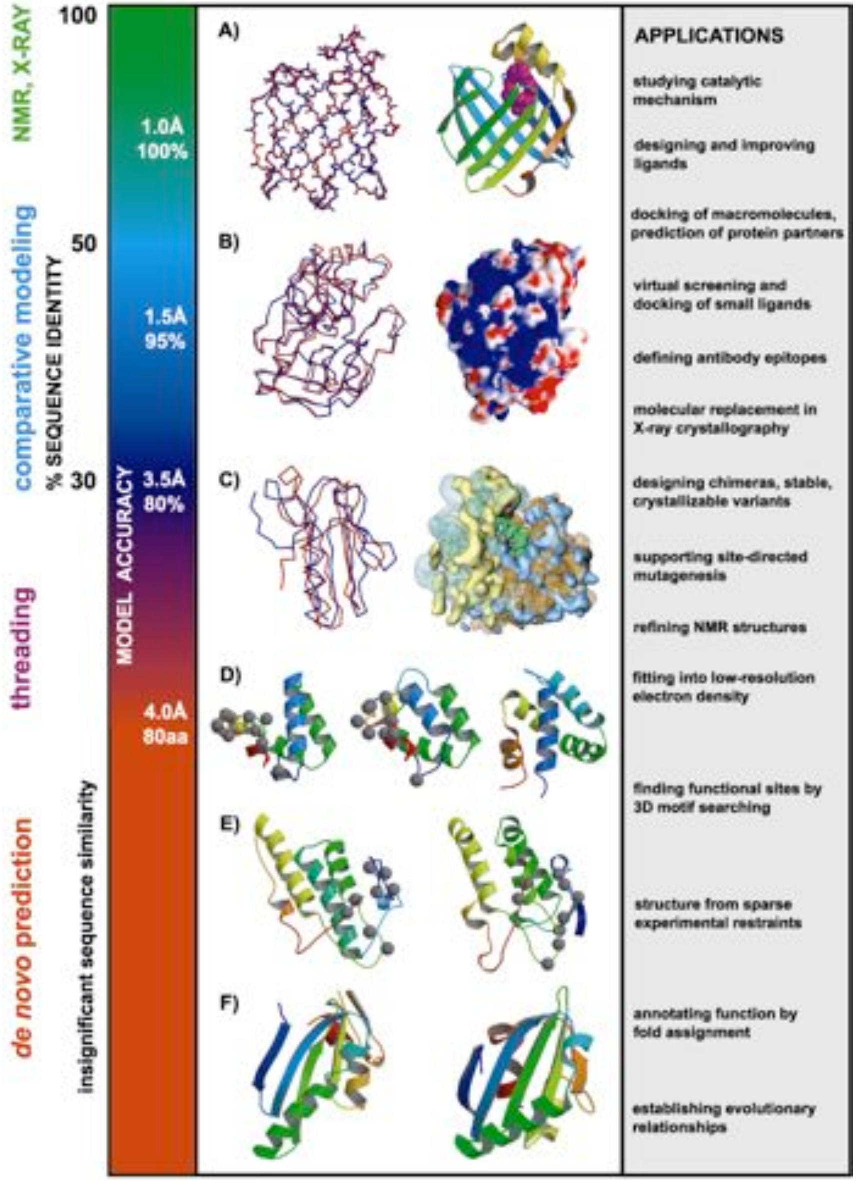


Sidechains
Core backbone
Loops
Alignment
Fold assignment

X-RAY / MODEL

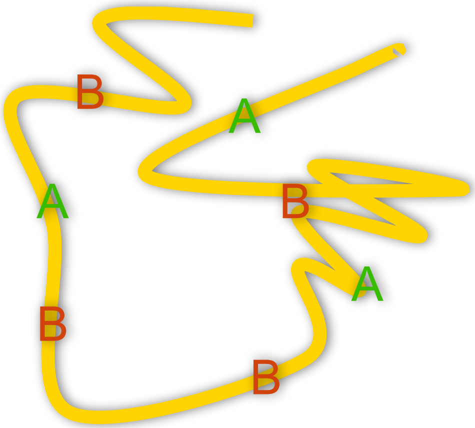
Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

Utility of protein structure models, despite errors


NMR, X-RAY
comparative modeling
threading
de novo prediction
% SEQUENCE IDENTITY
100
50
30
insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4.0Å
80aa
A)
B)
C)
D)
E)
F)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
finding functional sites by 3D motif searching
structure from sparse experimental restraints
annotating function by fold assignment
establishing evolutionary relationships

D. Baker & A. Sali. Science 294, 93, 2001.

Model Assessment (PMF)



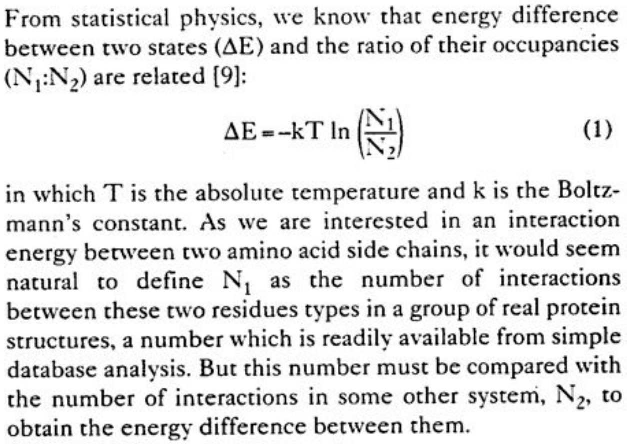
Scoring

Statistical Potential (inspiration)

$$K = \frac{[AB]}{[A]\cdot[B]}$$

$$\Delta G = -RT\ln(K) = -RT\ln\frac{[AB]}{[A]\cdot[B]}$$

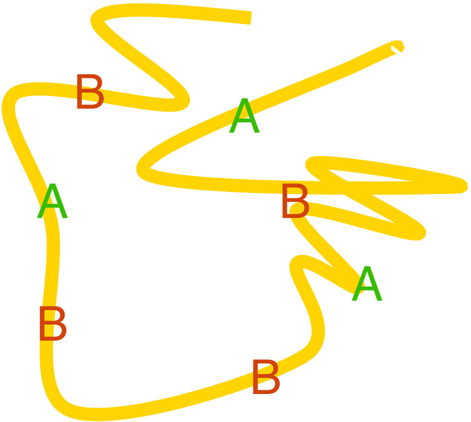
$$A + B \rightleftharpoons AB$$



From statistical physics, we know that energy difference between two states (ΔE) and the ratio of their occupancies (N_1:N_2) are related [9]:

$$\Delta E = -kT \ln\left(\frac{N_1}{N_2}\right) \qquad (1)$$

in which T is the absolute temperature and k is the Boltzmann's constant. As we are interested in an interaction energy between two amino acid side chains, it would seem natural to define N_1 as the number of interactions between these two residues types in a group of real protein structures, a number which is readily available from simple database analysis. But this number must be compared with the number of interactions in some other system, N_2, to obtain the energy difference between them.

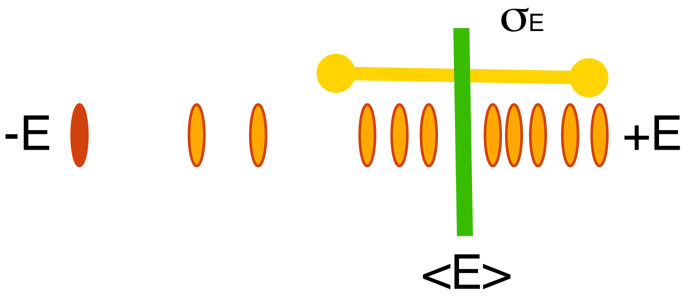


*Tanaka and Sheraga (1975) PNAS, **72** pp3802*
*Sippl, (1990) J.Mo.Biol. **213** pp859*
*Godzik, (1996) Structure **15** pp363*

Scoring

Significance of an alignment (score)

Energy Z-score the model with respect the energy of random models (or rest of decoys).



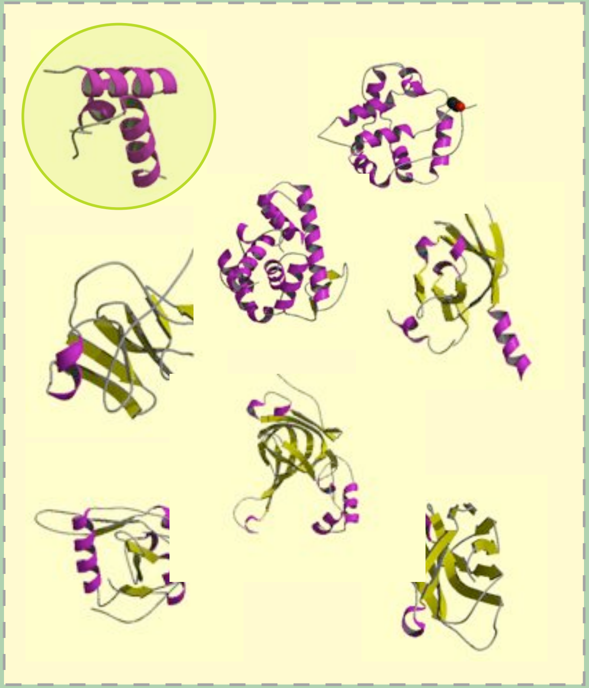
$$Zscore = \frac{\left(\langle E \rangle - E_m\right)}{\sigma_E}$$

Prosall

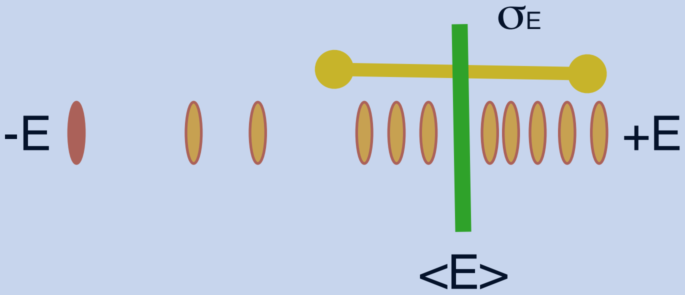
http://www.came.sbg.ac.at

Deriving

Structural space



Scoring



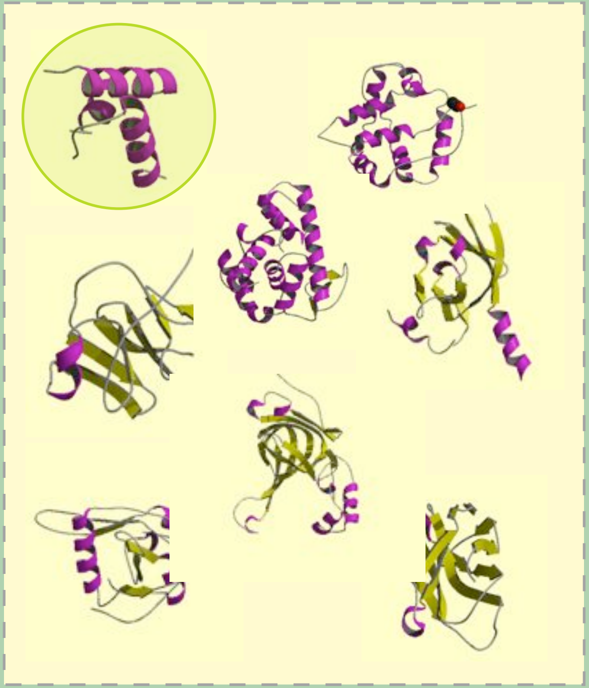
$$Zscore = \frac{\left(\langle E \rangle - E_{m}\right)}{\sigma_{E}}$$

ANOLEA

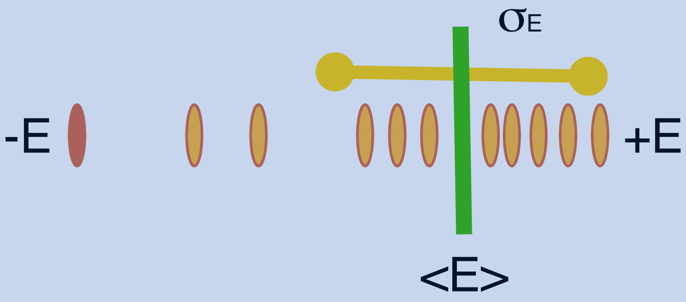
http://protein.bio.puc.cl/cardex/servers/anolea/

Deriving

Structural space



Scoring



$$Zscore = \frac{\left(\langle E \rangle - E_{m}\right)}{\sigma_{E}}$$

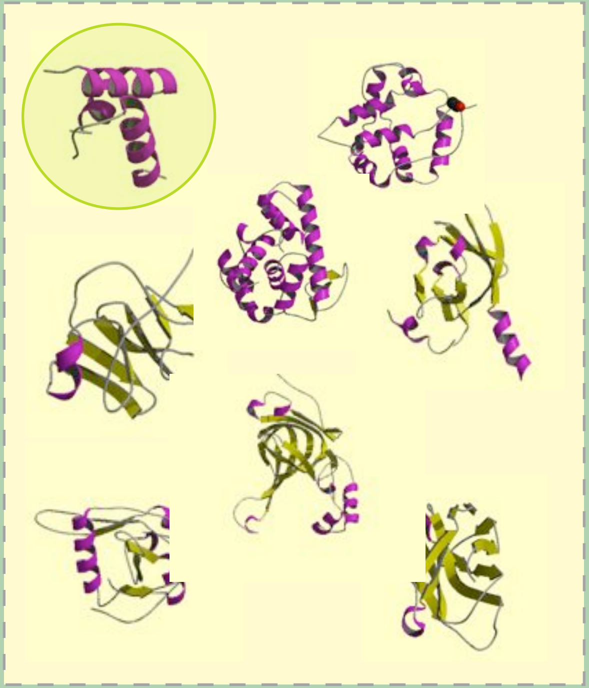
all atom potential

Verify3D

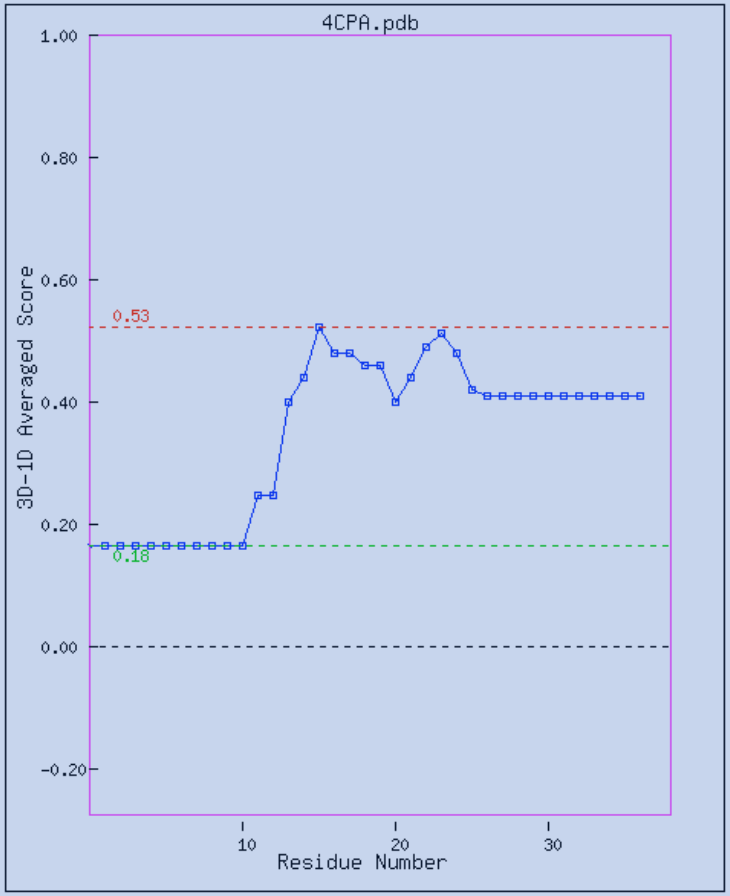
http://nihserver.mbi.ucla.edu/Verify_3D/

