

# Comparative Protein Structure Prediction



**Marc A. Marti-Renom**

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

Structural Genomics Unit  
Bioinformatics Department

Prince Felipe Research Center (CIPF), Valencia, Spain



**PRINCIPE FELIPE**  
CENTRO DE INVESTIGACION

# Program

Intro to comparative  
protein structure prediction

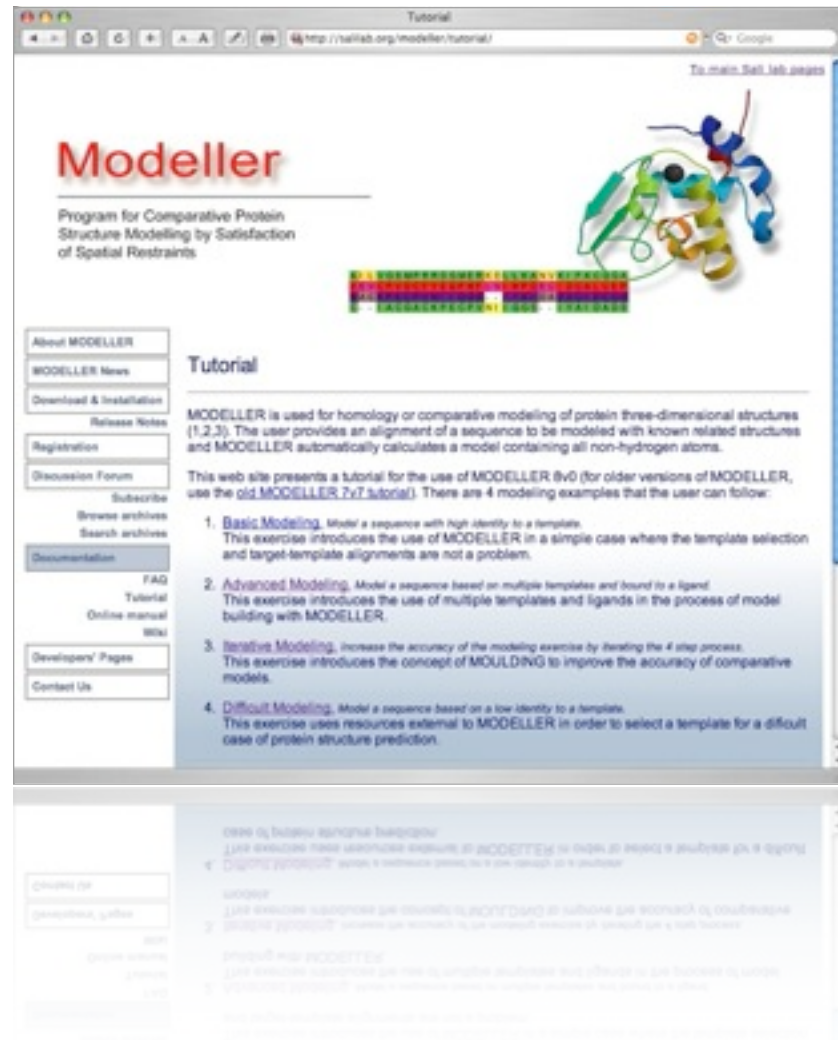
Template Search

Target – Template  
Alignment

Model Building

Model Evaluation

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



The screenshot shows the Modeller website's tutorial page. At the top, there's a navigation bar with links like 'About MODELLER', 'MODELLER News', 'Download & Installation', 'Registration', 'Discussion Forum', 'Browse archives', 'Search archives', 'Documentation', 'FAQ', 'Tutorial', 'Online manual', 'Bugs', 'Developers' Pages', and 'Contact Us'. The main content area is titled 'Modeller' and 'Tutorial'. It describes the program as 'Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints'. Below this, there's a 3D ribbon diagram of a protein structure. The tutorial text explains that MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2,3). It provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. The tutorial is divided into four sections: 1. Basic Modeling, 2. Advanced Modeling, 3. Iterative Modeling, and 4. Difficult Modeling. Each section describes a different modeling exercise.

**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 testing 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.

# Objective

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

# DISCLAIMER!

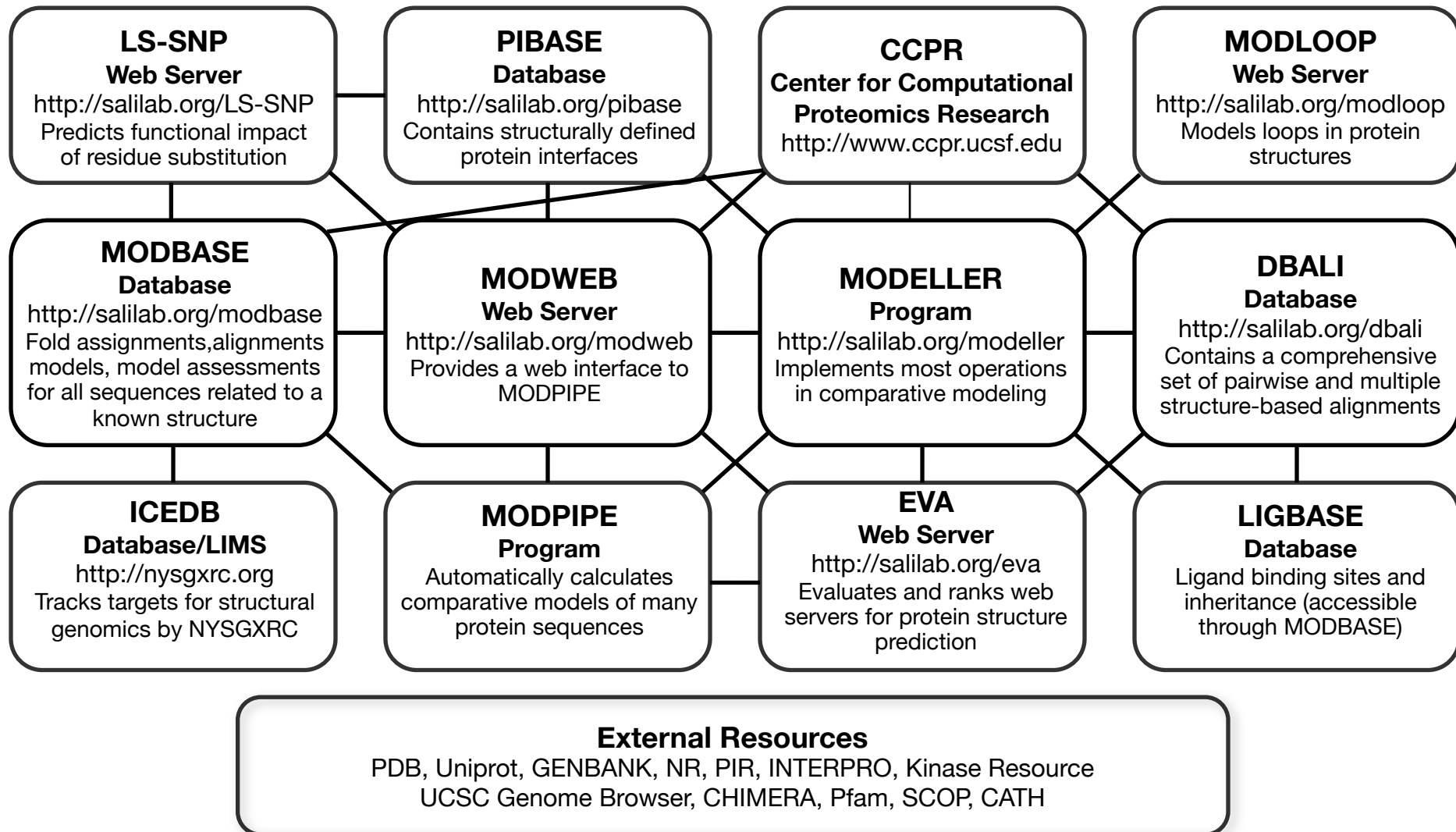
Name	Type <sup>a</sup>	World Wide Web address <sup>b</sup>
<b>DATABASES</b>		
CATH	S	<a href="http://www.biochem.ucl.ac.uk/bsm/cath/">http://www.biochem.ucl.ac.uk/bsm/cath/</a>
DBAII	S	<a href="http://www.sallab.org/DBAII/">http://www.sallab.org/DBAII/</a>
GenBank	S	<a href="http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html">http://www.ncbi.nlm.nih.gov/Genb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

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



# Programs, servers and databases

<http://salilab.org>



# 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

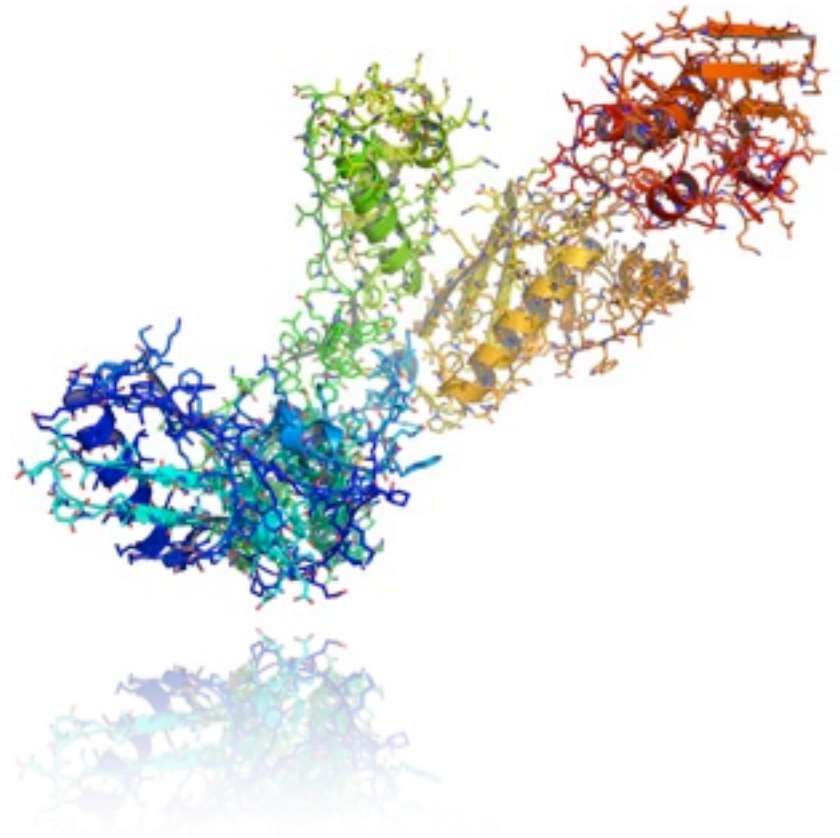
**Fold:** Three dimensional conformation of a protein sequence (usually at domain level).

**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:

Madhusudhan et al. The Proteomics Protocols Handbook. Ed. Walker. Humana Press Inc., Totowa, NJ. 831-860, 2005.

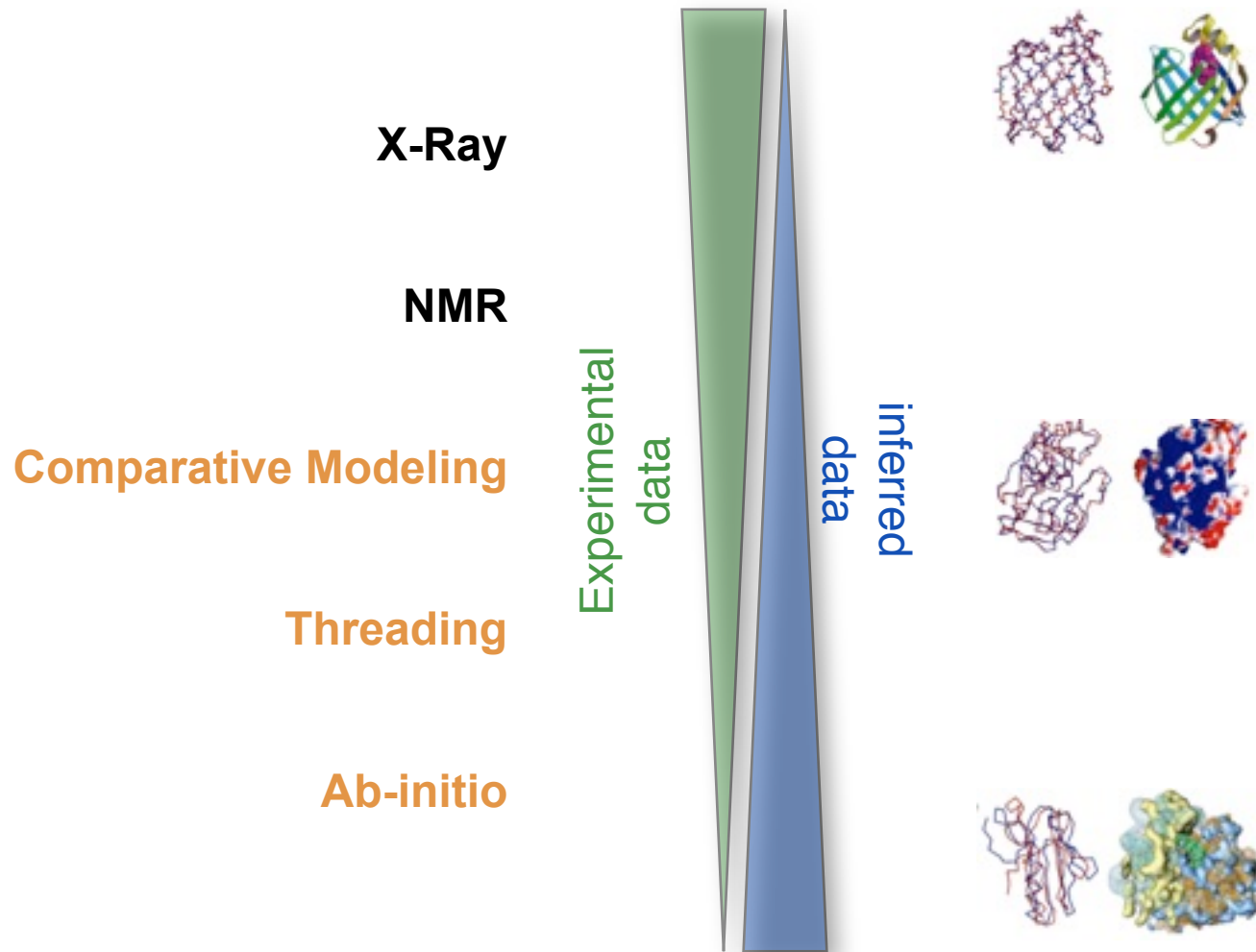
## MODELLER:

Sali & Blundell. J. Mol. Biol. 234, 779-815, 1993.  
Eswar, M. A. et al. Comparative Protein Structure Modeling With MODELLER.  
Current Protocols in Bioinformatics, John Wiley & Sons, Inc.,  
Supplement 15, 5.6.1-5.6.30, 2006.