Deriving

Structural space



Scoring

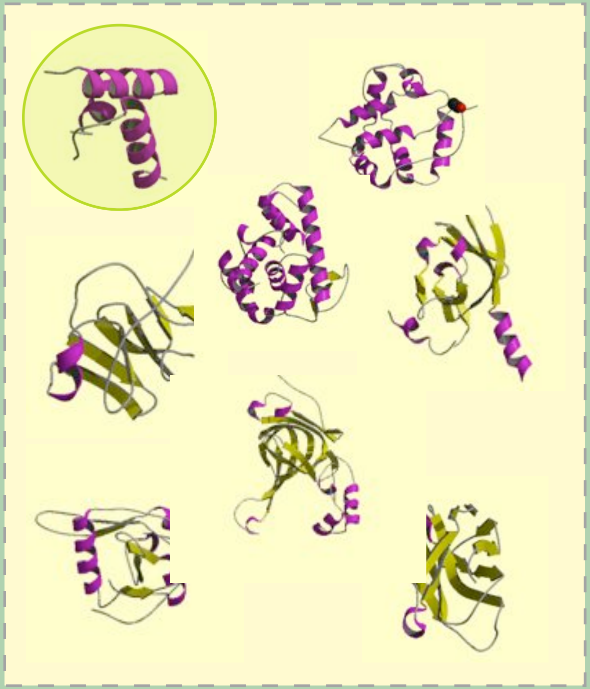


DFIRE

http://sparks.informatics.iupui.edu/

Deriving

Structural space



Scoring

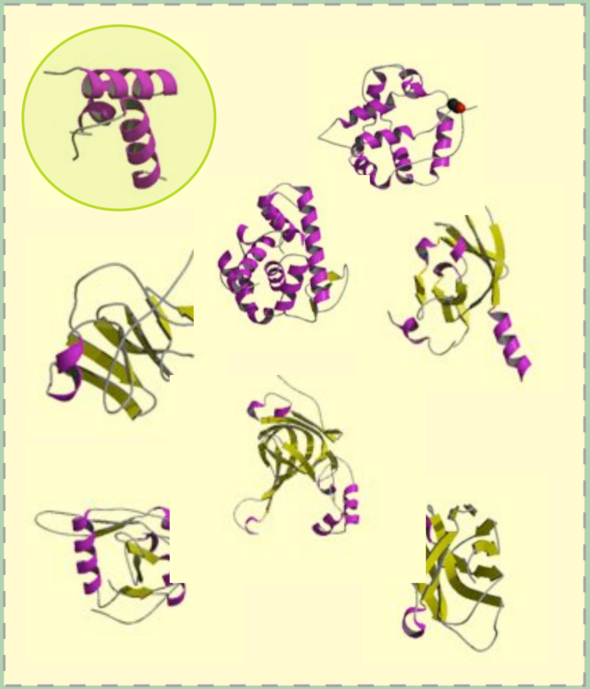
Pseudo-Energy with respect a ideal gas-phase reference state

DOPE (MODELLER)

http://www.salilab.org/modeller/

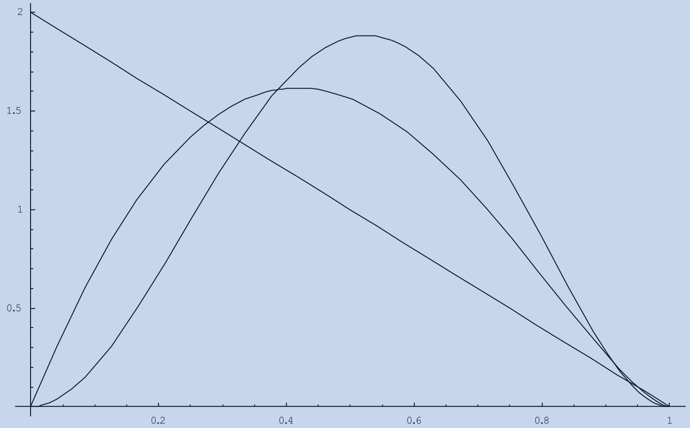
Deriving

Structural space



Scoring

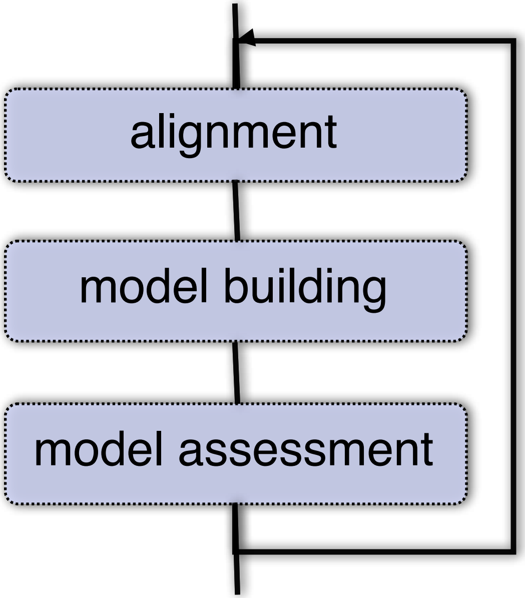
Pseudo-Energy with respect a ideal spherical protein as a reference state

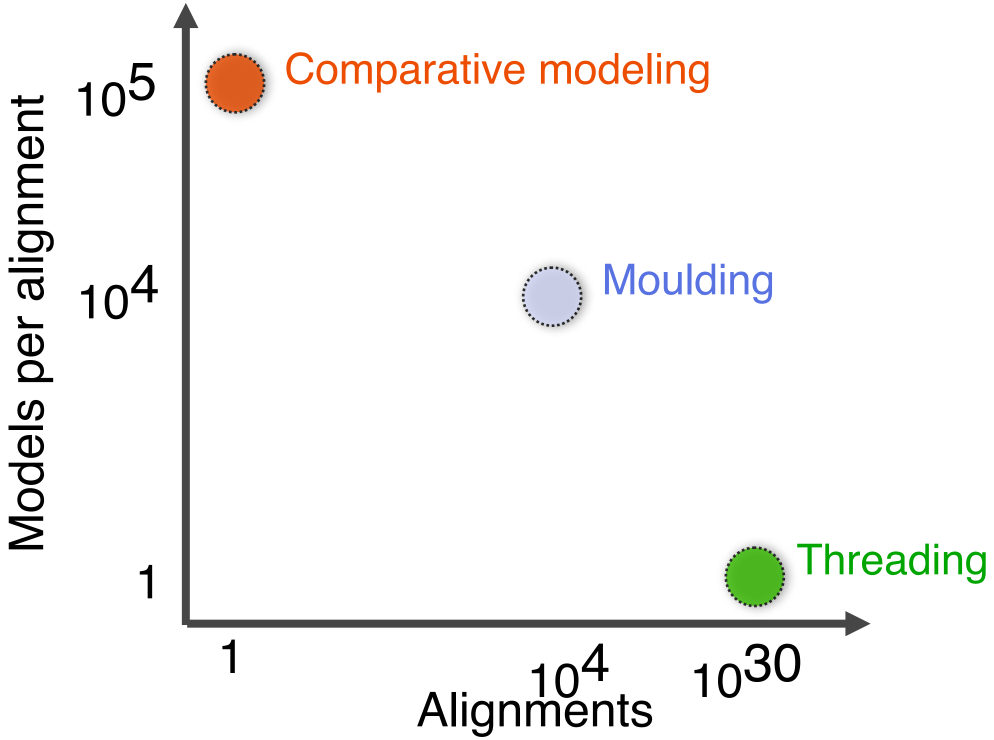


MOULDER

John, Sali (2003). NAR pp31 3982

Moulding: iterative alignment, model building, model assessment





Genetic algorithm operators

Single point cross-over

```
...TSSQ—NMKLGVFWGY——...
...V—SSCN——GDLHMKVGV...

...TSSQNMK——LGVFWGY...
...VSSCNGDLHMKV——GV...
```

→

```
...TSSQ—NMK——LGVFWGY...
...V—SSCNGDLHMKV——GV...

...TSSQNMKLGVFWGY——...
...VSSCN——GDLHMKVGV...
```

Gap insertion

```
...TSSQNMKLGVFWGY...
...VSSCNGDLHMKVGV...
```

→

```
...TSSQN——MKLGVFWGY...
...VSSCNGDLHMKVG——V...
```

Gap shift

```
...T——SSQNMKLGVFWGY...
...VSSCNGDLHMKVGV——...
```

→

```
...—T—SSQNMKLGVFWGY...
...VSSCNGDLHMKVGV——...

...T—S—SQNMKLGVFWGY...
...VSSCNGDLHMKVGV——...

...——TSSQNMKLGVFWGY...
...VSSCNGDLHMKVGV——...

...TS——SQNMKLGVFWGY...
...VSSCNGDLHMKVGV——...
```

Also, “two point crossover” and “gap deletion”.

Composite model assessment score

Weighted linear combination of several scores:

- Pair (P_p) and surface (P_s) statistical potentials;
- Structural compactness (S_c);
- Harmonic average distance score (H_a);
- Alignment score (A_s).

$$Z = 0.17\ Z(P_P) + 0.02\ Z(P_S) + 0.10\ Z(S_C) + 0.26\ Z(H_a) + 0.45\ (A_S)$$

$Z(\text{score}) = (\text{score}- \mu)/\sigma$

μ … average score of all models

σ … standard deviation of the scores

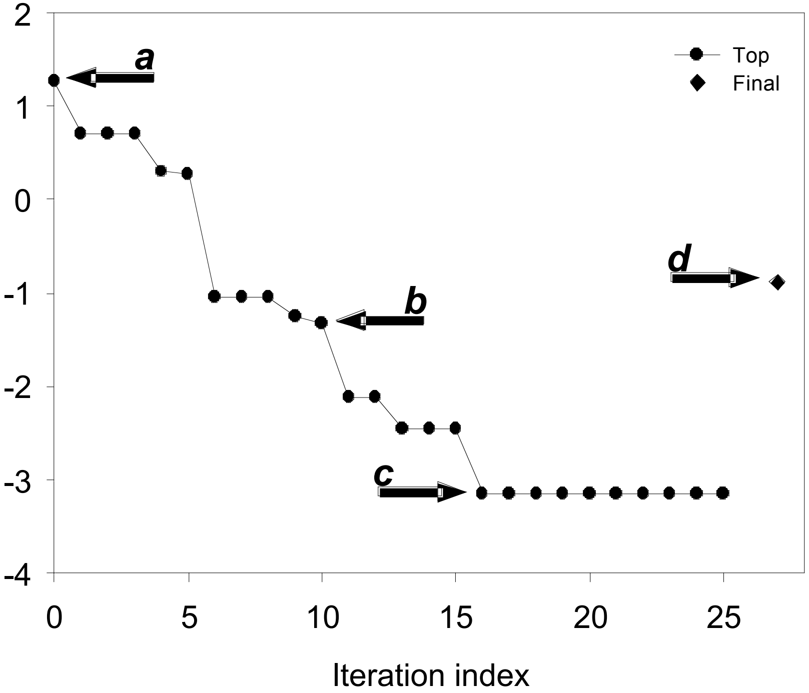
Benchmark with the “very difficult” test set

D. Fischer threading test set of 68 structural pairs (a subset of 19)

Target -template	Sequence identity [%]	Coverage [% aa]	Initial prediction		Final prediction		Best prediction	
			Cα RMSD [Å]	CE overlap [%]	Cα RMSD [Å]	CE overlap [%]	Cα RMSD [Å]	CE overlap [%]
1ATR-1ATN	13.8	94.3	19.2	20.2	18.8	20.2	17.1	24.6
1BOV-1LTS	4.4	83.5	10.1	29.4	3.6	79.4	3.1	92.6
1CAU-1CAU	18.8	96.7	11.7	15.6	10.0	27.4	7.6	47.4
1COL-1CPC	11.2	81.4	8.6	44.0	5.6	58.6	4.8	59.3
1LFB-1HOM	17.6	75.0	1.2	100.0	1.2	100.0	1.1	100.0
1NSB-2SIM	10.1	89.2	13.2	20.2	13.2	20.1	12.3	26.8
1RNH-1HRH	26.6	91.2	13.0	21.2	4.8	35.4	3.5	57.5
1YCC-2MTA	14.5	55.1	3.4	72.4	5.3	58.4	3.1	75.0
2AYH-1SAC	8.8	78.4	5.8	33.8	5.5	48.0	4.8	64.9
2CCY-1BBH	21.3	97.0	4.1	52.4	3.1	73.0	2.6	77.0
2PLV-1BBT	20.2	91.4	7.3	58.9	7.3	58.9	6.2	60.7
2POR-2OMF	13.2	97.3	18.3	11.3	11.4	14.7	10.5	25.9
2RHE-1CID	21.2	61.6	9.2	33.7	7.5	51.1	4.4	71.1
2RHE-3HLA	2.4	96.0	8.1	16.5	7.6	9.4	6.7	43.5
3ADK-1GKY	19.5	100.0	13.8	26.6	11.5	37.7	7.7	48.1
3HHR-1TEN	18.4	98.9	7.3	60.9	6.0	66.7	4.9	79.3
4FGF-81IB	14.1	98.6	11.3	24.0	9.3	30.6	5.4	41.2
6XIA-3RUB	8.7	44.1	10.5	14.5	10.1	11.0	9.0	34.3
9RNT-2SAR	13.1	88.5	5.8	41.7	5.1	51.2	4.8	69.0
AVERAGE	**14.2**	**85.2**	**9.6**	**36.7**	**7.7**	**44.8**	**6.3**	**57.8**

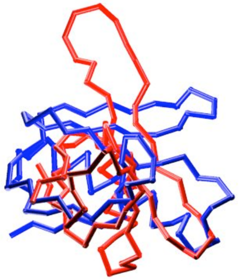
Application to a difficult modeling case

1BOV-1LTS

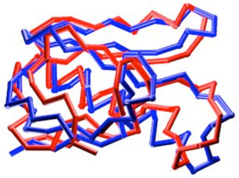


Sequence identity	4.4%
Initial model Cα RMSD	10.1Å
Final model Cα RMSD	3.6Å

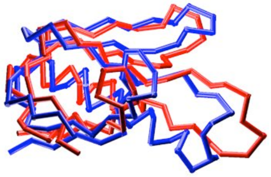
a



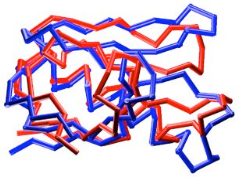
b



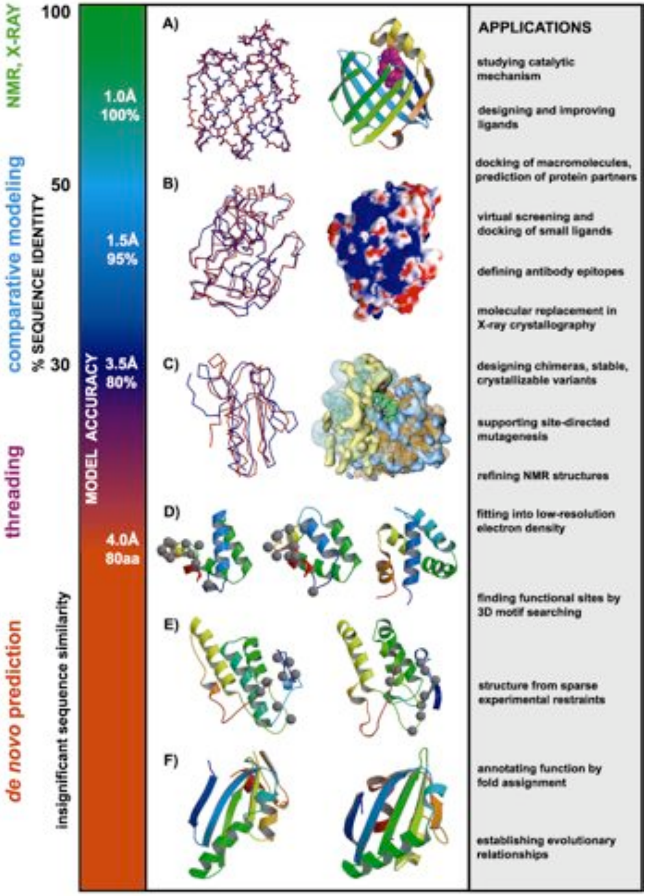
c

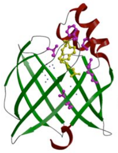


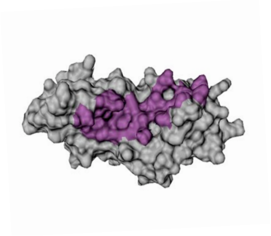
d

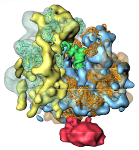


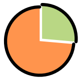
Can we use models to infer function?


NMR, X-RAY
comparative modeling
threading
de novo prediction
% SEQUENCE IDENTITY
100
50
30
insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4.0Å
80aa
A)
B)
C)
D)
E)
F)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
finding functional sites by 3D motif searching
structure from sparse experimental restraints
annotating function by fold assignment
establishing evolutionary relationships

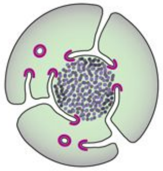


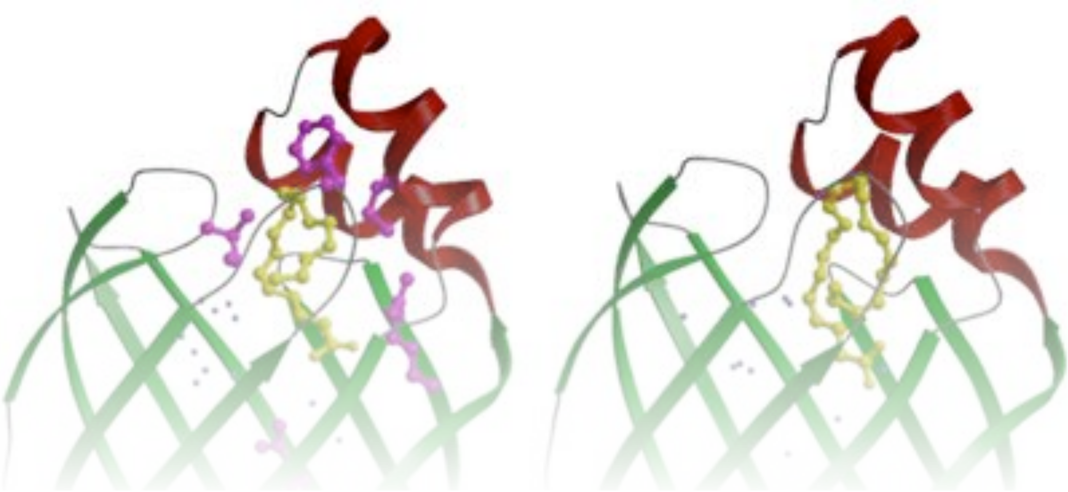






T. cruzi





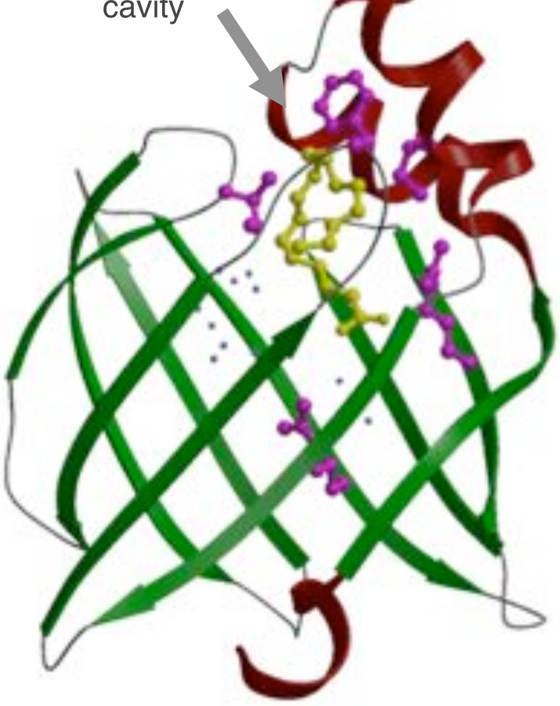
Modeling genes

What is the physiological ligand of Brain Lipid-Binding Protein?

Predicting features of a model that are not present in the template

BLBP/oleic acid

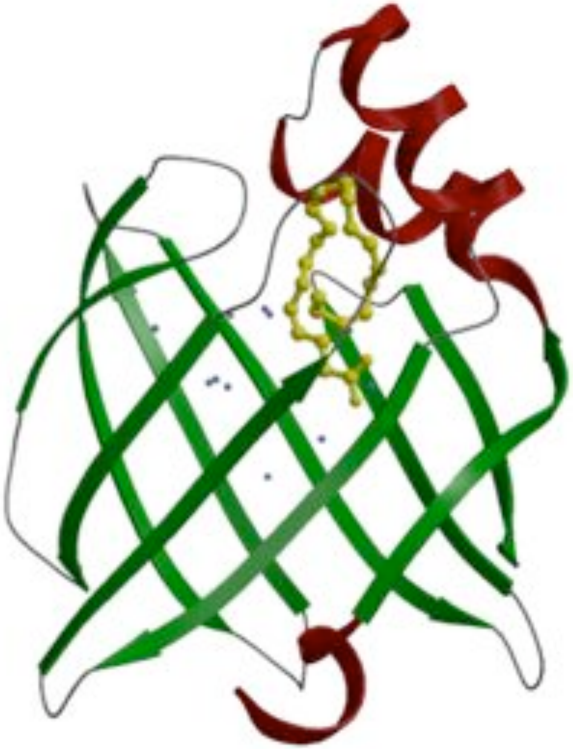
Cavity is not filled



Ligand binding cavity

BLBP/docosahexaenoic acid

Cavity is filled



1. BLBP binds fatty acids.

2. Build a 3D model.

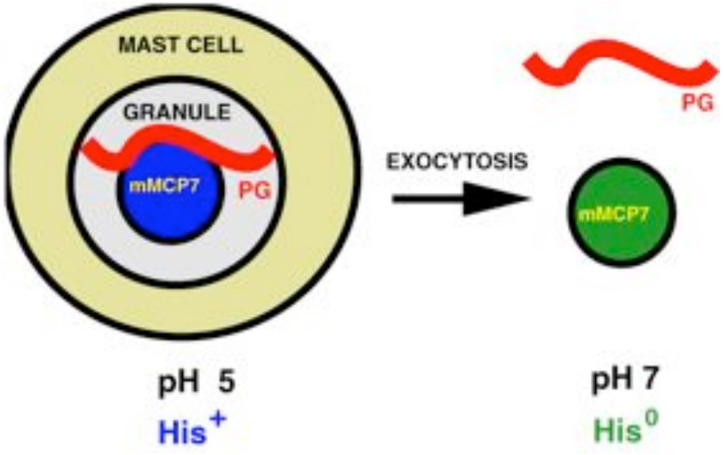
3. Find the fatty acid that fits most snuggly into the ligand binding cavity.

L. Xu, R. Sánchez, A. Šali, N. Heintz, J. Biol. Chem. 271, 24711, 1996.

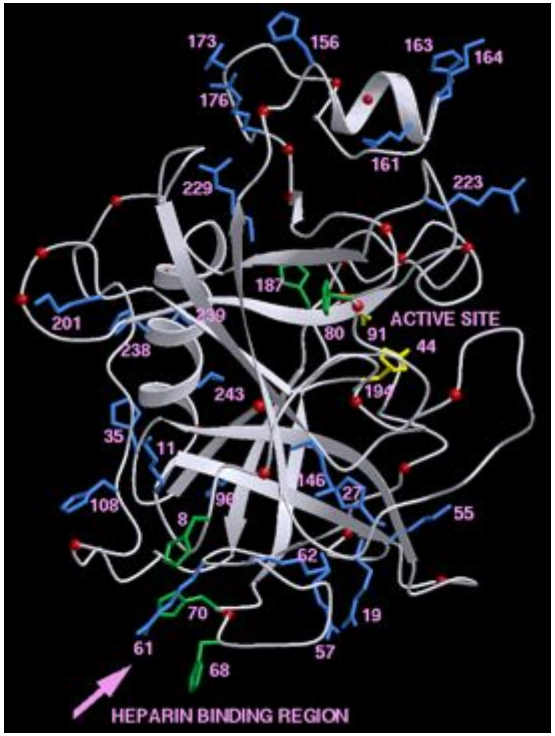
Do mast cell proteases bind proteoglycans? Where? When?

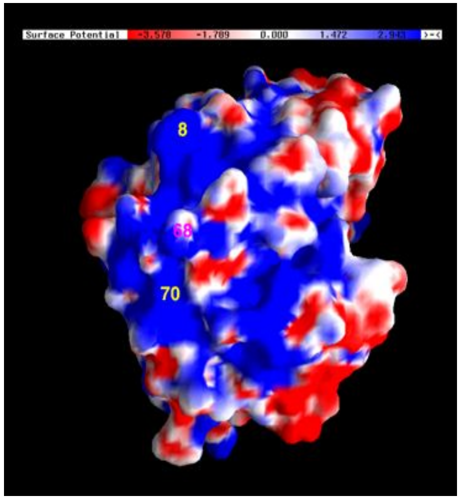
Predicting features of a model that are not present in the template

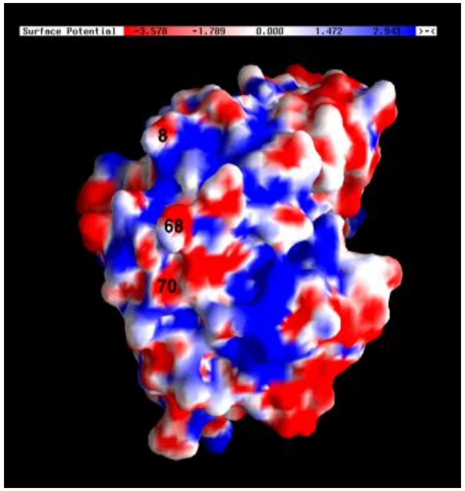
1. mMCPs bind negatively charged proteoglycans through electrostatic interactions
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.



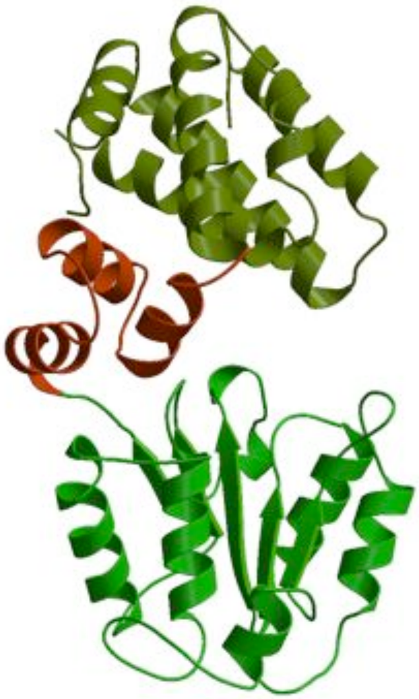
Huang *et al*. *J. Clin. Immunol*. **18**,169,1998.
Matsumoto *et al*. *J.Biol.Chem*. **270**,19524,1995.
Šali *et al*. *J. Biol. Chem*. **268**, 9023, 1993.