## 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



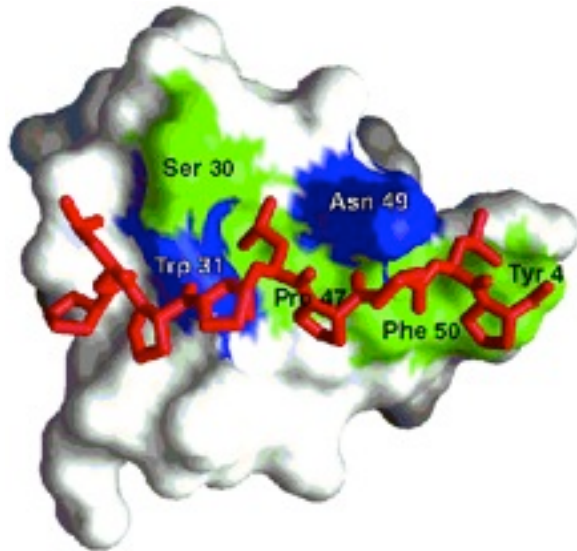
# Why is it useful to know the **structure** of a protein, not only its sequence?

- ◆ 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

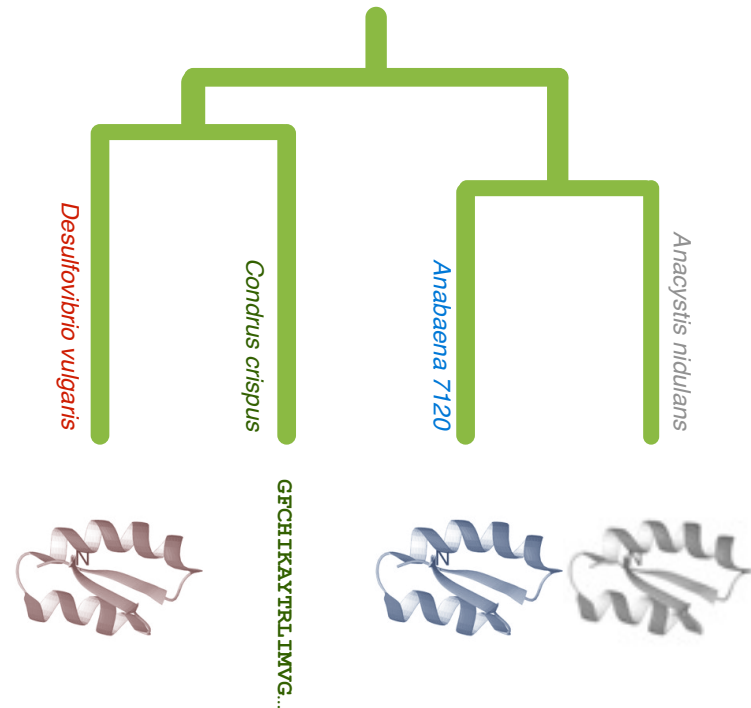
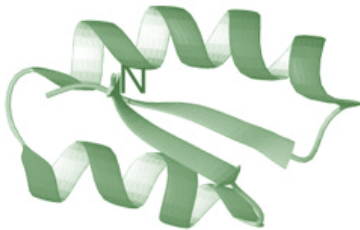


In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence.**

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

# Principles of protein structure

GFCHIKAYTRLIMVG...



Folding (physics)

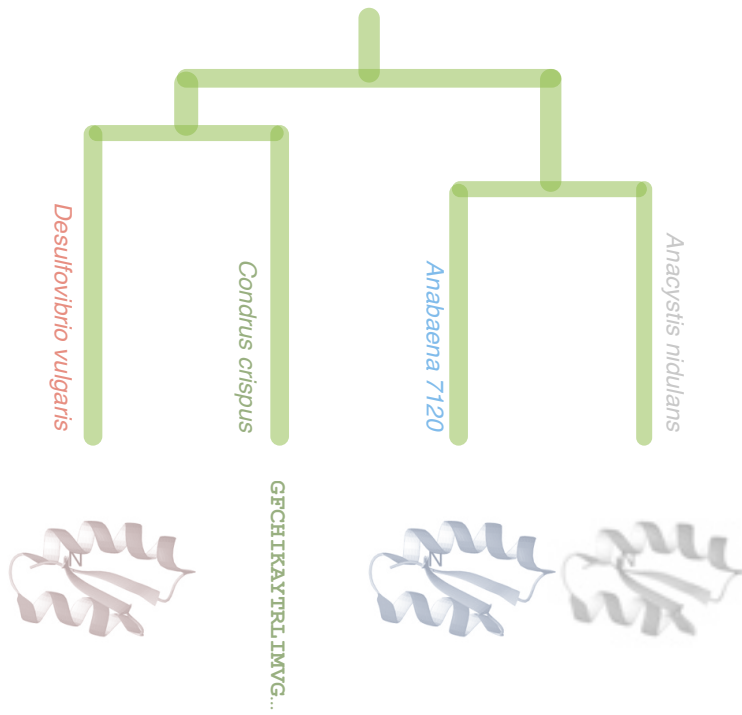
*Ab initio* prediction

Evolution (rules)

Threading  
Comparative Modeling

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

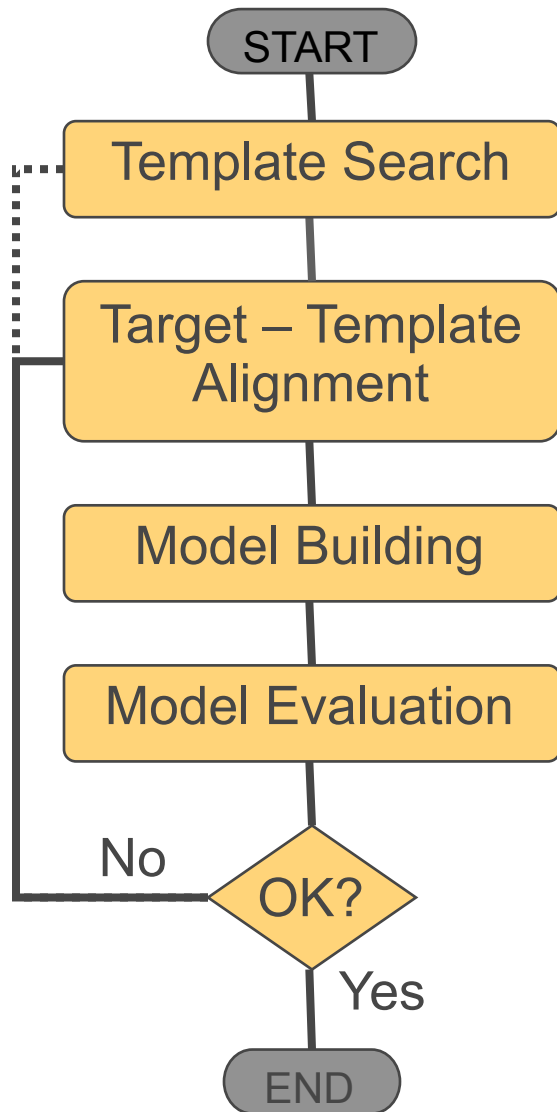




# MODELLER

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



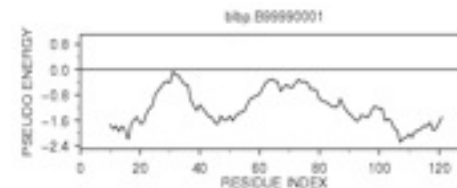
TARGET

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.

# Template Selection

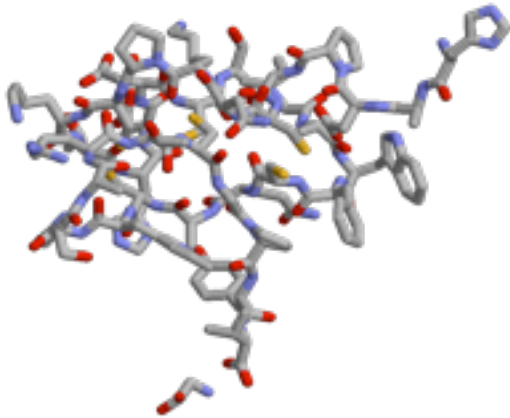
## “Structural Space”

# Structure-Structure alignments

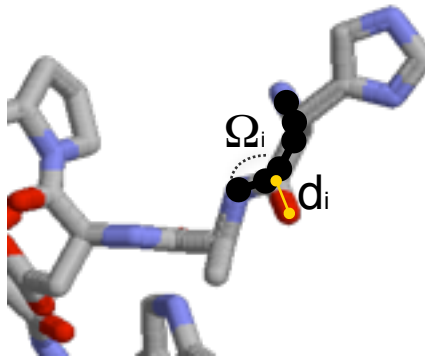
As any other bioinformatics problem...

- Representation
  - Scoring
  - Optimizer

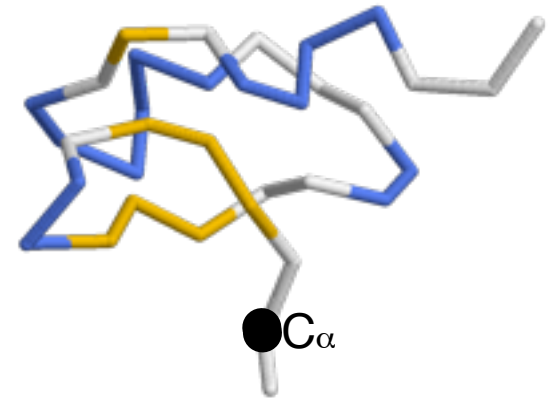
# Structures



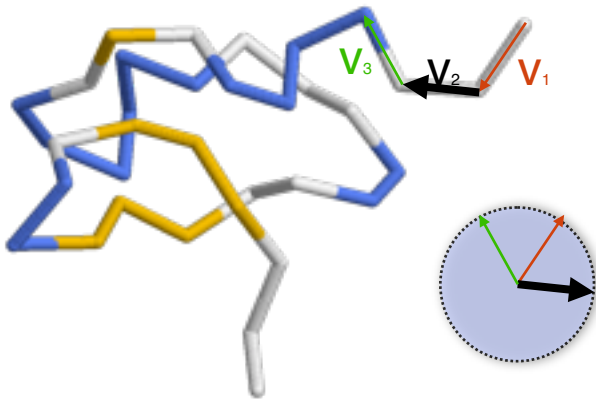
All atoms and coordinates



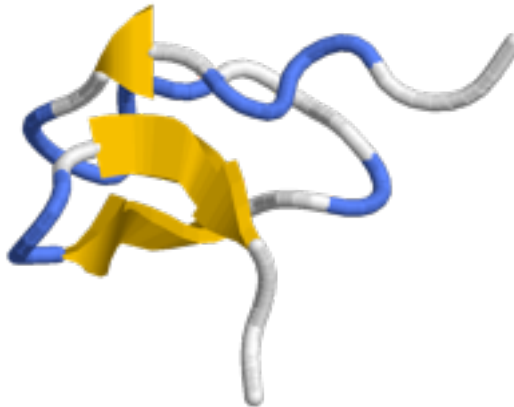
Dihedral space or distance space



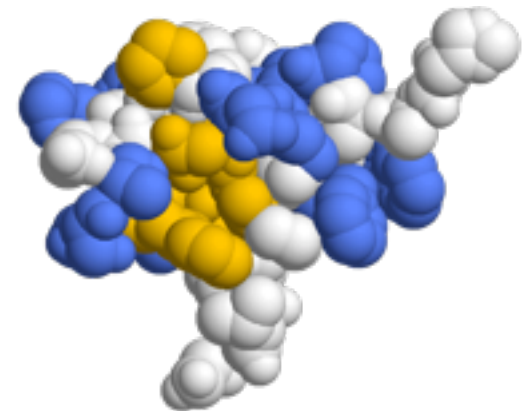
Reduced atom representation



Vector representation



Secondary Structure



Accessible surface (and others)

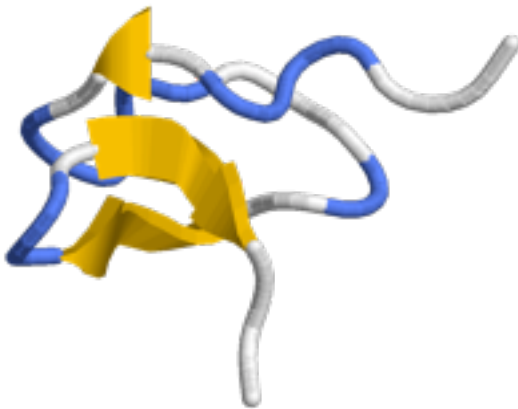
# Raw scores

	C	S	T	P	A	G	N	D	E	Q	R	K	L	M	I	V	F	Y	W
C	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
S	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
T	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
P	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
A	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
G	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
N	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
D	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
E	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Q	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1
R	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1
K	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1
L	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1
M	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1
I	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1
V	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1
F	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1
Y	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1
W	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0

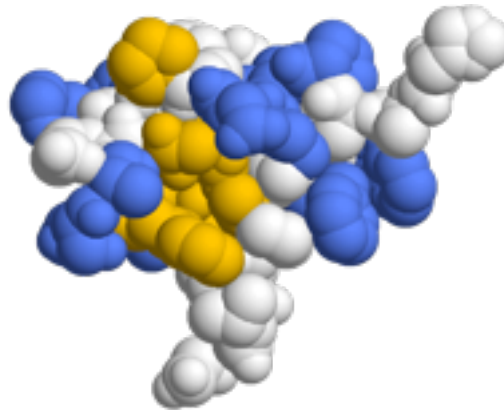
Aminoacid substitutions

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

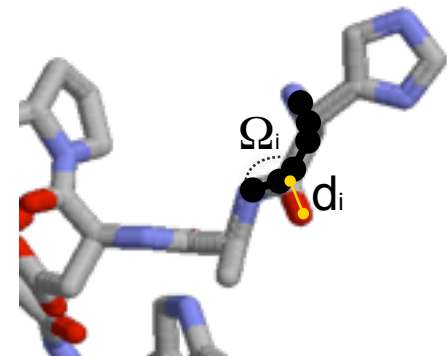
Root Mean Square Deviation



Secondary Structure (H,B,C)



Accessible surface (B,A [%])



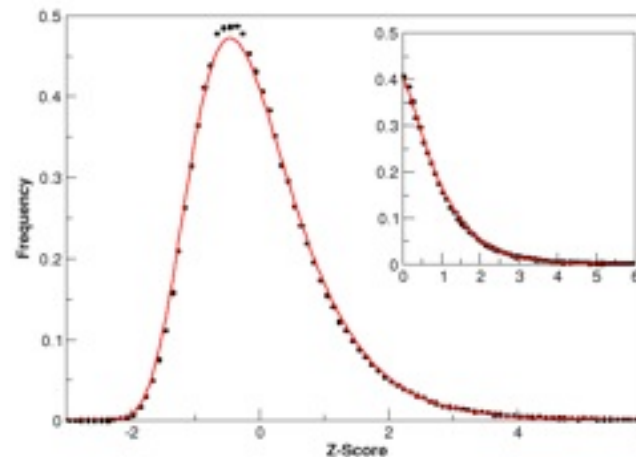
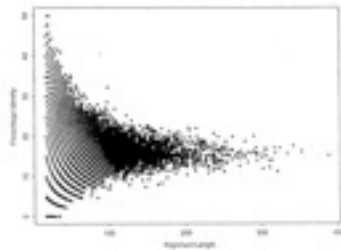
Angles or distances

## 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

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

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

*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	*	*	*	*	*

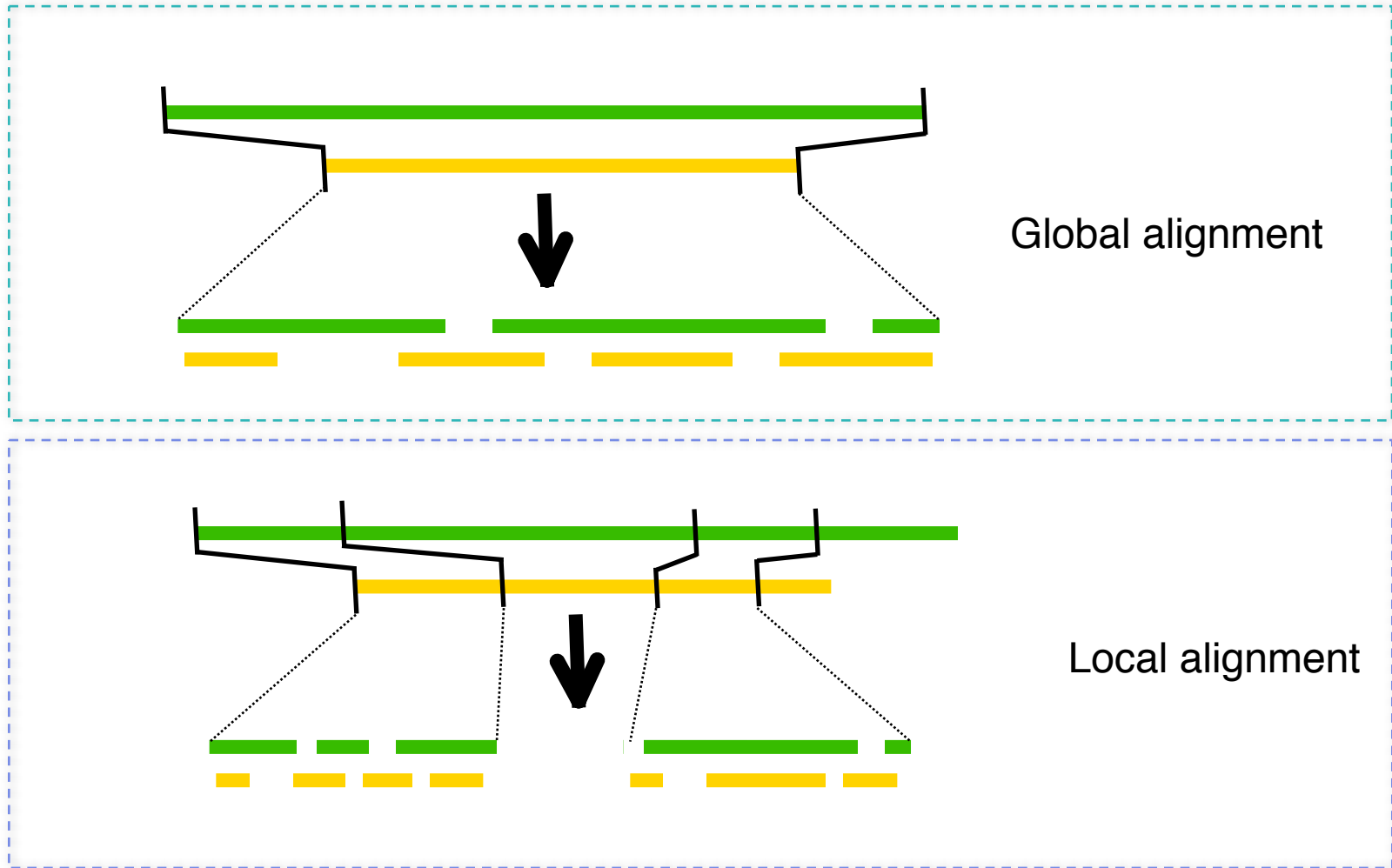
Best local alignment

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.



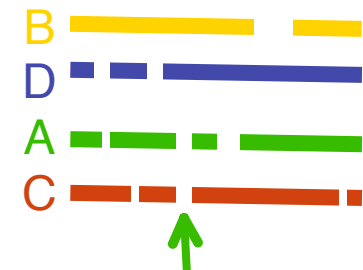
6 pairwise comparisons  
then cluster analysis

## Multiple alignments

Following the tree from step 1



Align B-D with A-C

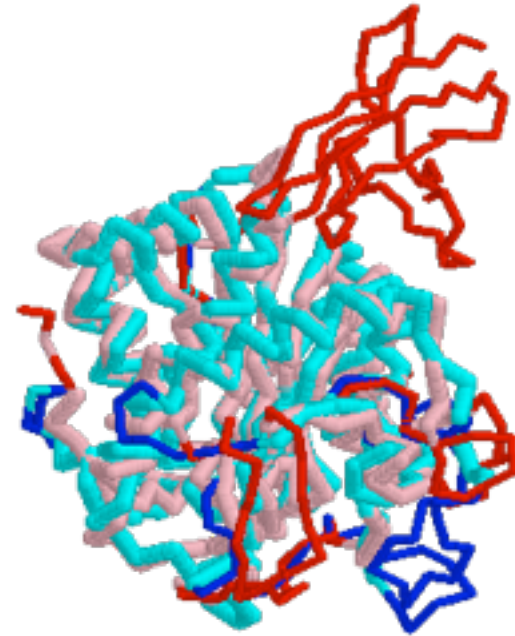


New gap in A-C to optimize  
its alignment with B-D

# Coverage .vs. Accuracy



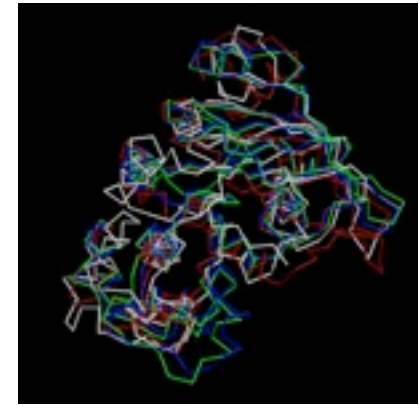
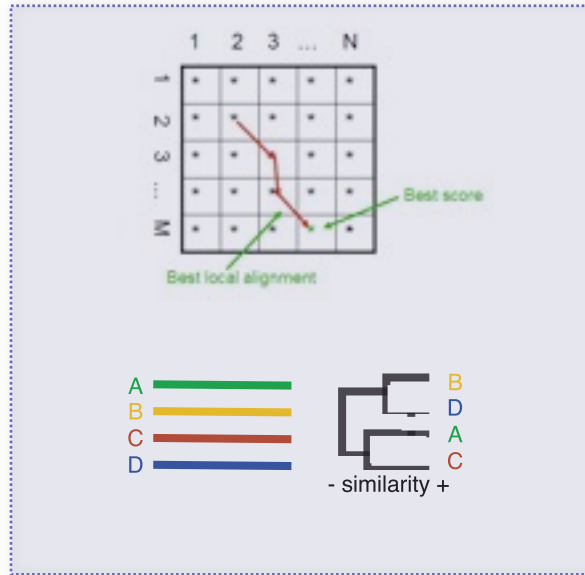
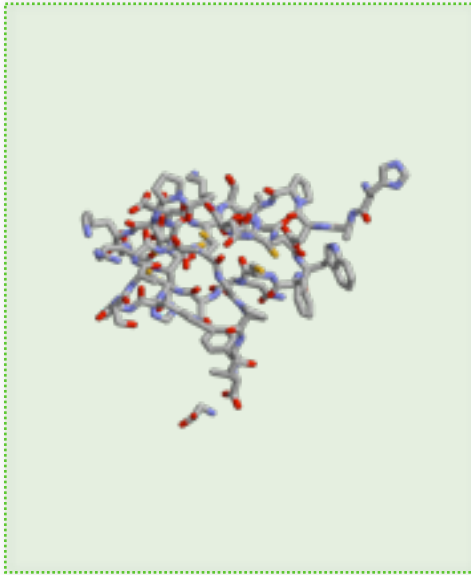
Coverage ~90%  $C\alpha$



Coverage ~75%  $C\alpha$

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

# Structural alignment by properties conservation (SALIGN-MODELLER)

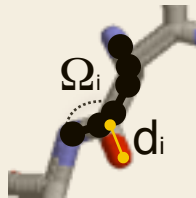


- ✓ Uses all available structural information
- ✓ Provides the optimal alignment

Computationally expensive



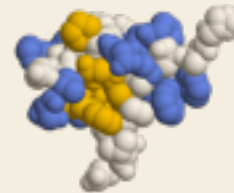
$R_{i,j}$



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



$S_{i,j}$



$B_{i,j}$

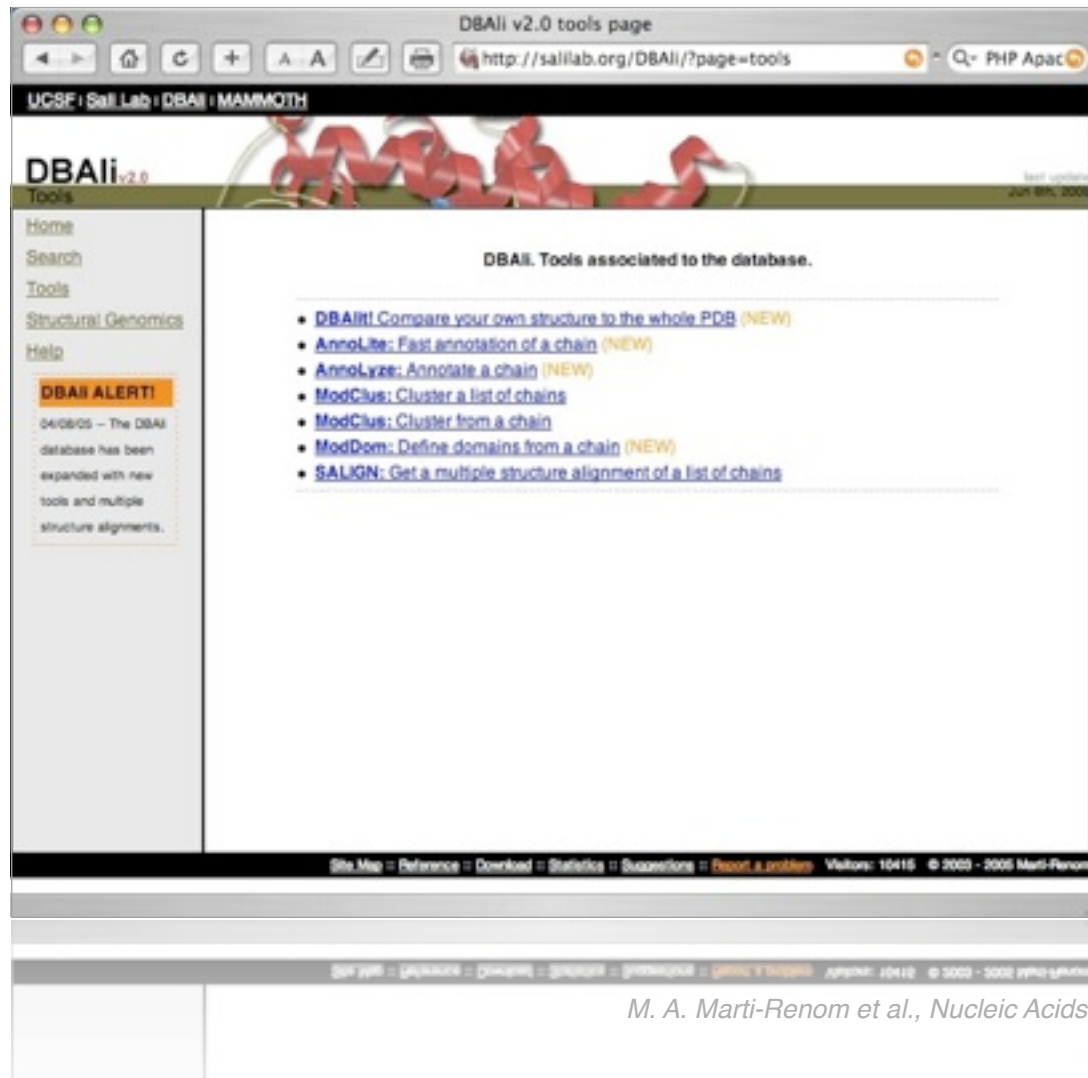
$$RMSD(x, y) = \sqrt{\left(\frac{1}{N}\right) \sum_{i=1}^N \|\mathbf{x}(i) - \mathbf{y}(i)\|^2}$$

$I_{i,j}$

*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)