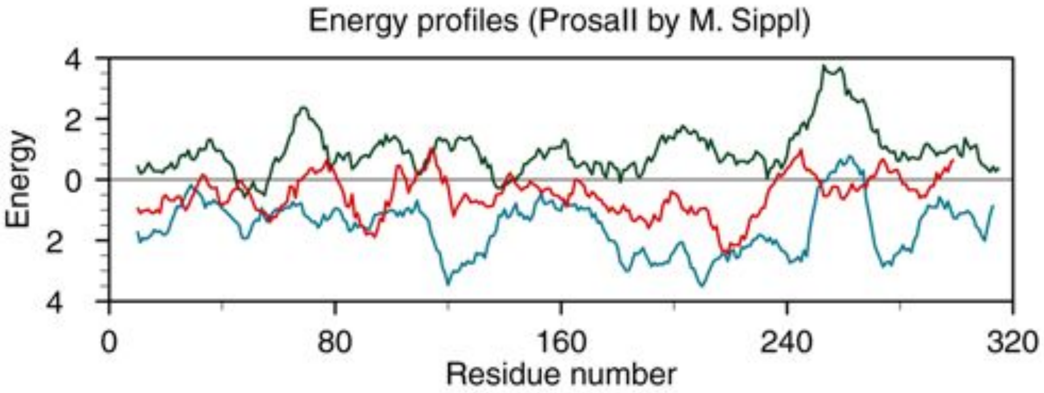




Does RuvB have the same fold as δ’ of E.coli DNA polymerase III?

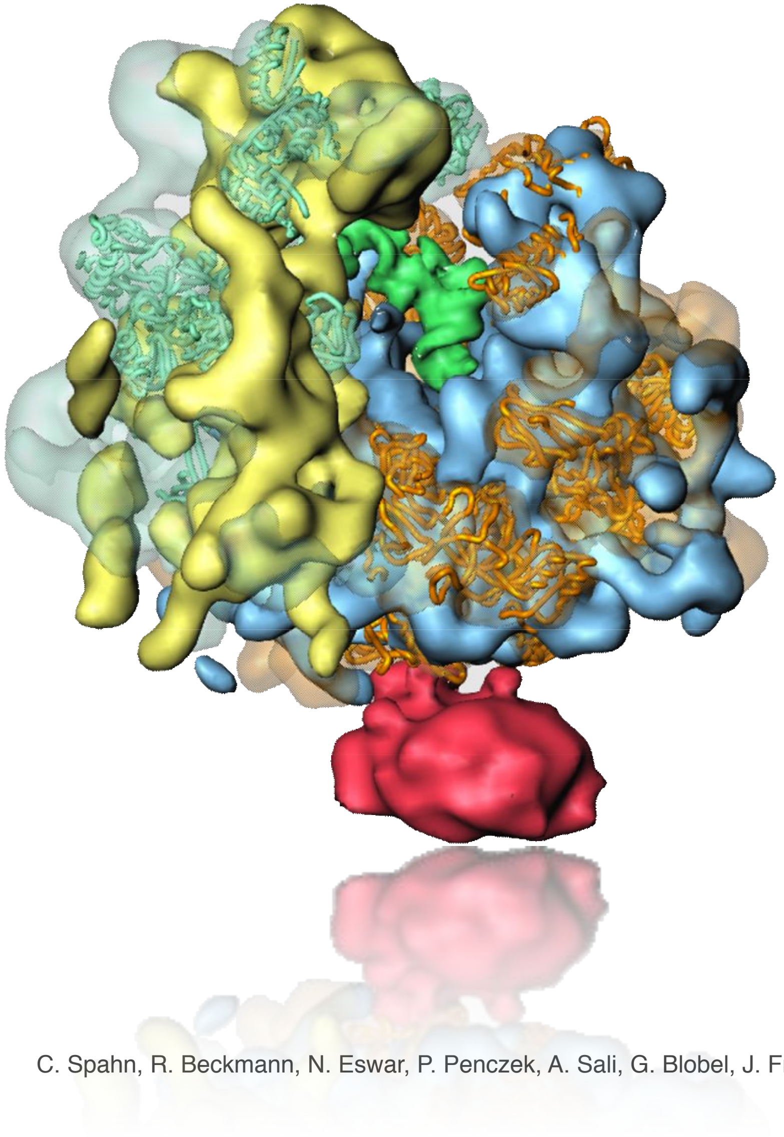


```
Ec d’   MRWYPWLRPDFEKLVASYQAGRG----HHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRG
RUVB    LEEYVGQPQVRSQMEIFIKAAKLRGDALDHLLIFGPPGLGKTTLANIVANEMG---------------

Ec d’   CQLMQAGTHPDYYTLAPEKGKATLGVDAVREVTEKLNEAARLGGAKVVWVTDAALLTDAAANALLKTL
RUVB    -----------VNLRTT-------SGPVLEKAGDLAAMLTNLEPHDVLFIDEIHRLSPVVEEVLYPAM

Ec d’   ------------------EEPPAETWFFLATREPERL---LATLRSRCRLHYLAPPPEQYAVTWLSRE
Ppdp    EDYQLDIMIGEGPAARSIKIDLPPFTLIGATTRAGSLTSPLRDRFGIVQRLEFY--QVPDLQYIVSRS

Ec d’   VTM-----SQDALLAALRLSAGSPGAALALFQ------------GDNWQARETLCQALAYSVPSGD--
RUVB    ARFMGLEMSDDGALEVARRARGTPRIANRLLRRVRDFAEVKHDGTISADIAAQALDMLNVDAEGFDYM

Ec d’   -WYSLLAALN---HEQAPARLHWLATLLMDALKR/VTNVDVPGLVAELANHL---SPSRLQAILGDVC
RUVB    DRKLLLAVIDKFF-GGPVGLDNLAAAIGEERETIE--DVLEPYLIQQGFLQRTPRGRMATTRAWNHFG

        Ec d’   HIREQLMSVAGANRELLITDLLLRIEHYLQPGVVLP
        RUVB    ITPPEMP-----------------------------
```



B. Guenther, et al. Cell 91, 335, 1997.
Yamada, K., et al. Proc.Nat.Acad.Sci.USA 98,1442, 2001.

S. cerevisiae ribosome



Fitting of comparative models into 15Å cryo-electron density map.

43 proteins could be modeled on 20-56% seq.id. to a known structure.

The modeled fraction of the proteins ranges from 34-99%.

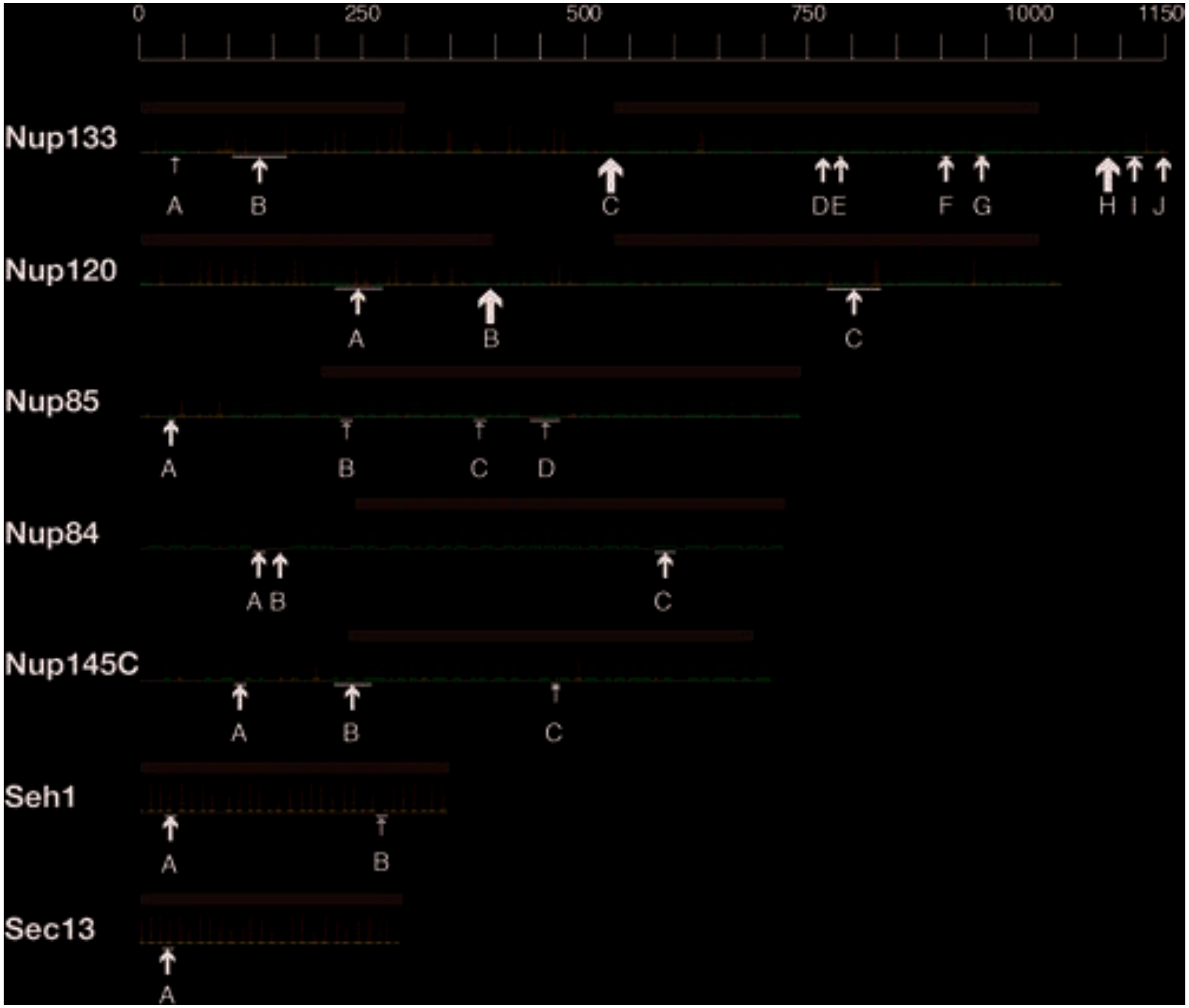
C. Spahn, R. Beckmann, N. Eswar, P. Penczek, A. Sali, G. Blobel, J. Frank. Cell 107, 361-372, 2001.

Common Evolutionary Origin of Coated Vesicles and Nuclear Pore Complexes

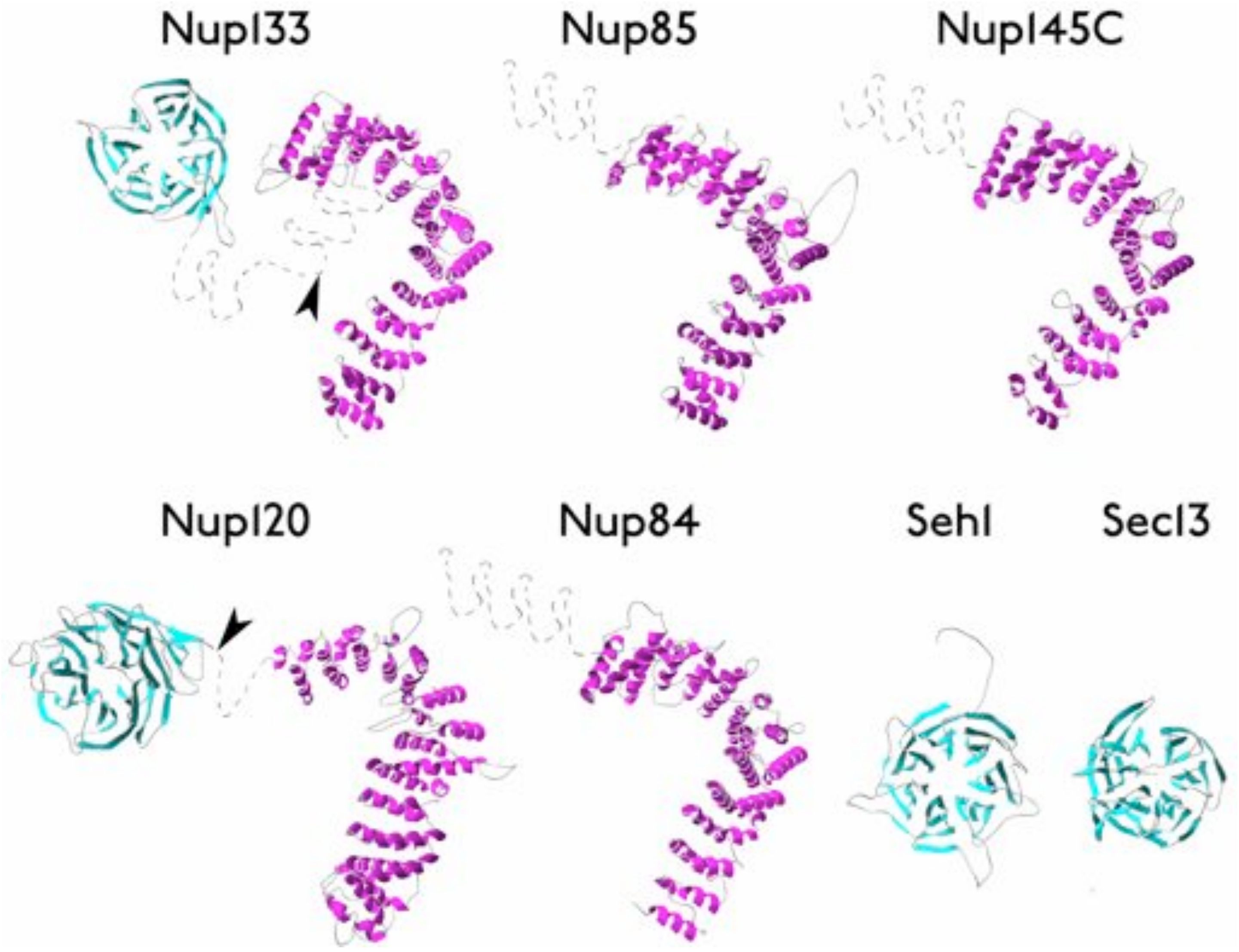
mGenThreader + SALIGN + MOULDER

D. Devos, S. Dokudovskaya, F. Alber, R. Williams, B.T. Chait, A. Sali, M.P. Rout.
Components of Coated Vesicles and Nuclear Pore Complexes Share a Common Molecular Architecture.
PLOS Biology **2(12):**e380, 2004