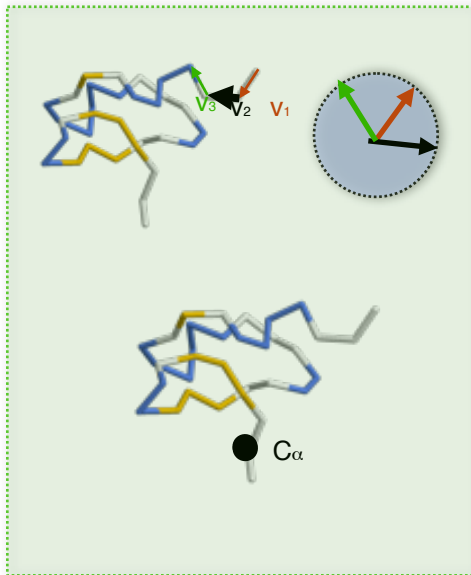
<http://dbali.org>



*M. A. Marti-Renom et al., Nucleic Acids Res 35, W393 (Jul 1, 2007)*



# Vector Alignment Search Tool (VAST)

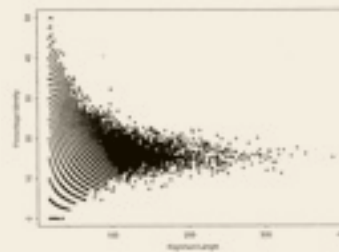


Graph theory search  
of similar SSE  
Refining by Monte Carlo  
at all atom resolution

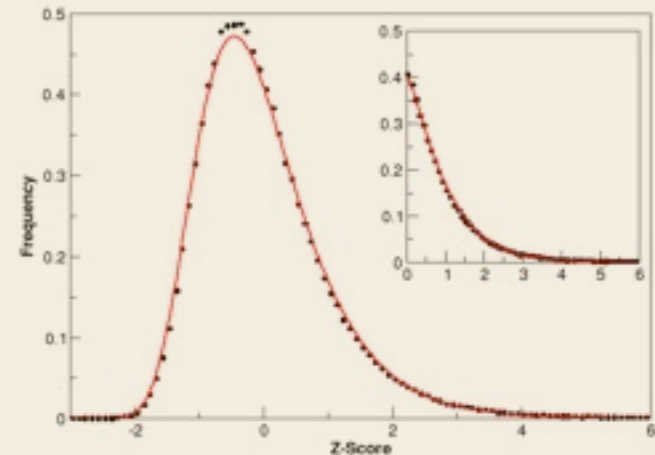


✓ Good scoring system with significance

Reduces the protein representation



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



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

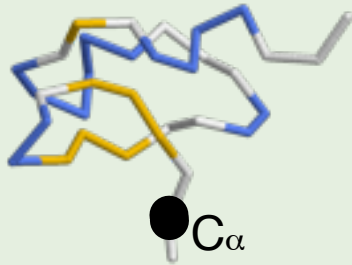
# Vector Alignment Search Tool (VAST)

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



The screenshot shows the NCBI VAST Home Page in a web browser. The browser's address bar displays the URL <http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml>. The page features the NCBI logo and a navigation bar with links to PubMed, Entrez, BLAST, OMIM, Books, TaxBrowser, and Entrez Structure. Below the navigation bar is a search bar labeled "Search Entrez" with the word "Structure" entered and a "Go" button. The main content area is titled "Vector Alignment Search Tool" and includes a "try:" button. The text explains that protein structure neighbors in Entrez are determined by direct comparison of 3-dimensional protein structures with the VAST algorithm. It mentions that more than 87,804 domains in MMDB are 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). A "VAST Search" is a service that allows searching for structural neighbors starting with a set of 3D coordinates specified by... The page also includes a "VAST Help" section with links to "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), and "nr-PDB" (non-redundant protein structure subsets). There are also links to "Get Cn3D v4.1 and look at this example to test!" and "Read a bit more about VAST...".

# Incremental combinatorial extension (CE)



Exhaustive combination  
of fragments

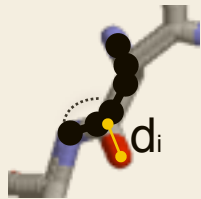
Longest combination of  
AFPs

Heuristic similar to  
PSI-BLAST



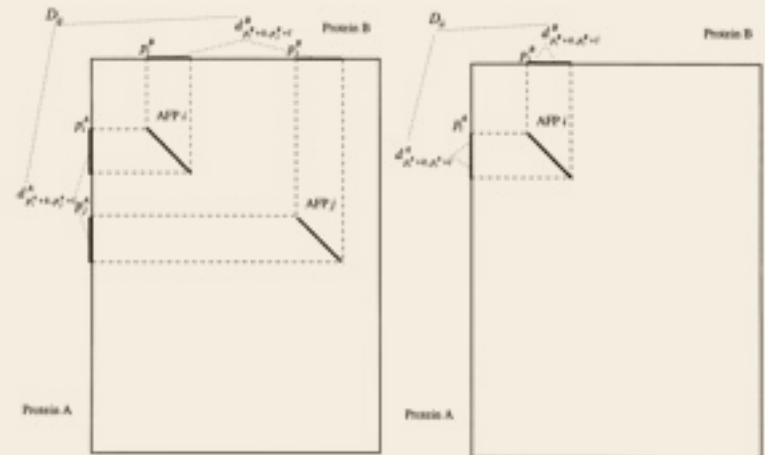
- ✓ FAST!
- ✓ Good quality of local alignments

Complicated scoring and heuristics



8 residues peptides

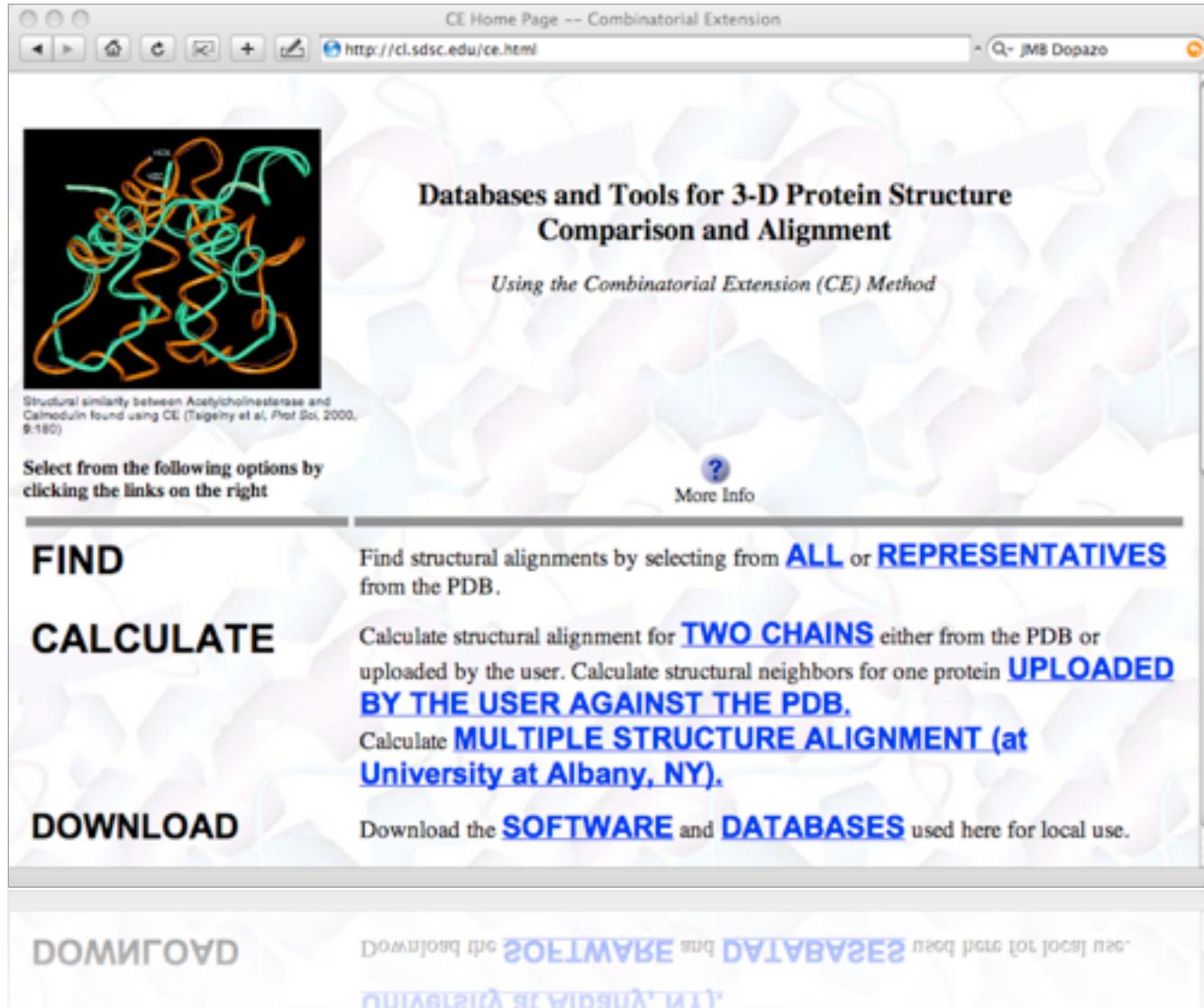
$$RMSD(x, y) = \sqrt{\left(\frac{1}{N}\right) \sum_{i=1}^N (\|x(i) - y(i)\|^2)}$$



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

# Incremental combinatorial extension (CE)

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

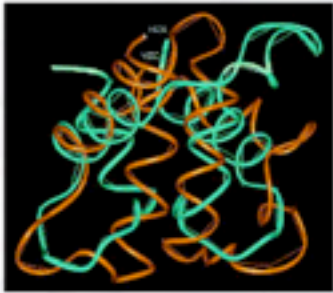


The screenshot shows a web browser window titled "CE Home Page -- Combinatorial Extension". The address bar displays "http://cl.sdsc.edu/ce.html". The page features a background image of protein structures. On the left, there is a 3D ribbon diagram of two protein chains, one in orange and one in green, showing structural similarity. Below this image, a caption reads: "Structural similarity between Acetylcholinesterase and Calmodulin found using CE (Tsigeim et al, Prot Sci, 2000, 9:180)". To the right of the image, the main heading is "Databases and Tools for 3-D Protein Structure Comparison and Alignment", followed by the subtitle "Using the Combinatorial Extension (CE) Method". Below the subtitle, there is a "More Info" link with a question mark icon. The page is divided into three main sections: "FIND", "CALCULATE", and "DOWNLOAD". The "FIND" section describes finding structural alignments by selecting from "ALL" or "REPRESENTATIVES" from the PDB. The "CALCULATE" section describes calculating structural alignment for "TWO CHAINS" (either from the PDB or uploaded by the user) and calculating structural neighbors for one protein "UPLOADED BY THE USER AGAINST THE PDB". It also mentions calculating "MULTIPLE STRUCTURE ALIGNMENT (at University at Albany, NY)". The "DOWNLOAD" section describes downloading the "SOFTWARE" and "DATABASES" used for local use. At the bottom of the page, there is a "DOWNLOAD" section that repeats the information about downloading the software and databases.

CE Home Page -- Combinatorial Extension

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

Q: JMB Dopazo



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

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

More Info

## 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)**.

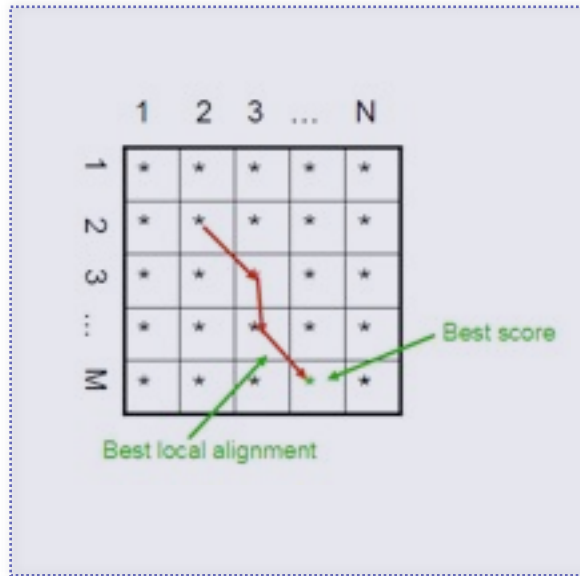
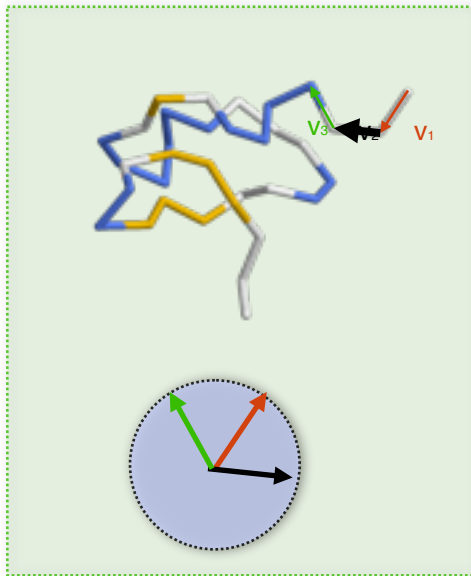
## DOWNLOAD

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

DOWNLOAD

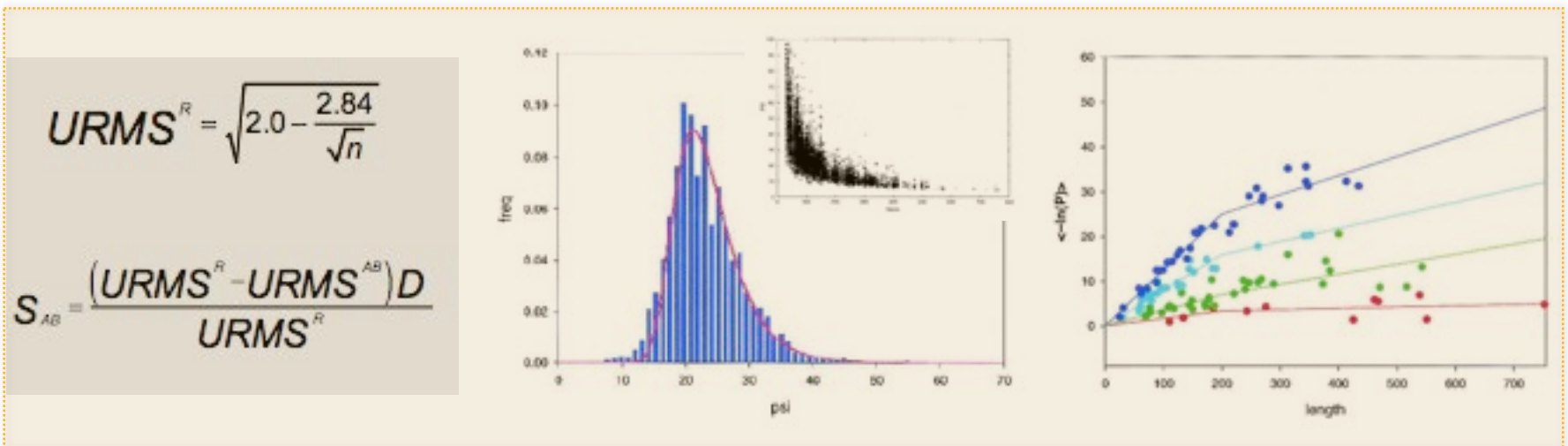
Download the **SOFTWARE** and **DATABASES** used here for local use.

# Matching molecular models obtained from theory (MAMMOTH)



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

Reduces the protein representation



Ortiz AR, (2002) *Protein Sci.* 11 pp2606

# Matching molecular models obtained from theory (MAMMOTH)

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

The screenshot shows a web browser window with the title "MAMMOTH Pairwise Protein Structure Alignment Server". The address bar shows the URL "http://ub.cbm.uam.es/mammoth/pair/index3.php". The page header includes the "Bioinformatics Unit - CBMSO" logo, the "MAMMOTH-mult" logo, and the text "Multiple Protein Structure Alignment Server". A date stamp "Madrid, Monday, November, 3th, 2008" is visible. The main content area has a left sidebar with "More information" and "Contact" links, and a "Webmaster" section with a "New Alignment" button. The main form area contains two file upload sections: "Upload the coordinates file (POB format) of your first protein:" and "Upload the coordinates file (POB format) of your second protein:", each with a "Choose File" button and "no file selected" text. Below these is an email input field with the label "Your e-mail for results to be sent back:" and a note: "\*some calculations may take upto few minutes, it is recommended that you include your email!". At the bottom of the form are "Align" and "Reset" buttons. The footer contains "CBMSO | Home" and "©2004 MAMMOTH Team". A status bar at the bottom indicates "Failed to open page (see Activity window for details)".

MAMMOTH Pairwise Protein Structure Alignment Server

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

Bioinformatics Unit - CBMSO

**MAMMOTH-mult**  
Multiple Protein Structure Alignment Server

Madrid, Monday, November, 3th, 2008

More information  
Contact

Webmaster

Upload the coordinates file (POB format) of your **first protein**:  no file selected

Upload the coordinates file (POB format) of your **second protein**:  no file selected

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

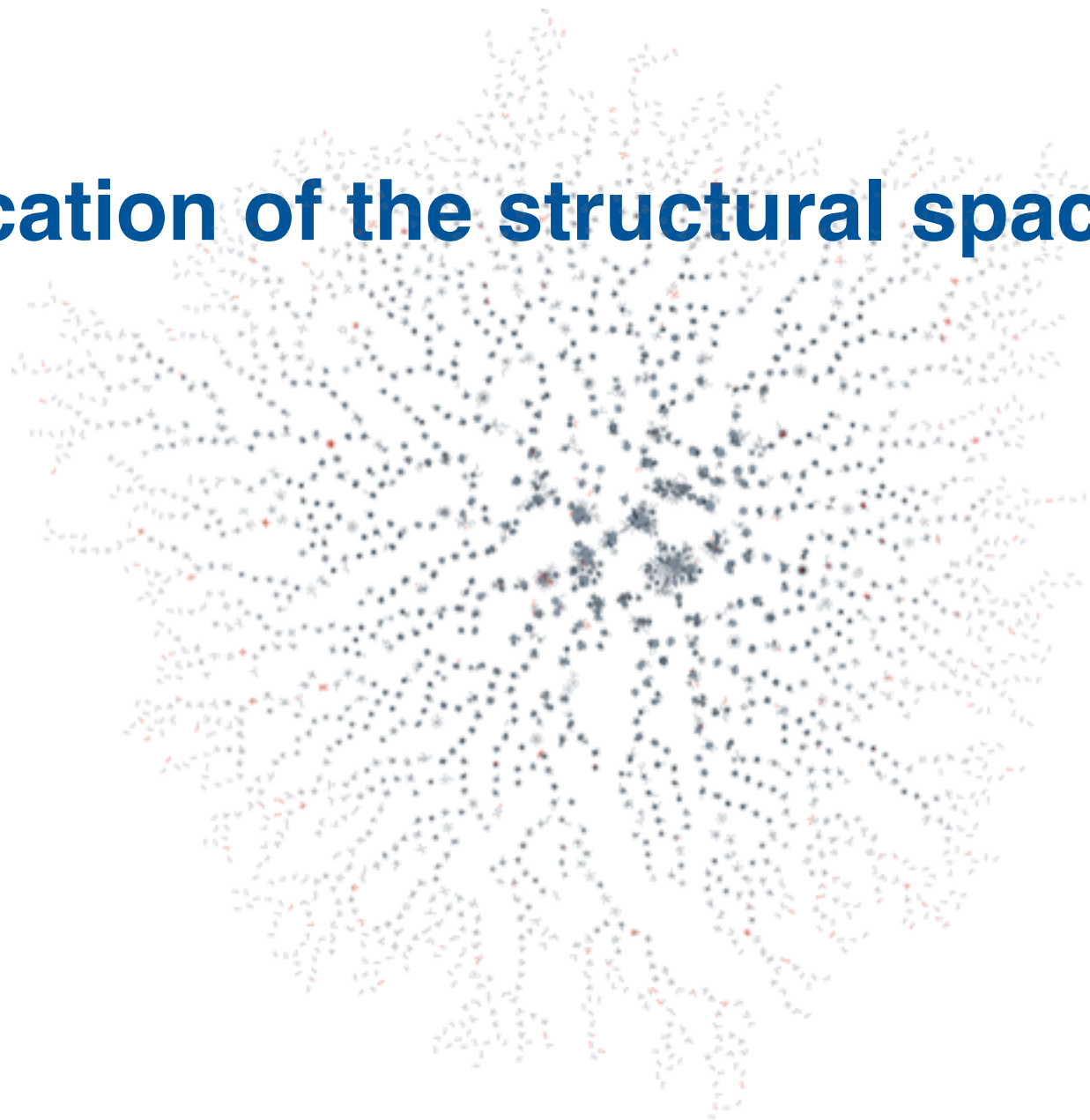
*\*some calculations may take upto few minutes, it is recommended that you include your email!*

CBMSO | Home  
©2004 MAMMOTH Team

Failed to open page (see Activity window for details)



# Classification of the structural space





# SCOP<sub>1.75</sub> database

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



- ✓ Largely recognized as “standard of gold”
- ✓ Manually classification
- ✓ Clear classification of structures in:  
 CLASS  
 FOLD  
 SUPER-FAMILY  
 FAMILY
- ✓ Some large number of tools already available

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

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

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

# CATH<sub>3.2</sub> 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**



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<sub>v2.0</sub> database

<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:	95,804
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

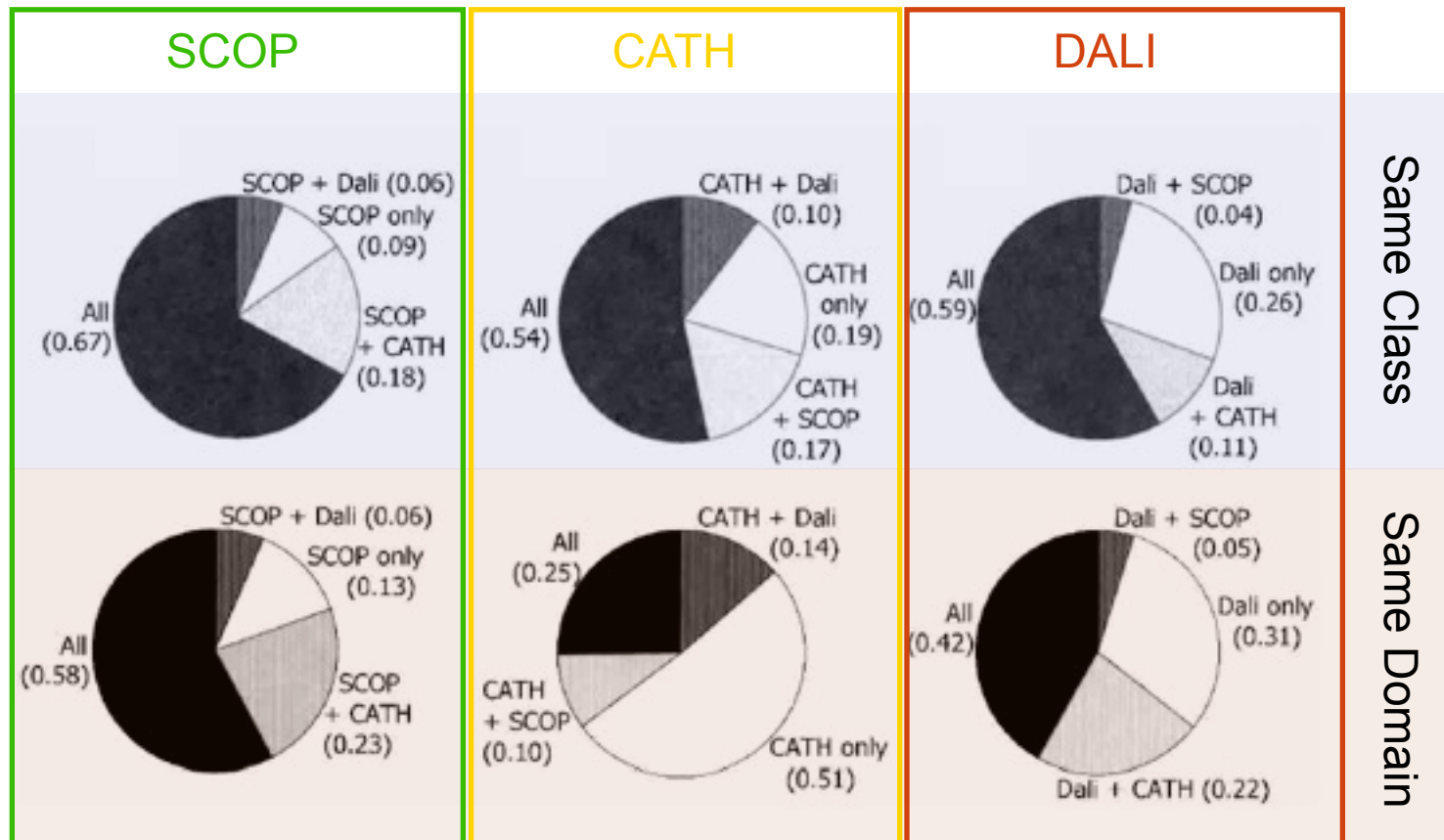
Marti-Renom et al. 2007. *BMC BMC Bioinformatics* (2007) **8** (Suppl 4) S4

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



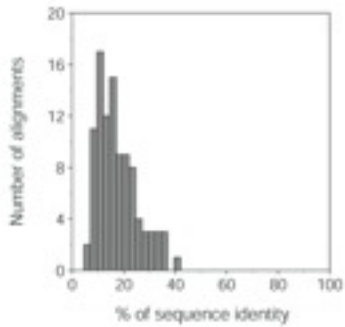
Day, et al. (2003) Protein Sciences, 12 pp2150

# template search and template-target alignment (pp\_scan)

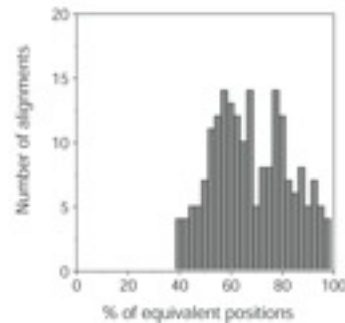
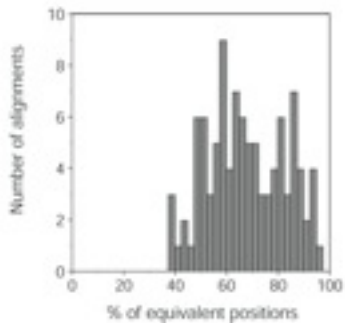
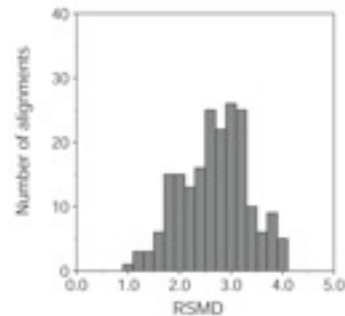
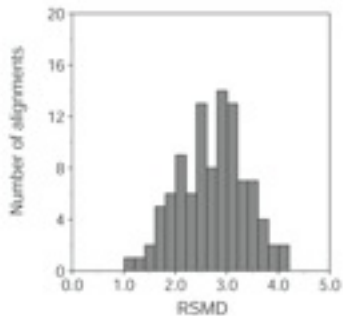
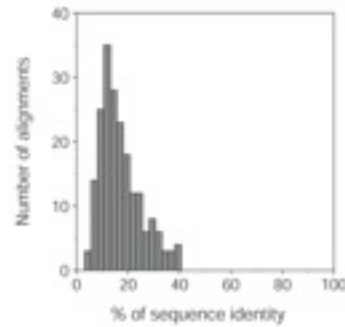
*Marti-Renom, et al. (2004) Prot. Sci. 13 pp1071*  
*Narayanan, et al. in prepration*

# PP\_SCAN or profile-profile alignments

A) Training Set



B) Testing Set



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:** HMM + 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)