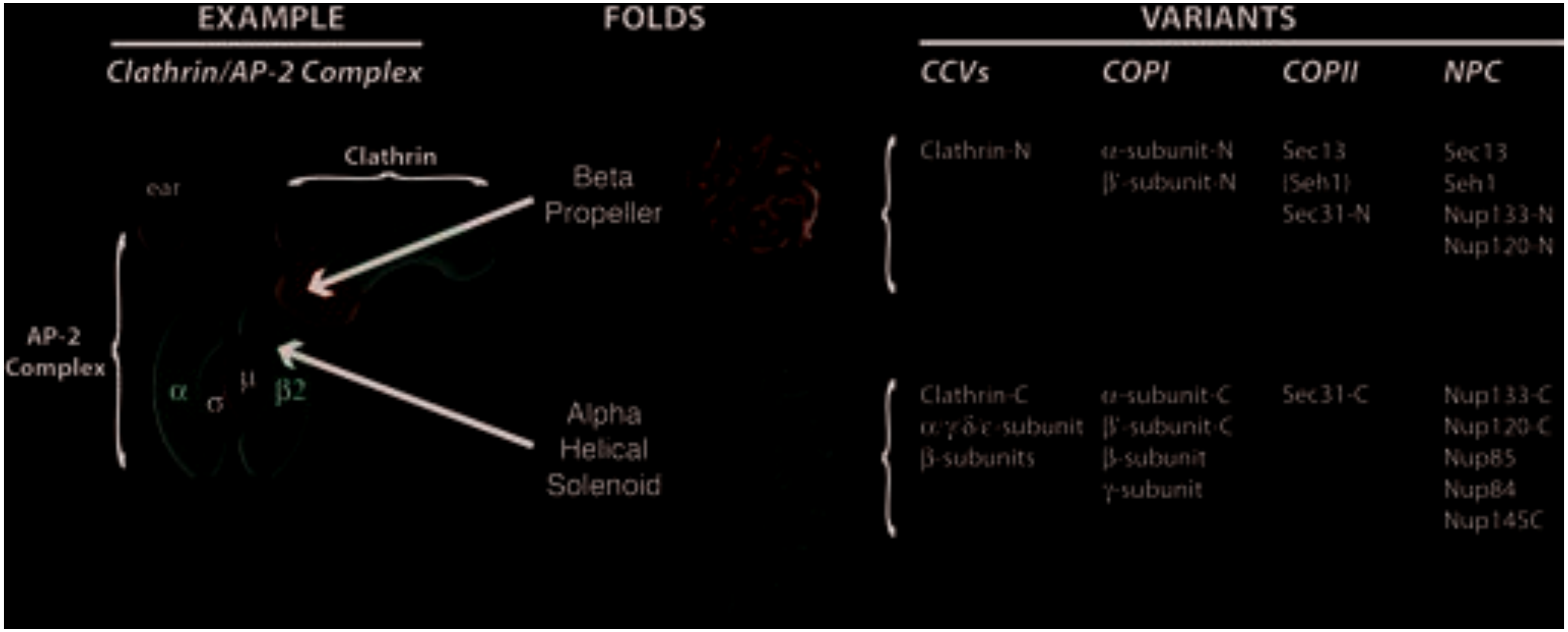
yNup84 complex proteins



All Nucleoporins in the Nup84 Complex are Predicted to Contain β-Propeller and/or α-Solenoid Folds



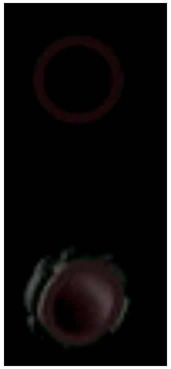
NPC and Coated Vesicles Share the β-**Propeller and** α-**Solenoid** Folds and Associate with Membranes



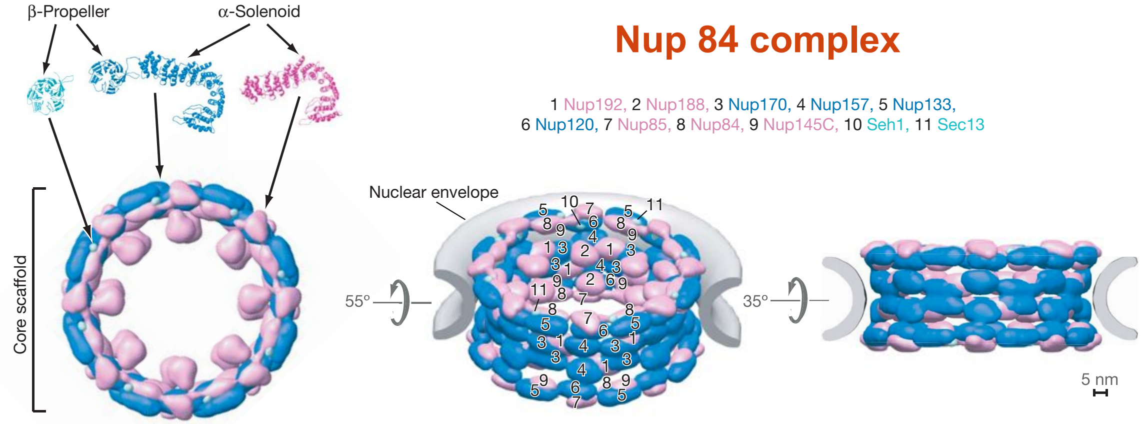
CCVs	COPI	COPII	NPC
Clathrin-N	α-subunit-N β'-subunit-N	Sec13 (Seh1) Sec31-N	Sec13 Seh1 Nup133-N Nup120-N
Clathrin-C α/γ/δ/ε-subunit β-subunits	α-subunit-C β'-subunit-C β-subunit γ-subunit	Sec31-C	Nup133-C Nup120-C Nup85 Nup84 Nup145C

NPC and Coated Vesicles Both Associate with Membranes

Coated Vesicle



NPC model



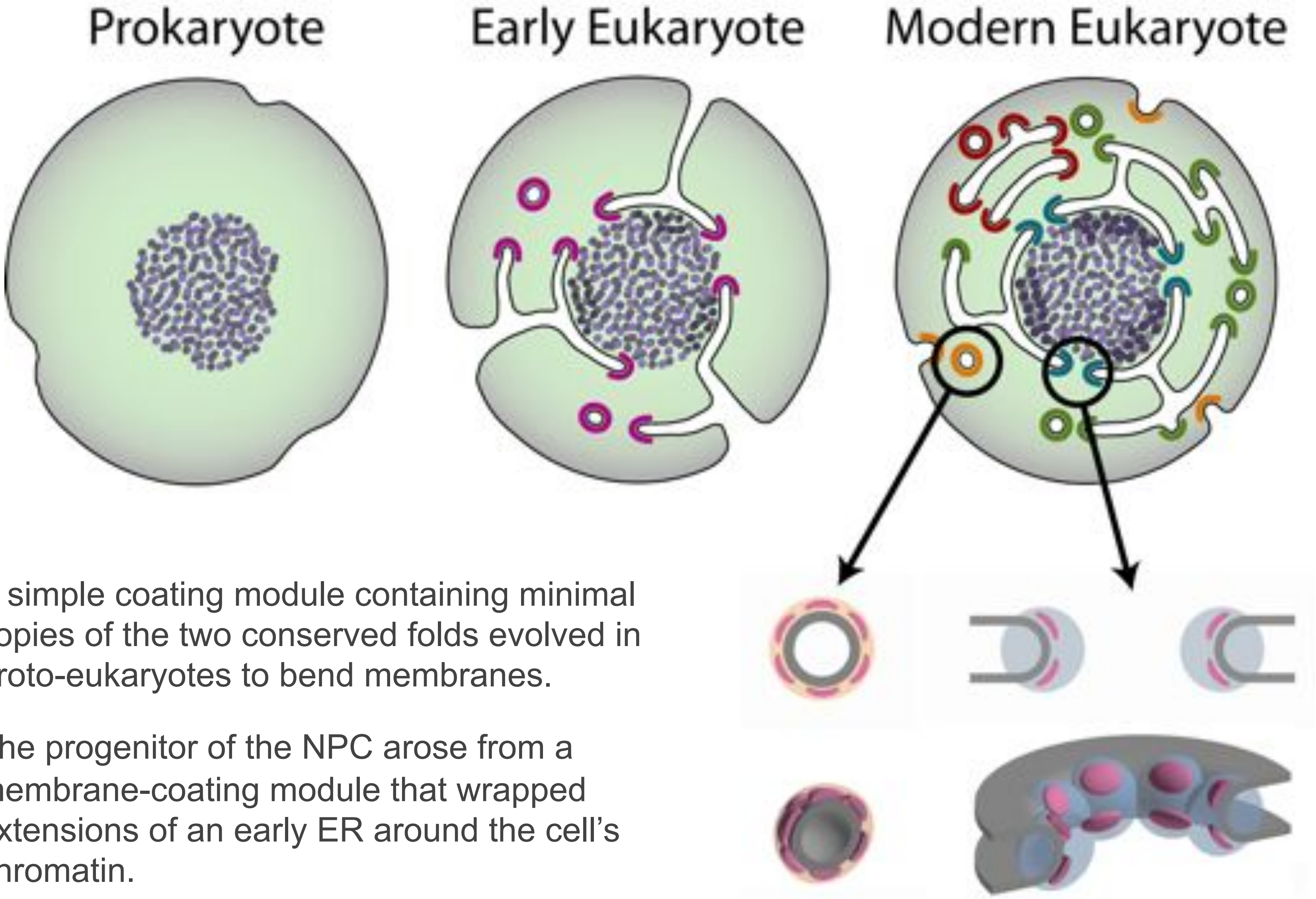
Nup 84 complex

1 Nup192, 2 Nup188, 3 Nup170, 4 Nup157, 5 Nup133,
6 Nup120, 7 Nup85, 8 Nup84, 9 Nup145C, 10 Seh1, 11 Sec13

Alber et al. The molecular architecture of the nuclear pore complex. Nature (2007) vol. 450 (7170) pp. 695-701

A Common Evolutionary Origin for Nuclear Pore Complexes and Coated Vesicles?

The proto-coatomer hypothesis

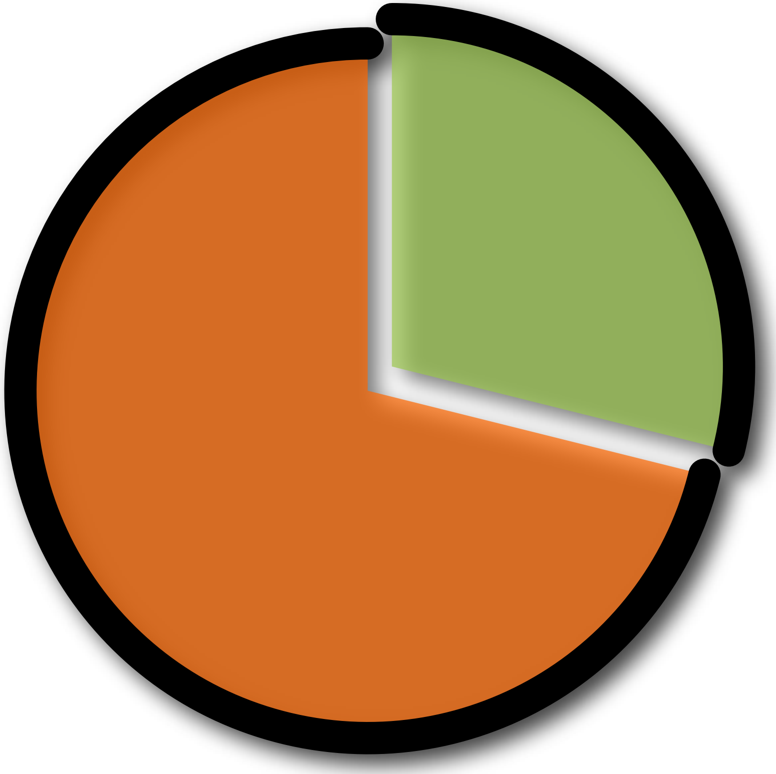


A simple coating module containing minimal copies of the two conserved folds evolved in proto-eukaryotes to bend membranes.

The progenitor of the NPC arose from a membrane-coating module that wrapped extensions of an early ER around the cell's chromatin.

Tropical Disease Initiative (TDI)

Predicting binding sites in protein structure models.