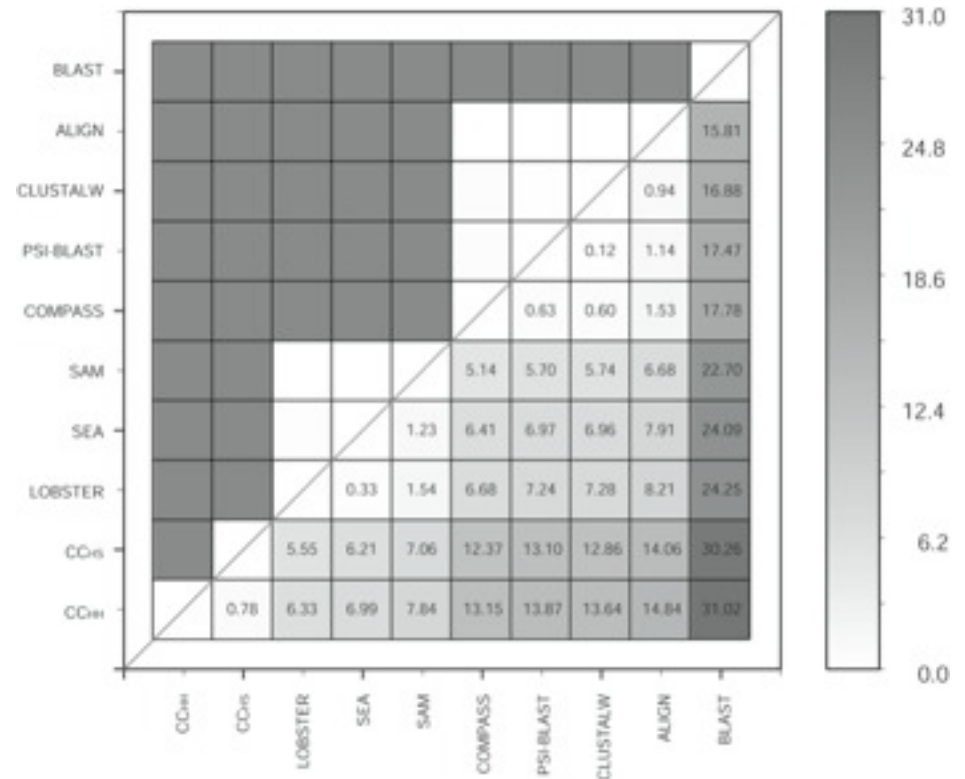
# PP\_SCAN protocols accuracy

SALIGN protocol	CE overlap [%]	Shift score
CC <sub>PBP</sub>	55 ± 23	0.61 ± 0.24
CC <sub>HH</sub>	<b>56 ± 23</b>	<b>0.61 ± 0.24</b>
CC <sub>HS</sub>	<b>56 ± 24</b>	<b>0.62 ± 0.23</b>
CC <sub>MAT</sub>	51 ± 25	0.55 ± 0.27
ED <sub>PBP</sub>	54 ± 24	0.60 ± 0.25
ED <sub>HH</sub>	54 ± 24	0.59 ± 0.26
ED <sub>HS</sub>	55 ± 24	0.59 ± 0.26
DP <sub>PBP</sub>	55 ± 23	0.61 ± 0.24
DP <sub>HH</sub>	56 ± 23	0.60 ± 0.25
DP <sub>HS</sub>	55 ± 24	0.61 ± 0.24
JS <sub>HH</sub>	53 ± 24	0.60 ± 0.24
JS <sub>HS</sub>	54 ± 24	0.60 ± 0.24
Ave <sub>MAT</sub>	49 ± 26	0.52 ± 0.29
<b>TOP</b>	<b>62 ± 20</b>	<b>0.67 ± 0.20</b>

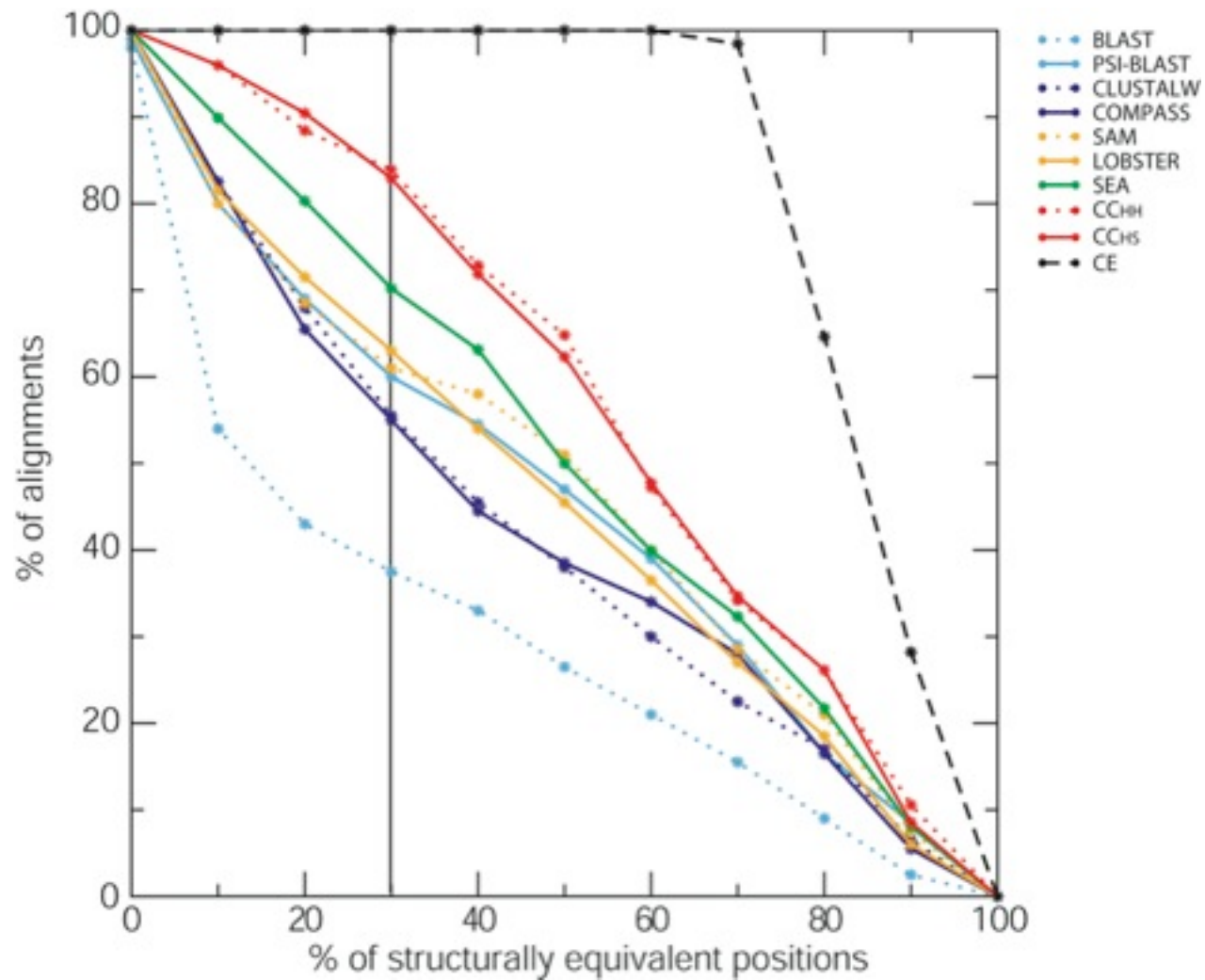


# PP\_SCAN accuracy

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



# 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%

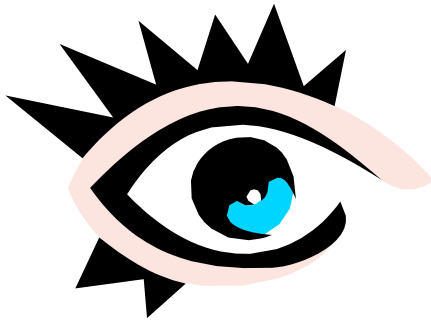


APL 000000

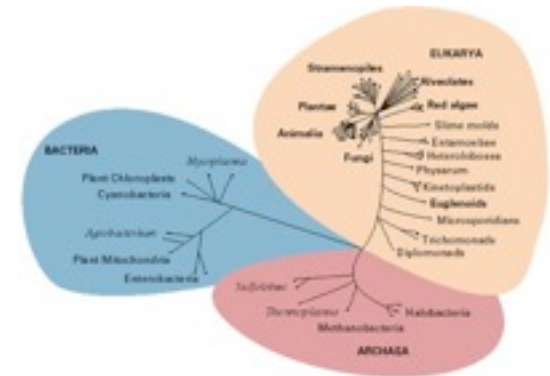
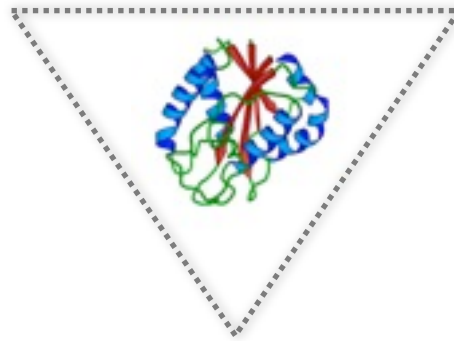
# **model building and model assessment**

# Information about a protein can come from three distinct sources

<http://www.integrativemodeling.org/>



Experimental observations



Statistical rules



Laws of physics

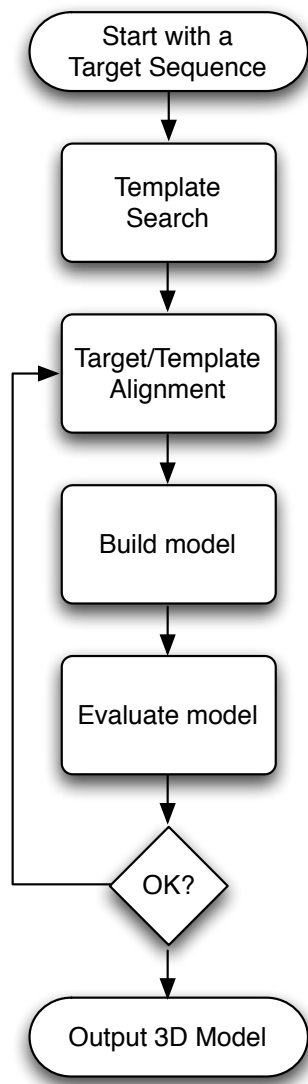
# Classes of methods for comparative protein structure modeling

- ◆ Model building by assembly of rigid bodies  
core, loops, side-chains.
- ◆ 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



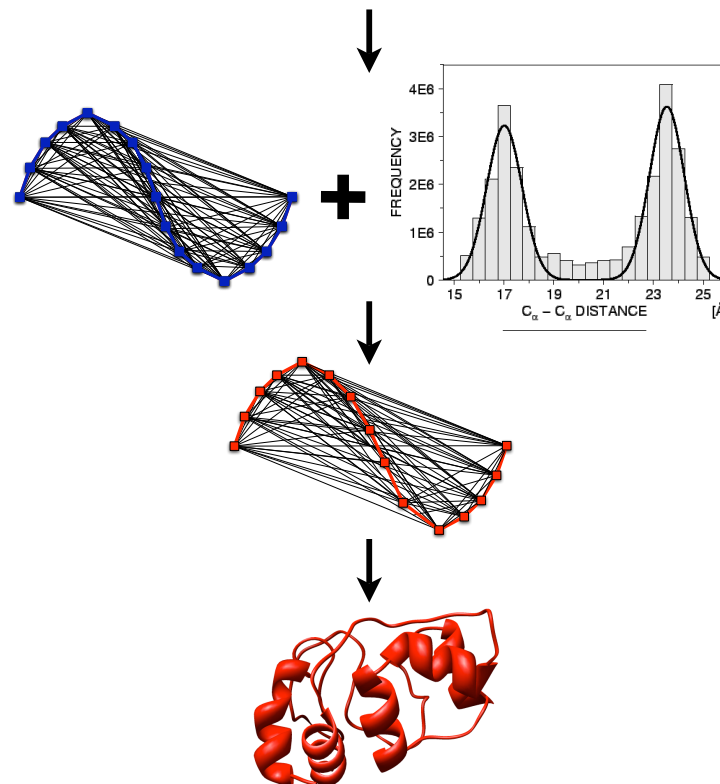
Given an alignment...

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

MSVIPKR--GNCEQTSE  
ASILPKRLFGNCEQTSD



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

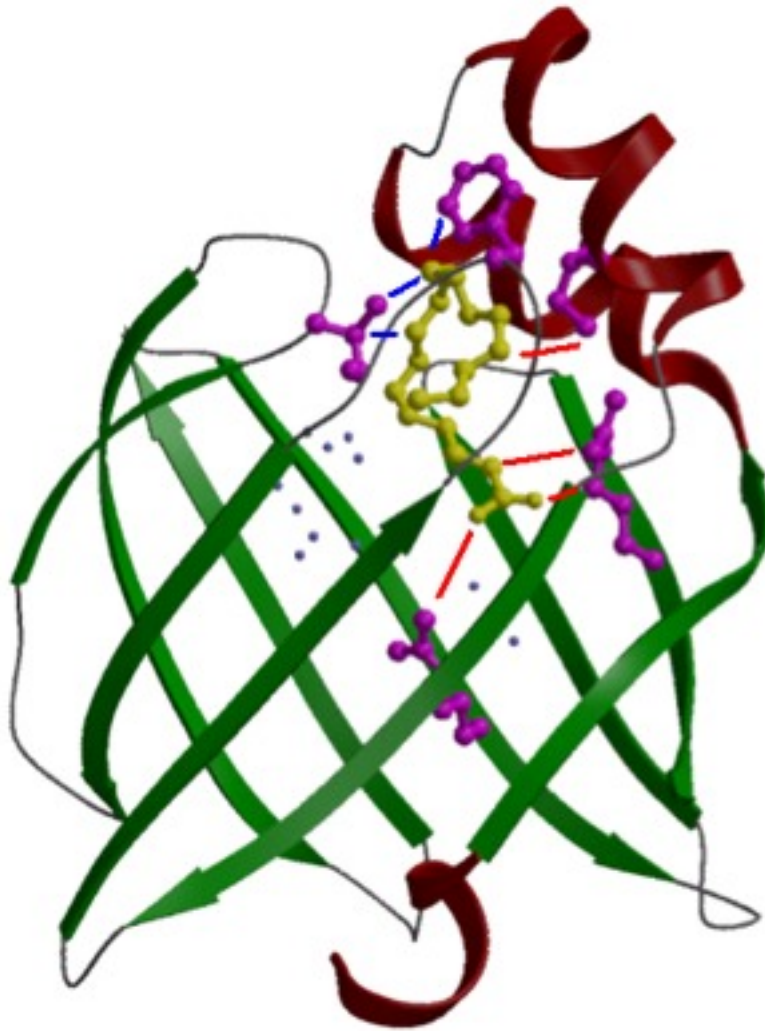
KSINPIHGDNCEQTSDEGLKIERTPL-----QWLKSSICDMRGLIPE

Target

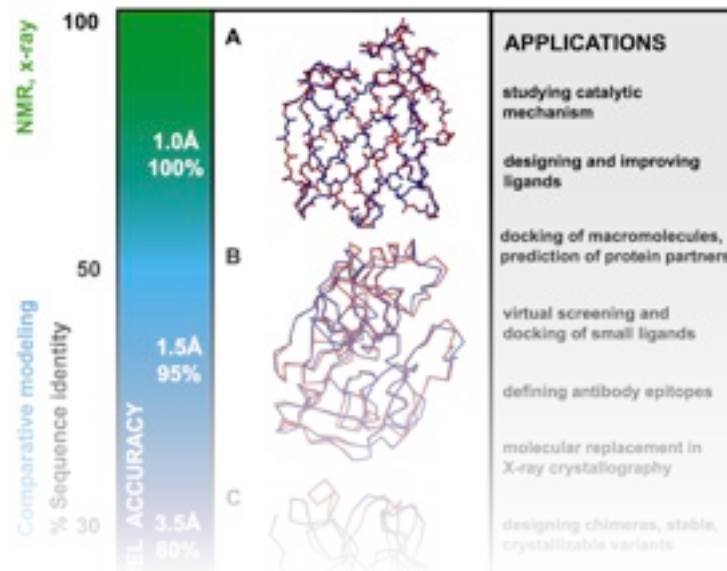
ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE