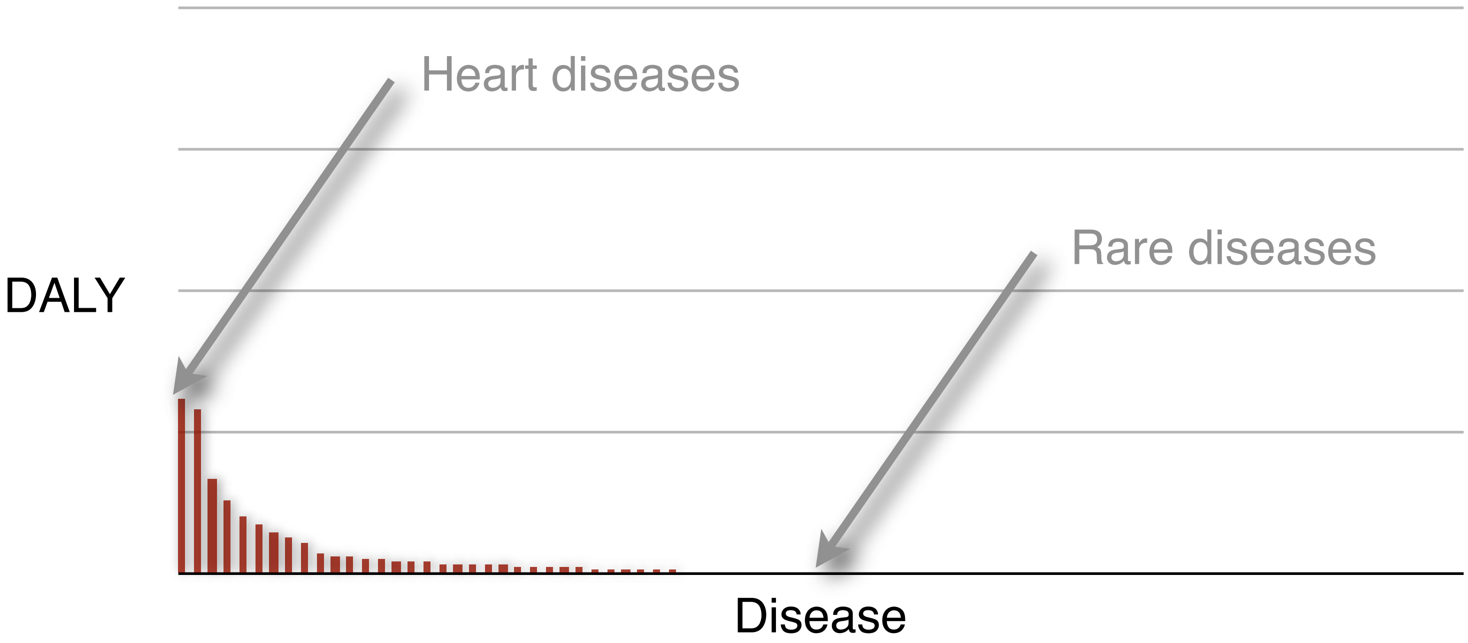


http://www.tropicaldisease.org

Need is High in the Tail

🟥 DALY Burden Per Disease in Developed Countries

🟩 DALY Burden Per Disease in Developing Countries



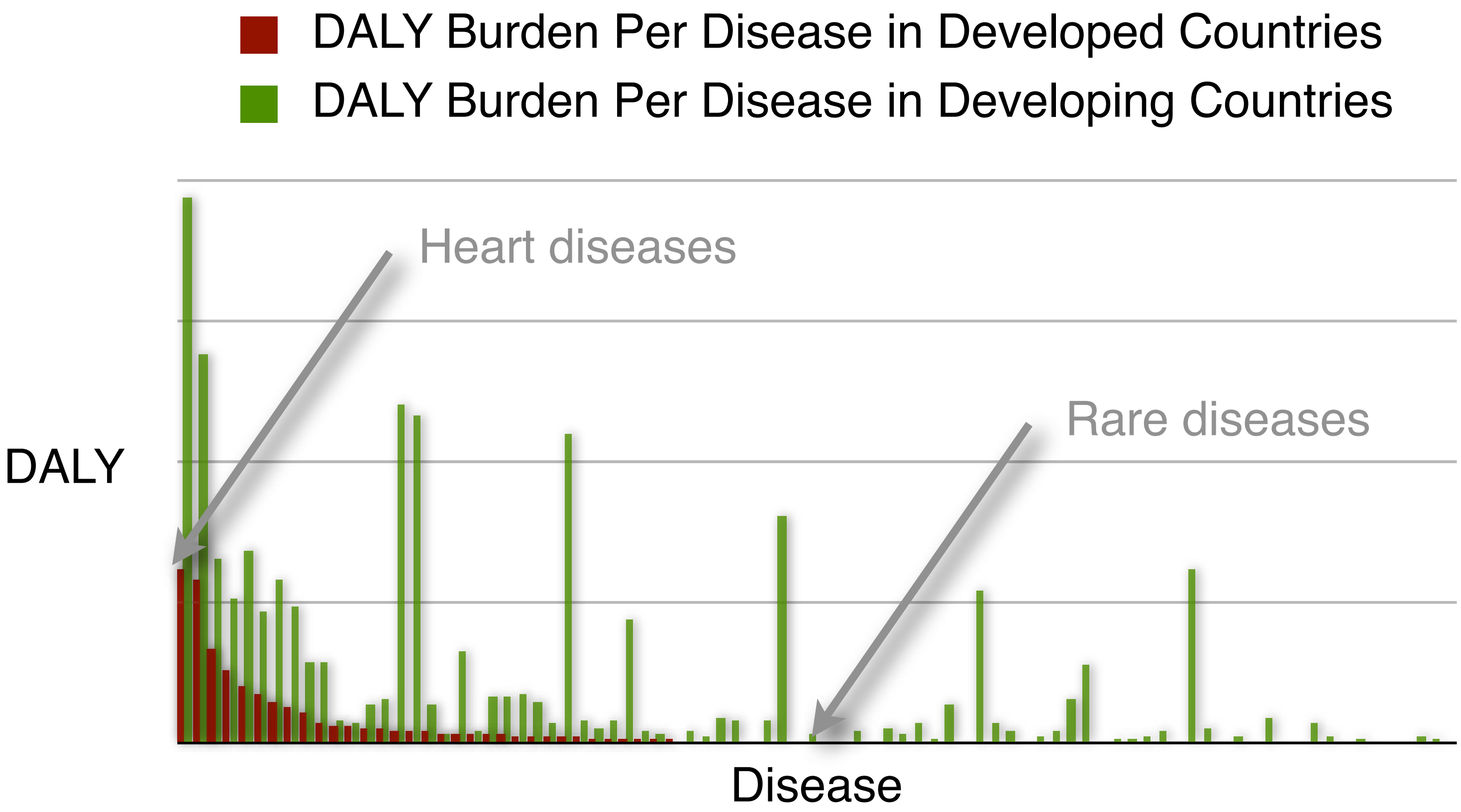
Disease data taken from WHO, *World Health Report 2004*

DALY - Disability adjusted life years

DALY is not a perfect measure of market size, but is certainly a good measure for importance.

DALYs for a disease are the sum of the years of life lost due to premature mortality (YLL) in the population and the years lost due to disability (YLD) for incident cases of the health condition. The DALY is a health gap measure that extends the concept of potential years of life lost due to premature death (PYLL) to include equivalent years of 'healthy' life lost in states of less than full health, broadly termed disability. One DALY represents the loss of one year of equivalent full health.

Need is High in the Tail



Disease data taken from WHO, *World Health Report 2004*

DALY - Disability adjusted life years

DALY is not a perfect measure of market size, but is certainly a good measure for importance.

DALYs for a disease are the sum of the years of life lost due to premature mortality (YLL) in the population and the years lost due to disability (YLD) for incident cases of the health condition. The DALY is a health gap measure that extends the concept of potential years of life lost due to premature death (PYLL) to include equivalent years of 'healthy' life lost in states of less than full health, broadly termed disability. One DALY represents the loss of one year of equivalent full health.

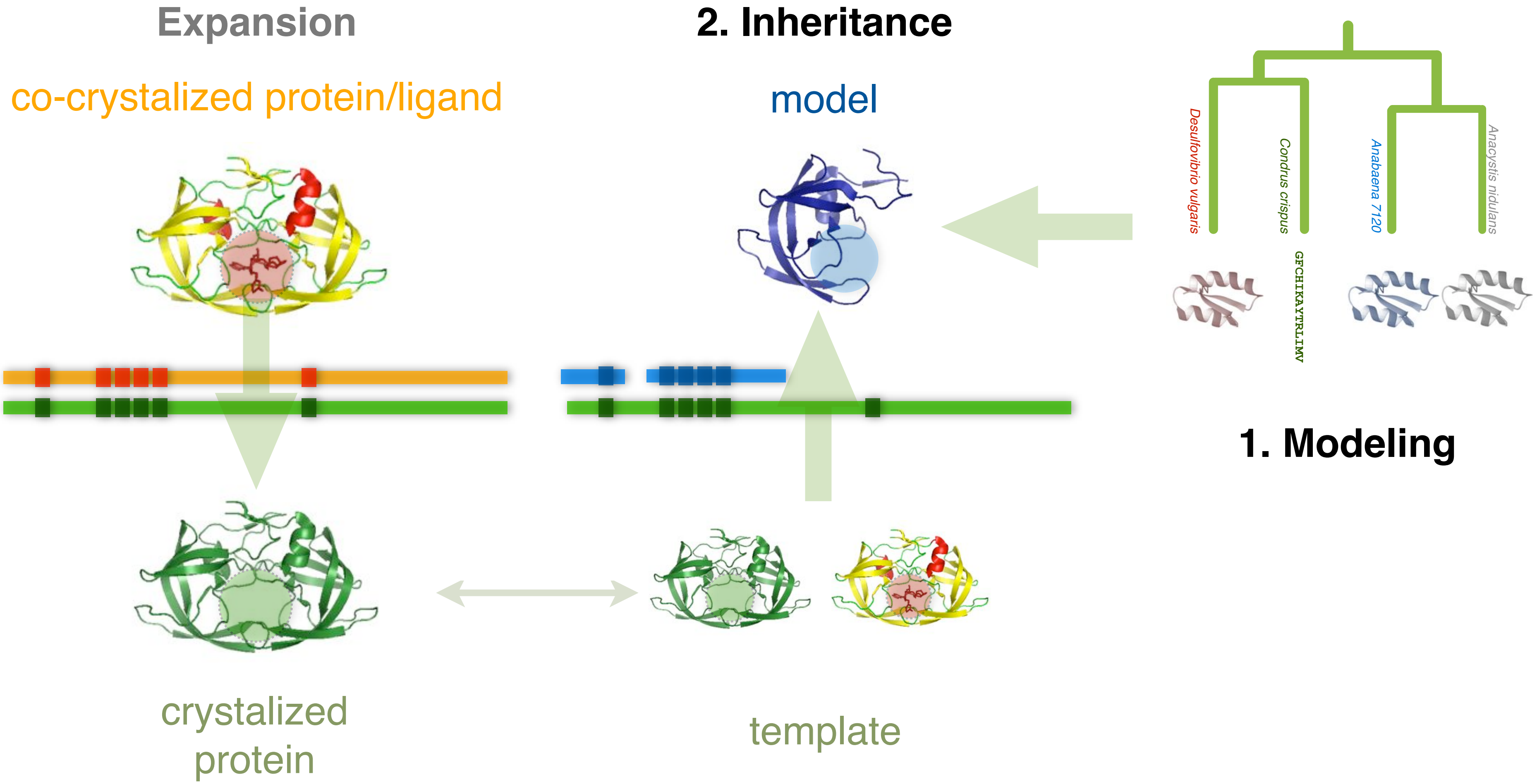
“Unprofitable” Diseases and Global DALY (in 1000’s)

Malaria*	**46,486**
Tetanus	7,074
Lymphatic filariasis*	**5,777**
Syphilis	4,200
Trachoma	2,329
Leishmaniasis*	**2,090**
Ascariasis	1,817
Schistosomiasis*	**1,702**
Trypanosomiasis*	**1,525**

Trichuriasis	1,006
Japanese encephalitis	709
Chagas Disease*	**667**
Dengue*	**616**
Onchocerciasis*	**484**
Leprosy*	**199**
Diphtheria	185
Poliomyelitise	151
Hookworm disease	59

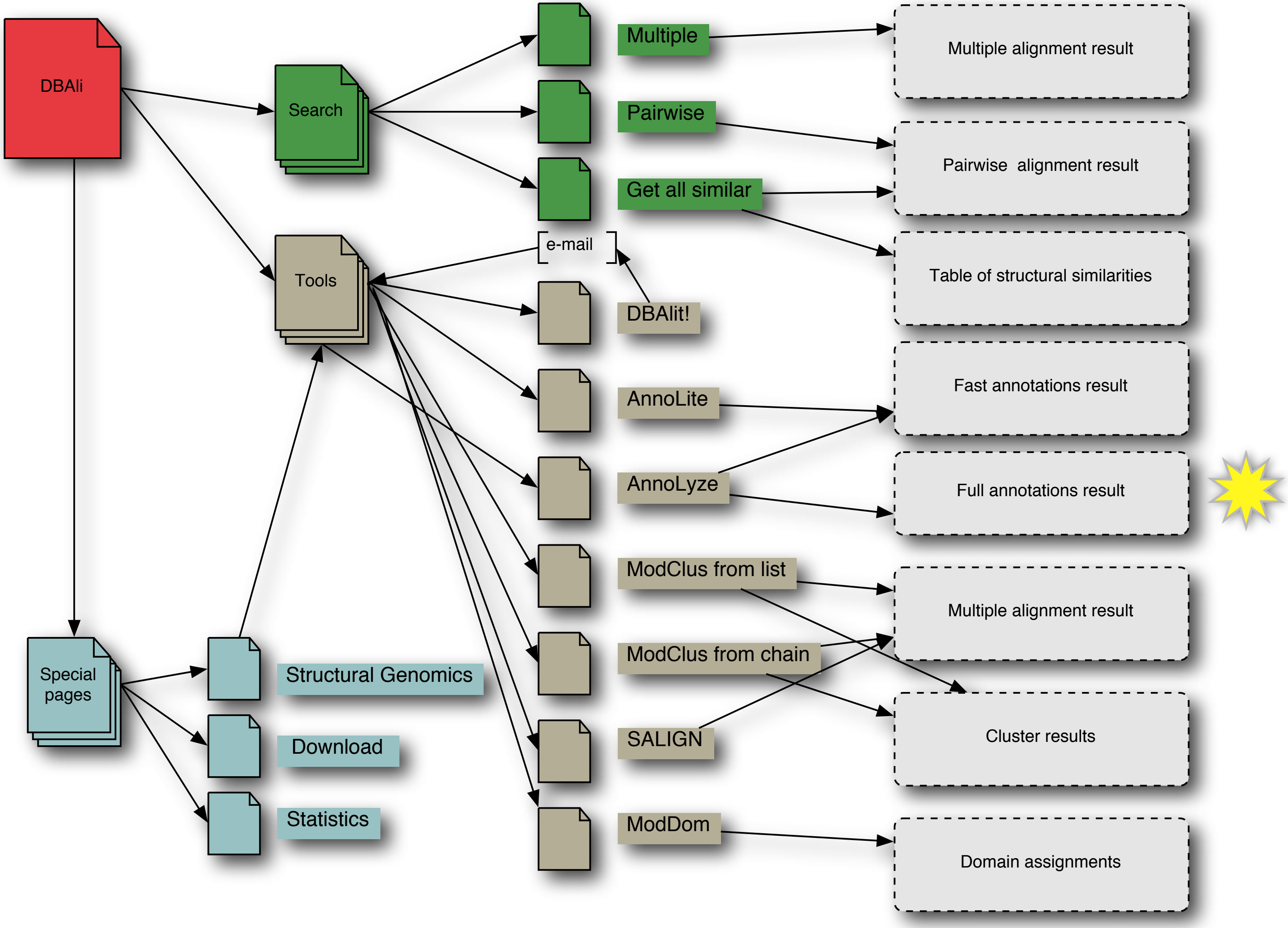
Disease data taken from WHO, *World Health Report 2004*
DALY - Disability adjusted life year in 1000’s.
* Officially listed in the WHO Tropical Disease Research disease portfolio.

Comparative docking



DBAli$_{\text{v2.0}}$ database

http://www.dbali.org



DBAli

Search

Multiple

Pairwise

Get all similar

Tools

e-mail

DBAlit!

AnnoLite

AnnoLyze

ModClus from list

ModClus from chain

SALIGN

ModDom

Special pages

Structural Genomics

Download

Statistics

Multiple alignment result

Pairwise alignment result

Table of structural similarities

Fast annotations result

Full annotations result

Multiple alignment result

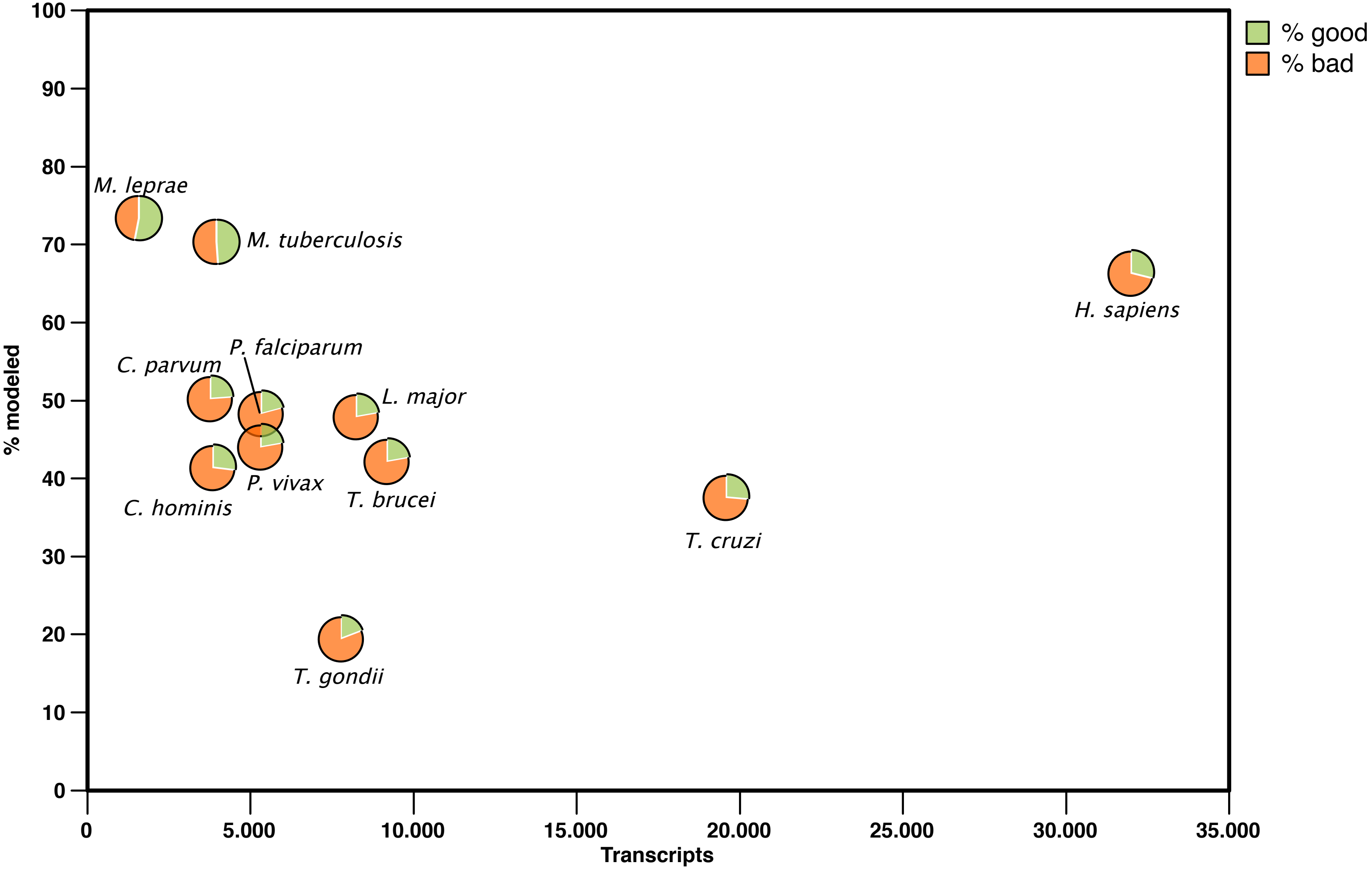
Cluster results

Domain assignments

Marti-Renom et al. BMC Bioinformatics (2007) Volume 8. Suppl S4

Modeling Genomes

data from models generated by ModPipe (Eswar, Pieper & Sali)



A good model has MPQS of 1.0 or higher

Summary table

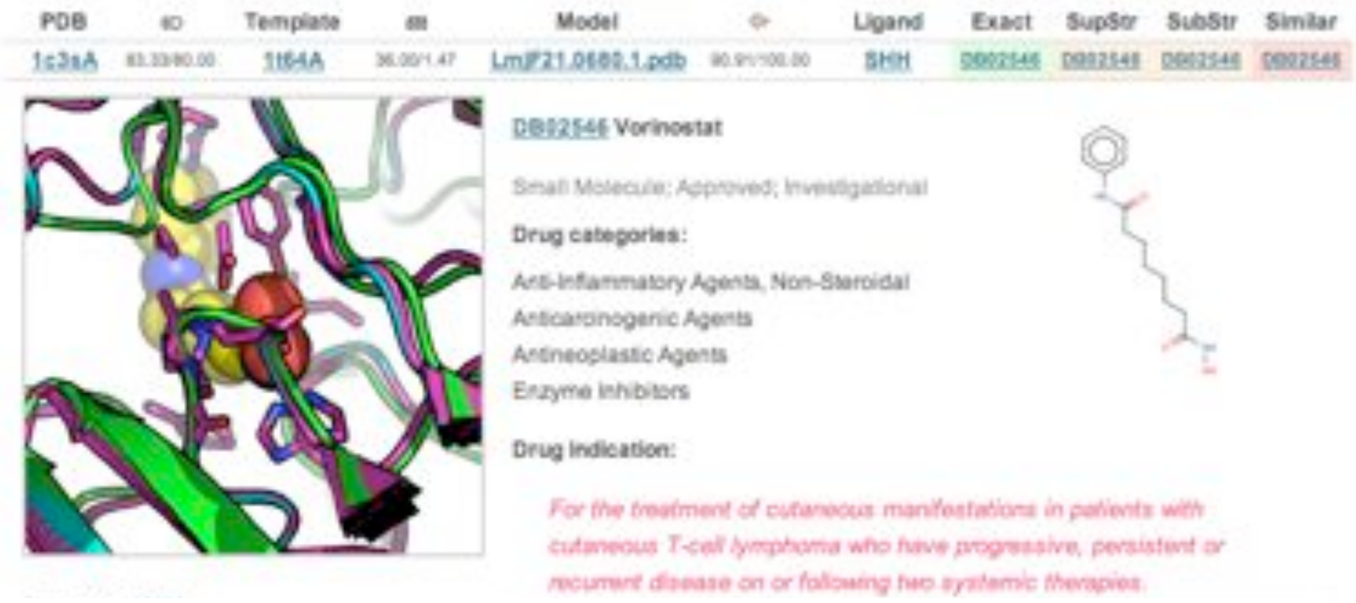
models with inherited ligands

29,271 targets with good models, 297 inherited a ligand/substance similar to a known drug in DrugBank

	Transcripts	Modeled targets	Selected models	Inherited ligands	Similar to a drug	Drugs
C. hominis	3,886	1,614	666	197	20	13
C. parvum	3,806	1,918	742	232	24	13
L. major	8,274	3,975	1,409	478	43	20
M. leprae	1,605	1,178	893	310	25	6
M. tuberculosis	3,991	2,808	1,608	365	30	10
P. falciparum	5,363	2,599	818	284	28	13
P. vivax	5,342	2,359	822	268	24	13
T. brucei	7,793	1,530	300	138	13	6
T. cruzi	19,607	7,390	3,070	769	51	28
T. gondii	9,210	3,900	1,386	458	39	21
TOTAL	**68,877**	**29,271**	**11,714**	**3,499**	**297**	**143**

L. major Histone deacetylase 2 + Vorinostat

Template 1t64A a human HDAC8 protein.



PDB	ID	Template	ID	Model		Ligand	Exact	SupStr	SubStr	Similar
1c3sA	83.33/90.00	1t64A	36.00/1.47	LmjF21.0680.1.pdb	90.91/100.00	SHH	DB02546	DB02546	DB02546	DB02546

DB02546 Vorinostat

Small Molecule; Approved; Investigational

Drug categories:

Anti-Inflammatory Agents, Non-Steroidal

Anticarcinogenic Agents

Antineoplastic Agents

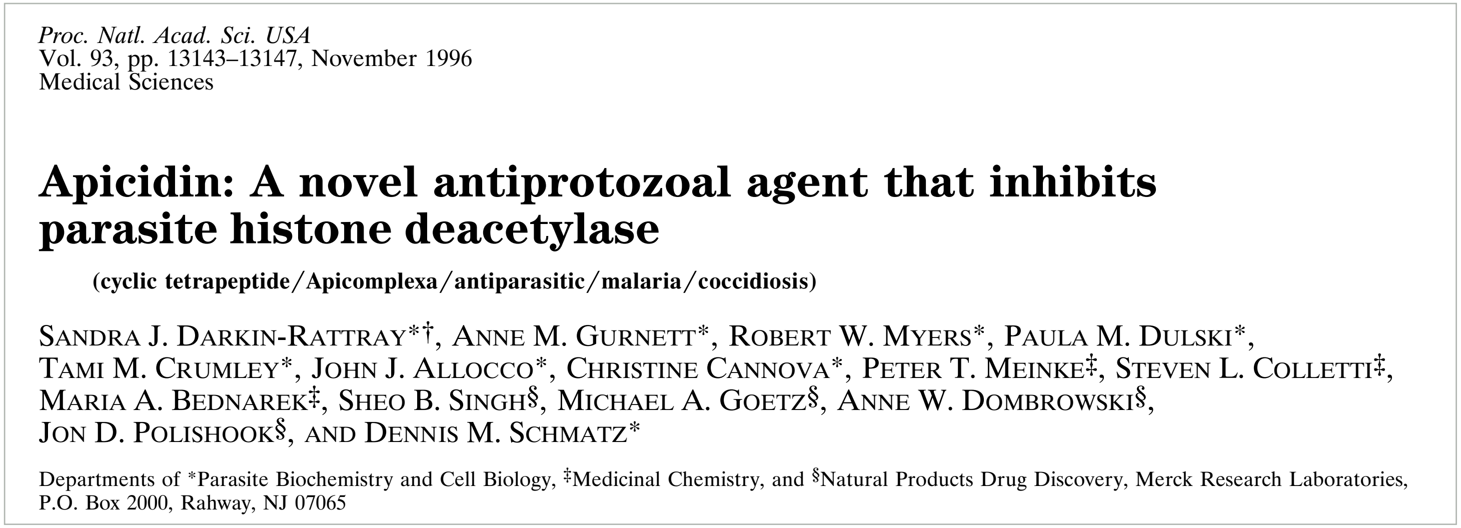
Enzyme Inhibitors

Drug indication:

For the treatment of cutaneous manifestations in patients with cutaneous T-cell lymphoma who have progressive, persistent or recurrent disease on or following two systemic therapies.

L. major Histone deacetylase 2 + Vorinostat

Literature



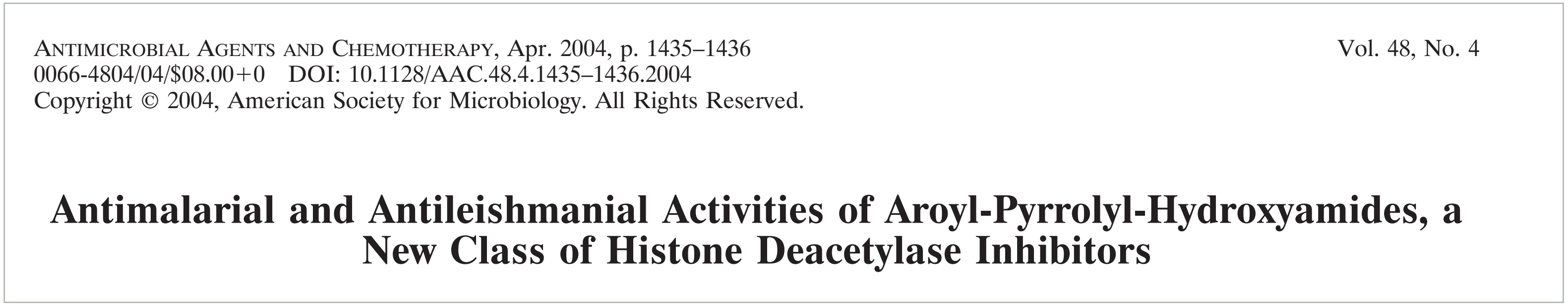
Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 13143–13147, November 1996
Medical Sciences

Apicidin: A novel antiprotozoal agent that inhibits parasite histone deacetylase

(cyclic tetrapeptide/Apicomplexa/antiparasitic/malaria/coccidiosis)

SANDRA J. DARKIN-RATTRAY[*†], ANNE M. GURNETT[*], ROBERT W. MYERS[*], PAULA M. DULSKI[*], TAMI M. CRUMLEY[*], JOHN J. ALLOCCO[*], CHRISTINE CANNOVA[*], PETER T. MEINKE[‡], STEVEN L. COLLETTI[‡], MARIA A. BEDNAREK[‡], SHEO B. SINGH[§], MICHAEL A. GOETZ[§], ANNE W. DOMBROWSKI[§], JON D. POLISHOOK[§], AND DENNIS M. SCHMATZ[*]

Departments of [*]Parasite Biochemistry and Cell Biology, [‡]Medicinal Chemistry, and [§]Natural Products Drug Discovery, Merck Research Laboratories, P.O. Box 2000, Rahway, NJ 07065

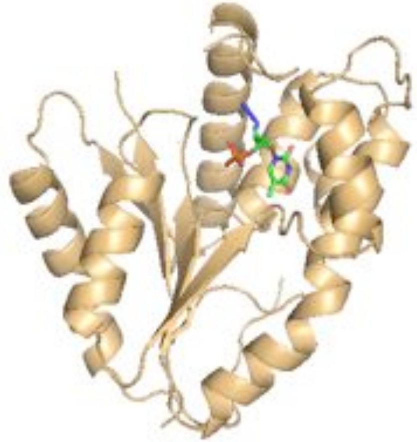


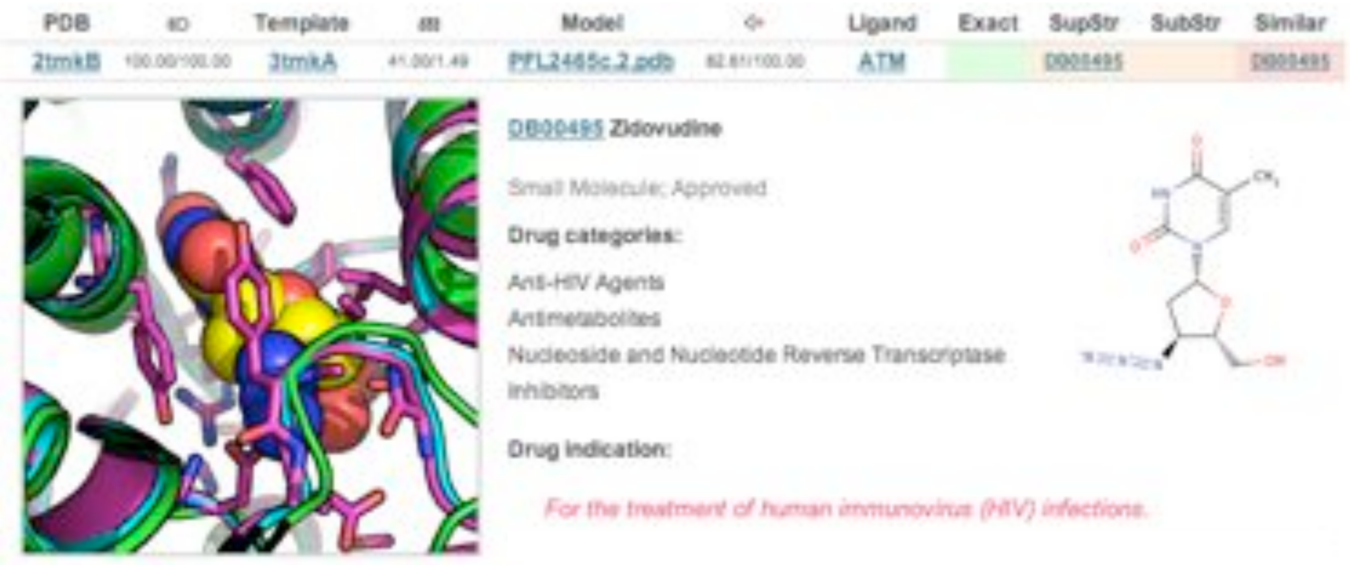
ANTIMICROBIAL AGENTS AND CHEMOTHERAPY, Apr. 2004, p. 1435–1436 Vol. 48, No. 4
0066-4804/04/$08.00+0 DOI: 10.1128/AAC.48.4.1435–1436.2004
Copyright © 2004, American Society for Microbiology. All Rights Reserved.