MSVIPKRLYGNCEQTSEEAIRIEDSPIVRWISAQLVCLKIDEIPERLVGE

# Modeling ligands and using external restraints

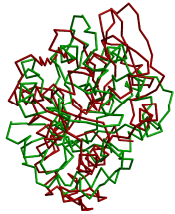


- Homology derived restraint
- External Restraint

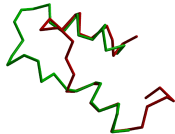


# Accuracy and applicability of comparative models

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



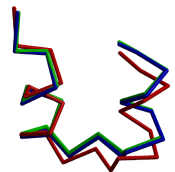
**Wrong fold**



**Miss alignments**



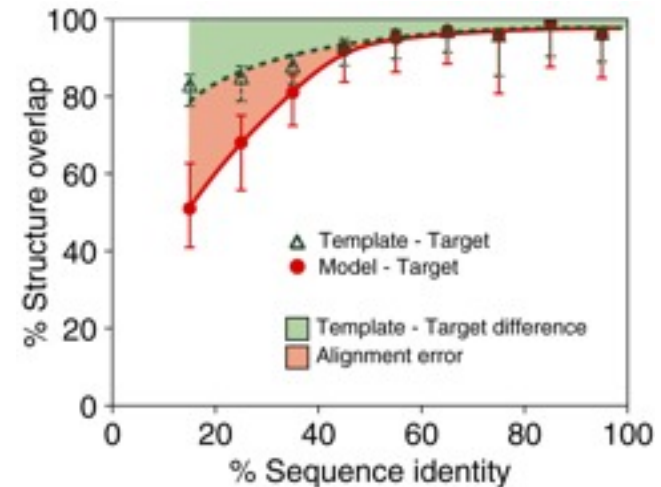
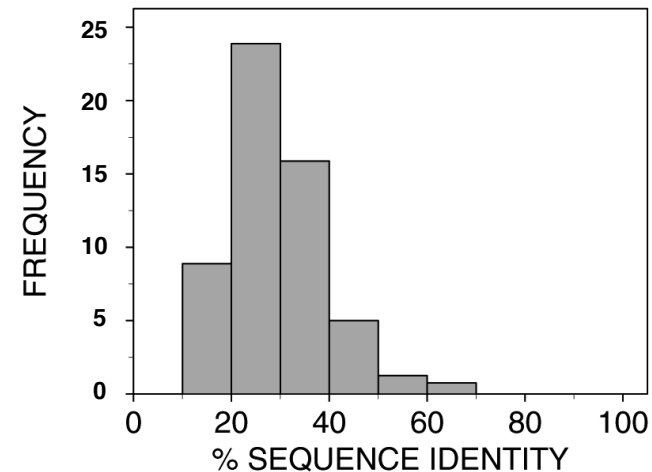
**Loop regions**



**Rigid body distortions**

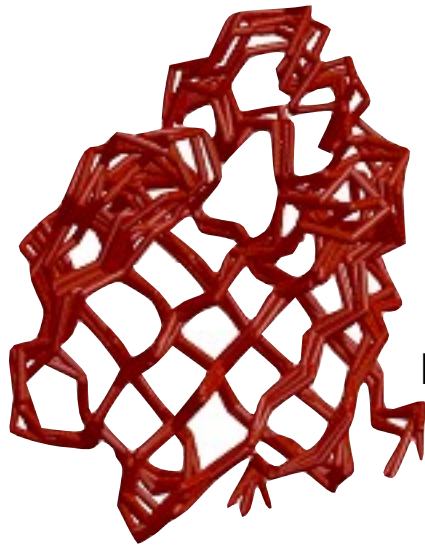


**Side-chain packing**



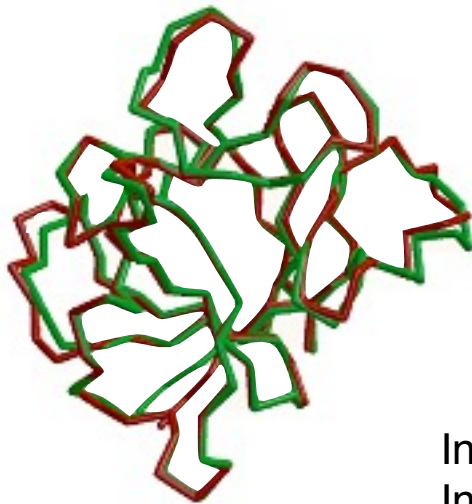
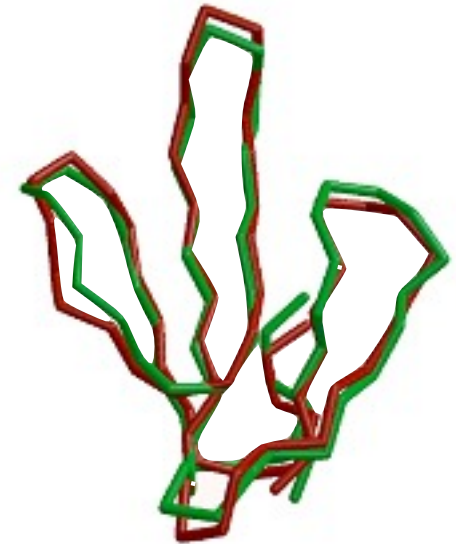
*Marti-Renom et al. Ann Rev Biophys Biomol Struct (2000) 29, 291*

# “Biological” significance of modeling errors



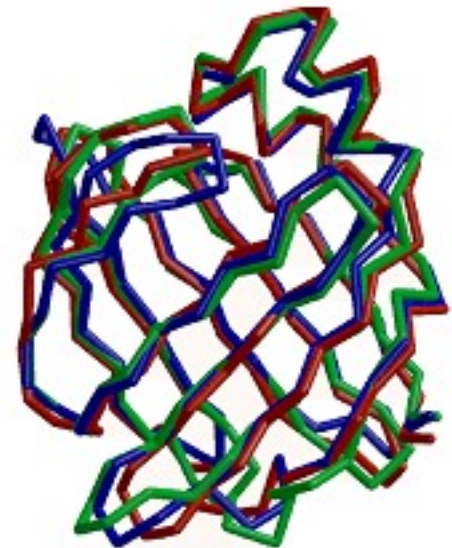
**NMR**  
Ileal lipid-binding protein  
1eal

**NMR – X-RAY**  
Erabutoxin 3ebx  
Erabutoxin 1era



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

**CRABP II** 1opbB  
**FABP** 1ftpA  
**ALBP** 1lib  
40% seq. id.

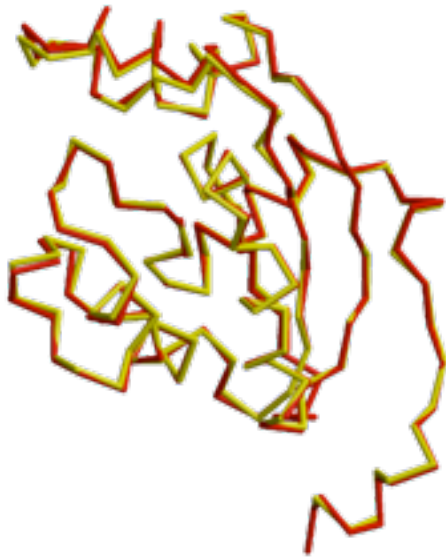


# Model Accuracy

## HIGH ACCURACY

NM23 Seq id 77%

C $\alpha$  equiv 147/148  
RMSD 0.41Å

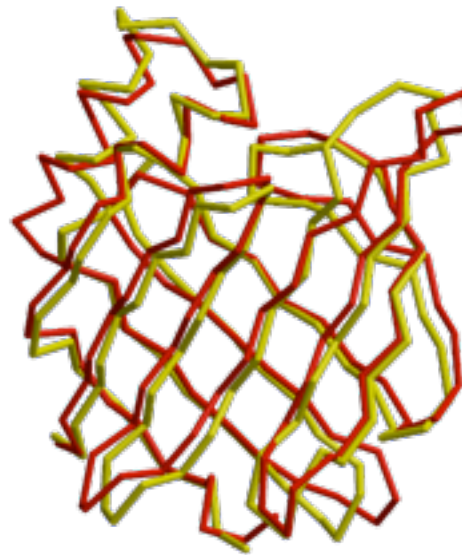


Sidechains  
Core backbone  
Loops

## 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



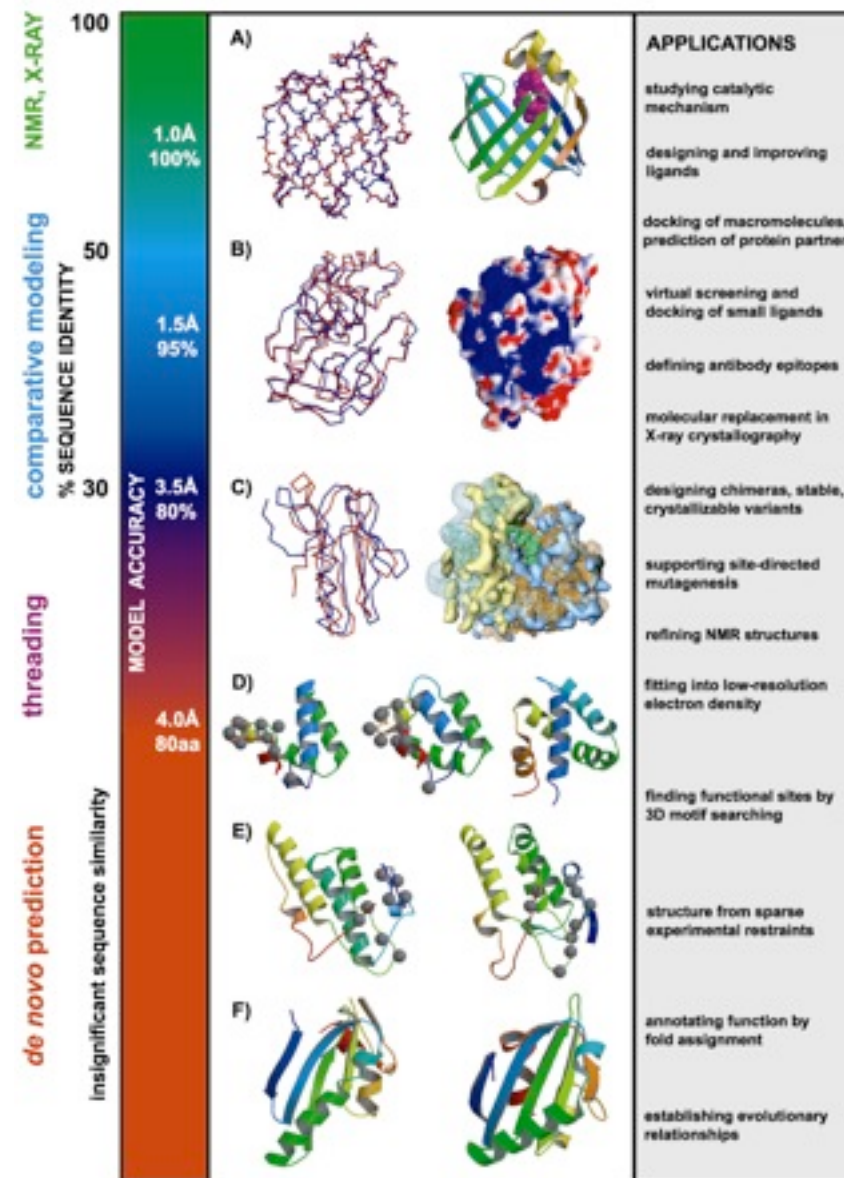
Sidechains  
Core backbone  
Loops  
Alignment  
Fold assignment

X-RAY / MODEL

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

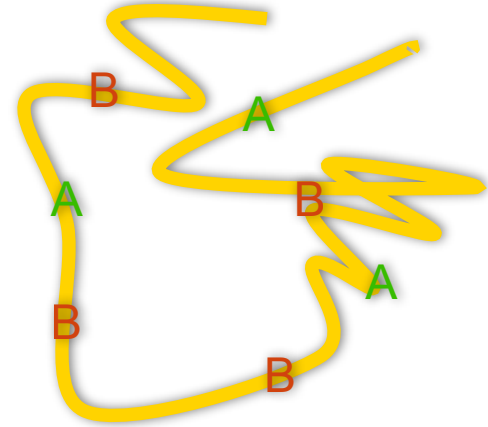


# Utility of protein structure models, despite errors



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

# Model Assessment (PMF)





# Scoring

## Statistical Potential (inspiration)

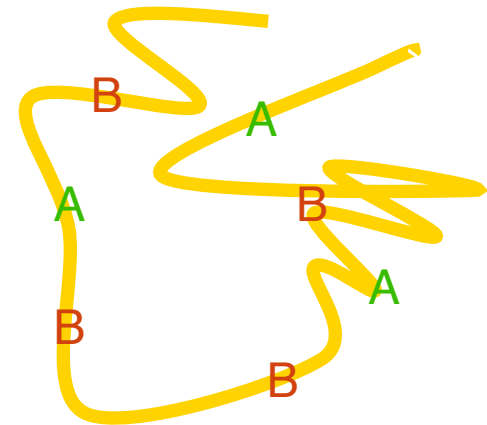
$$K = \frac{[AB]}{[A] \cdot [B]}$$

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

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

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

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.