Antimalarial and Antileishmanial Activities of Aroyl-Pyrrolyl-Hydroxyamides, a New Class of Histone Deacetylase Inhibitors

P. falciparum tymidylate kinase + zidovudine

Template 3tmkA a yeast tymidylate kinase.





PDB	ID	Template	[illegible]	Model	⇦	Ligand	Exact	SupStr	SubStr	Similar
2tmkB	100.00/100.00	3tmkA	41.00/1.49	PFL2465c.2.pdb	82.81/100.00	ATM	🟩	DB00495		DB00495

DB00495 Zidovudine

Small Molecule; Approved

Drug categories:

Anti-HIV Agents

Antimetabolites

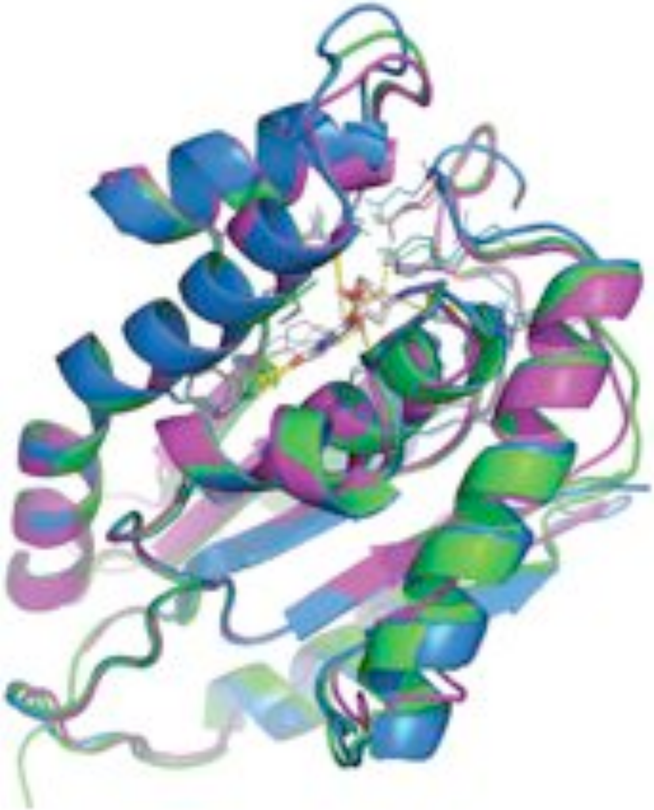
Nucleoside and Nucleotide Reverse Transcriptase Inhibitors

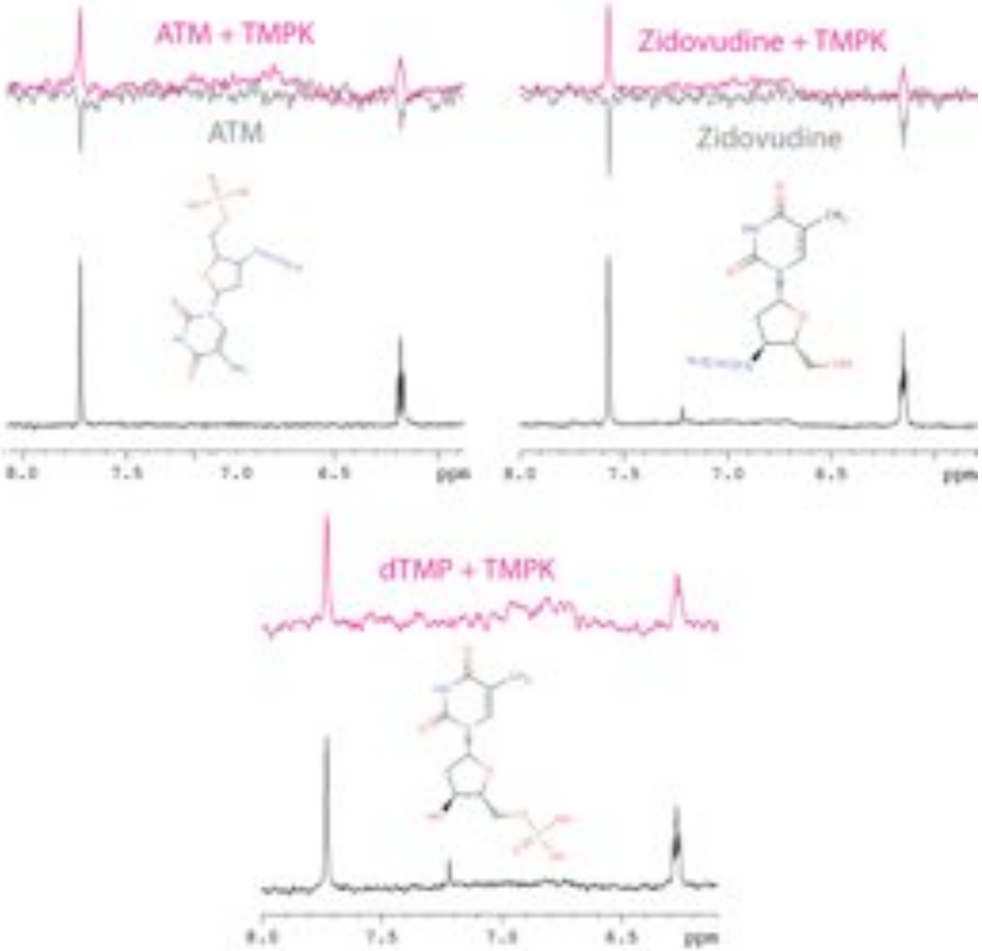
Drug indication:

For the treatment of human immunovirus (HIV) infections.

P. falciparum tymydilate kinase + zidovudine

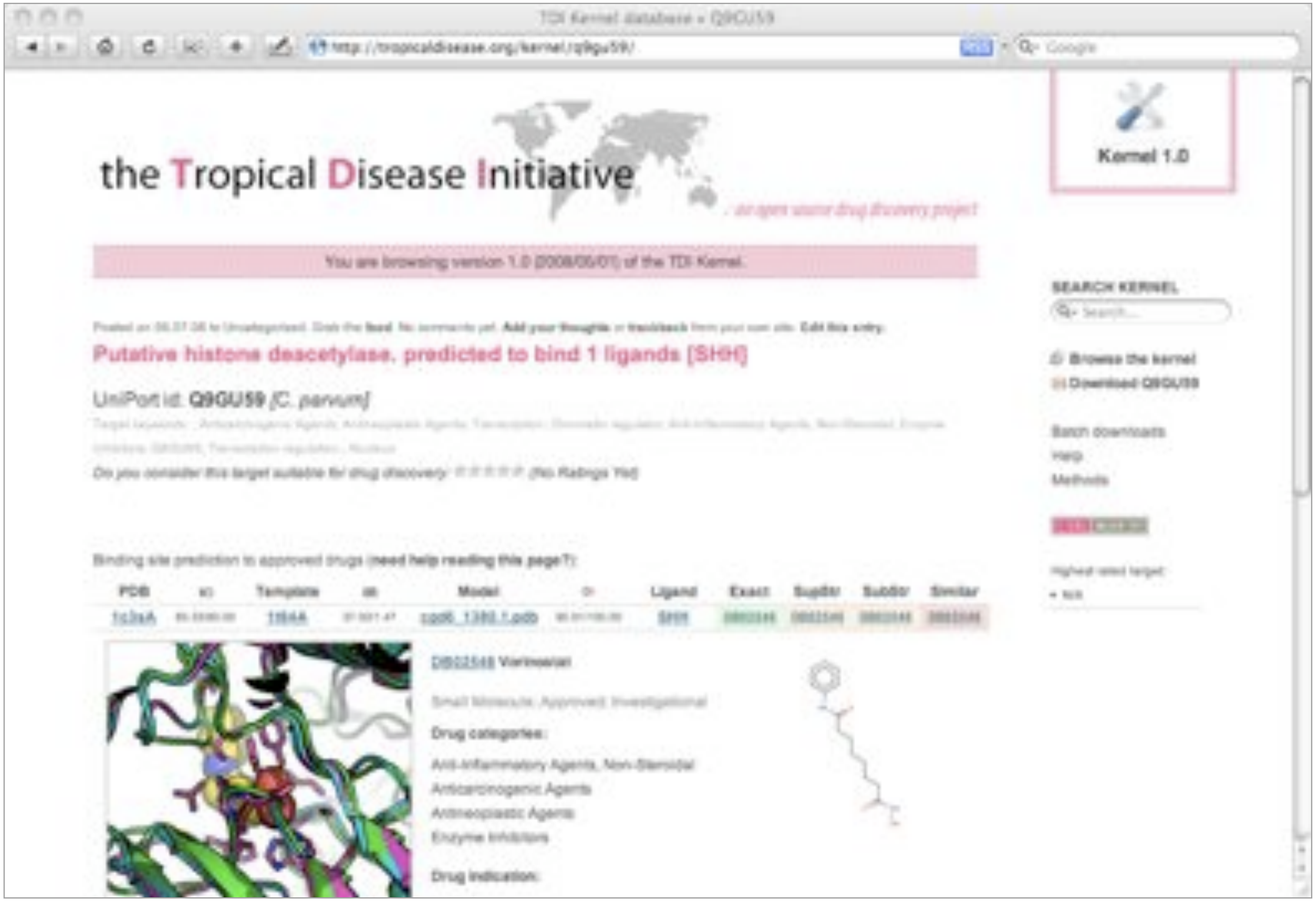
NMR Water-LOGSY experiments





TDI’s kernel

http://tropicaldisease.org/kernel



TDI Kernel database » Q9GU59
http://tropicaldisease.org/kernel/q9gu59/
RSS
Google

the Tropical Disease Initiative

. an open source drug discovery project

You are browsing version 1.0 (2008/06/01) of the TDI Kernel.

Posted on 06.07.08 to Uncategorized. Grab the feed. No comments yet. Add your thoughts or trackback from your own site. Edit this entry.

Putative histone deacetylase, predicted to bind 1 ligands [SHH]

UniPort id: Q9GU59 *[C. parvum]*

Target keywords: , Anticarcinogenic Agents, Antineoplastic Agents, Transcription, Chromatin regulator, Anti-Inflammatory Agents, Non-Steroidal, Enzyme Inhibitors, [illegible], Transcription regulation, Nucleus

Do you consider this target suitable for drug discovery: ☆☆☆☆☆ (No Ratings Yet)

Binding site prediction to approved drugs (need help reading this page?):

PDB	[illegible]	Template	[illegible]	Model	[illegible]	Ligand	Exact	SupStr	SubStr	Similar
1c3sA	[illegible]	1t64A	[illegible]	cpd6_1380.1.pdb	[illegible]	SHH	DB02546	DB02546	DB02546	DB02546

DB02546 Vorinostat

Small Molecule; Approved; Investigational

Drug categories:

Anti-Inflammatory Agents, Non-Steroidal
Anticarcinogenic Agents
Antineoplastic Agents
Enzyme Inhibitors

Drug indication:

Acknowledgments

Structural Genomics Unit (CIPF)
Marc A. Marti-Renom
Emidio Capriotti
Davide Baù

Comparative Genomics Unit (CIPF)
Hernán Dopazo
Leo Arbiza
Francisco García

Functional Genomics Unit (CIPF)
Joaquín Dopazo
Fátima Al-Shahrour
José Carbonell
Ignacio Medina
David Montaner
Joaquin Tárraga
Ana Conesa
Toni Gabaldón
Eva Alloza
Lucía Conde
Stefan Goetz
Jaime Huerta Cepas
Marina Marcet
Pablo Minguez
Jordi Burguet Castell

FUNDING
Prince Felipe Research Center
Marie Curie Reintegration Grant
STREP EU Grant
Ministerio de Educación y Ciencia (Spain)

Tropical Disease Initiative
Stephen Maurer (UC Berkeley)
Arti Rai (Duke U)
Andrej Sali (UCSF)
Ginger Taylor (TSL)
Barri Bunin (CDD)

STRUCTURAL GENOMICS
Stephen Burley (SGX)
John Kuriyan (UCB)
NY-SGXRC

MAMMOTH
Angel R. Ortiz

BIOLOGY
Jeff Friedman (RU)
James Hudsped (RU)
Partho Ghosh (UCSD)
Alvaro Monteiro (Cornell U)
Stephen Krilis (St.George H)

FUNCTIONAL ANNOTATION
Fatima Al-Shahrour
Joaquin Dopazo

NMR STRUCTURE BIOLOGY
Antonio Pineda-Lucena
Leticia Ortí
Rodrigo Carbajo

COMPARATIVE MODELING
Andrej Sali
Narayanan Eswar
Ursula Pieper

FUNCTIONAL ANNOTATION
Andrea Rossi
Fred Davis

http://bioinfo.cipf.es
http://sgu.bioinfo.cipf.es