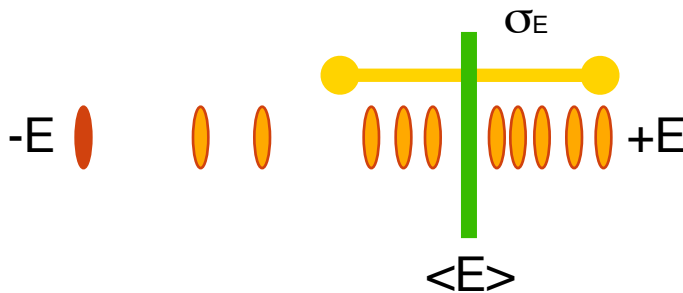


*Tanaka and Sheraga (1975) PNAS, 72 pp3802*  
*Sippl, (1990) J.Mo.Biol. 213 pp859*  
*Godzik, (1996) Structure 15 pp363*

## Scoring

# Significance of an alignment (score)

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



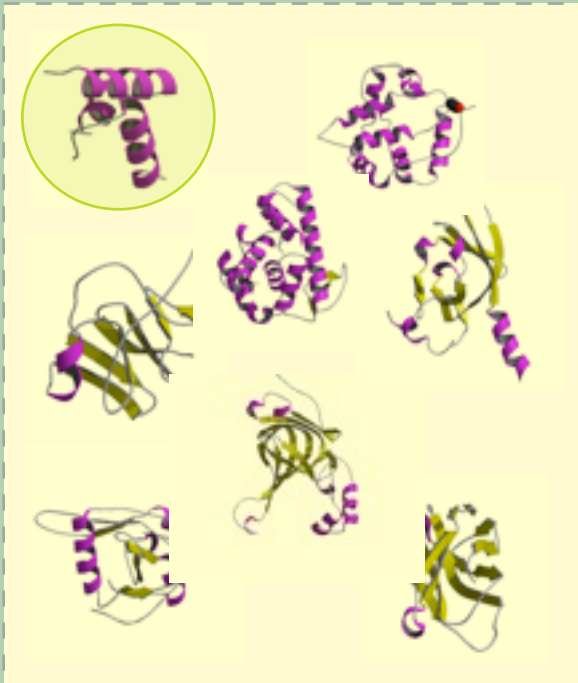
$$Zscore = \frac{(\langle E \rangle - E_m)}{\sigma_E}$$

# Prosall

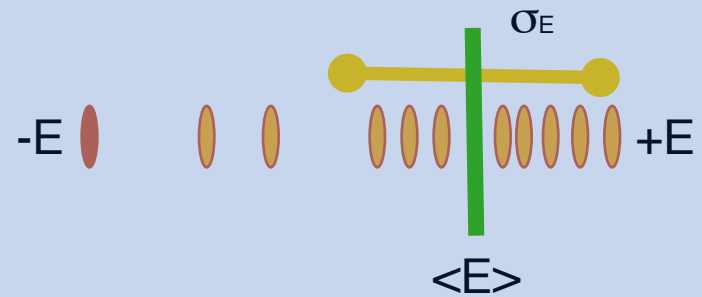
<http://www.came.sbg.ac.at>

## Deriving

Structural space



## Scoring



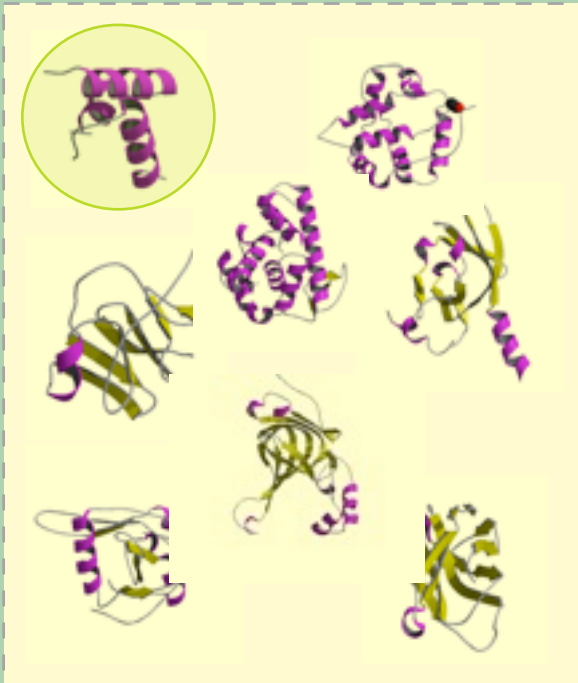
$$Zscore = \frac{(\langle E \rangle - E_m)}{\sigma_E}$$

# ANOLEA

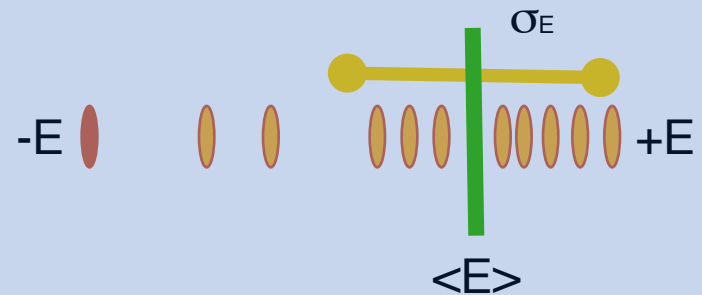
<http://protein.bio.puc.cl/cardex/servers/anolea/>

## Deriving

Structural space



## Scoring



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

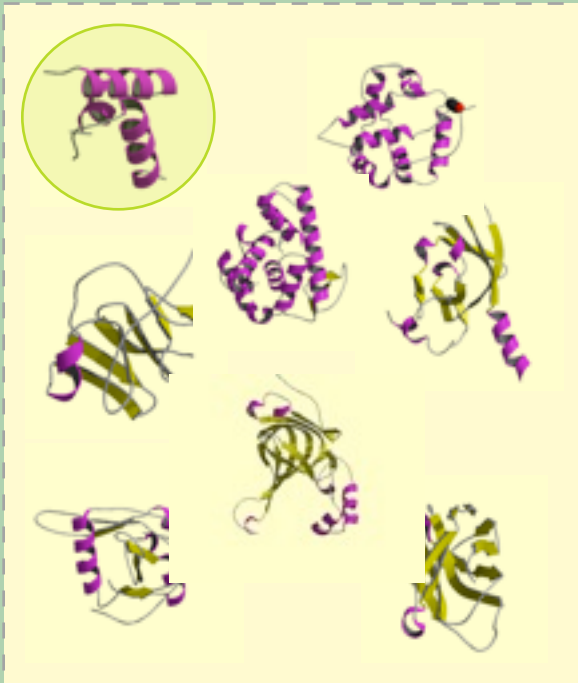
all atom potential

# Verify3D

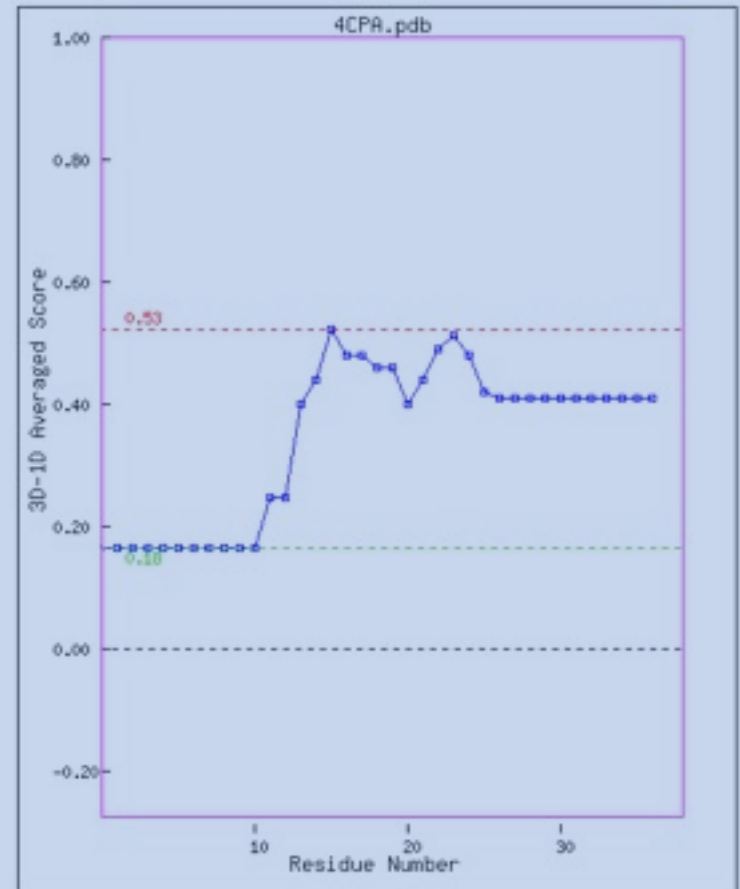
[http://nihserver.mbi.ucla.edu/Verify\\_3D/](http://nihserver.mbi.ucla.edu/Verify_3D/)

## Deriving

Structural space



## Scoring

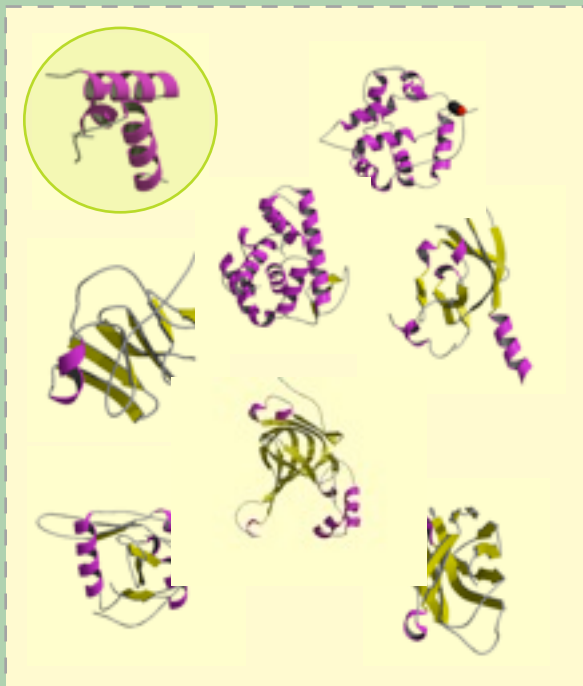


# DFIRE

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

## Deriving

Structural space



## Scoring

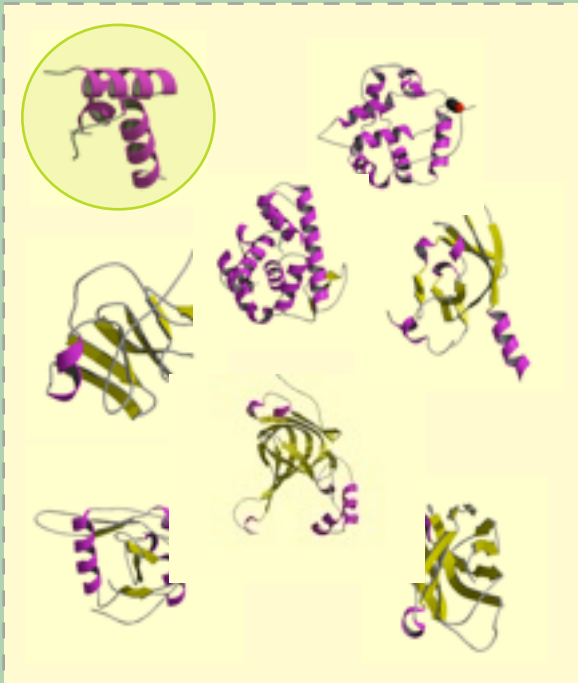
Pseudo-Energy  
with respect a  
ideal gas-phase  
reference state

# DOPE (MODELLER)

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

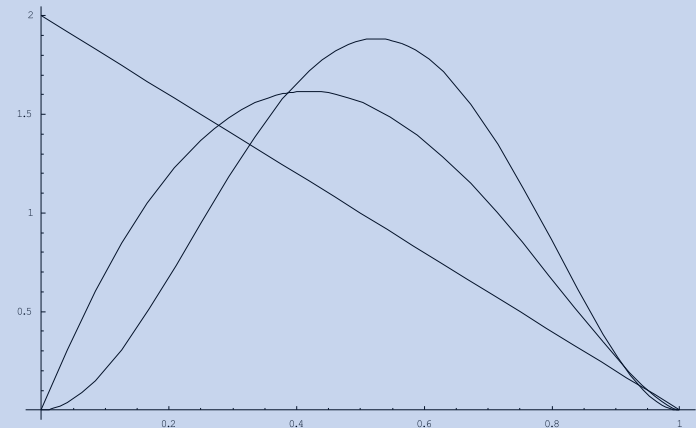
## Deriving

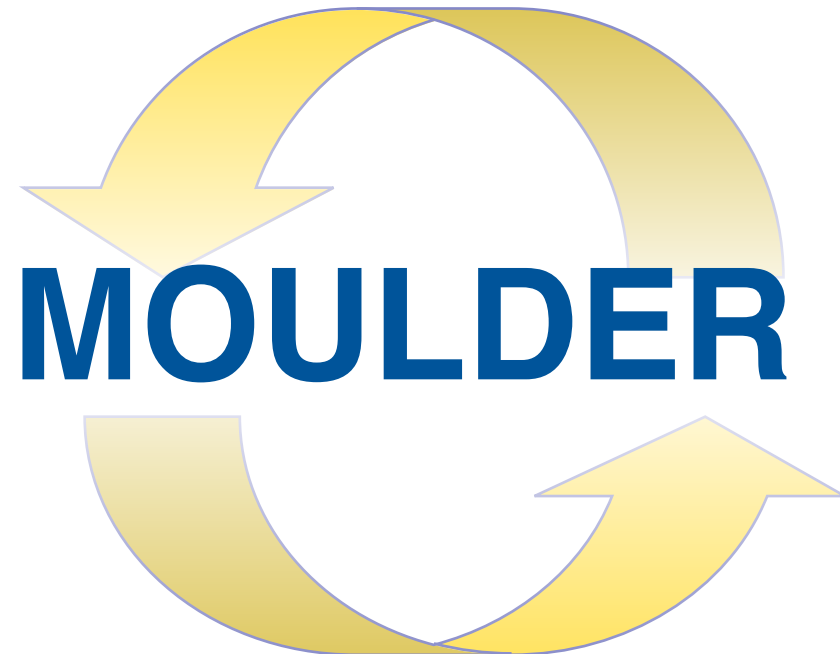
Structural space



## Scoring

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

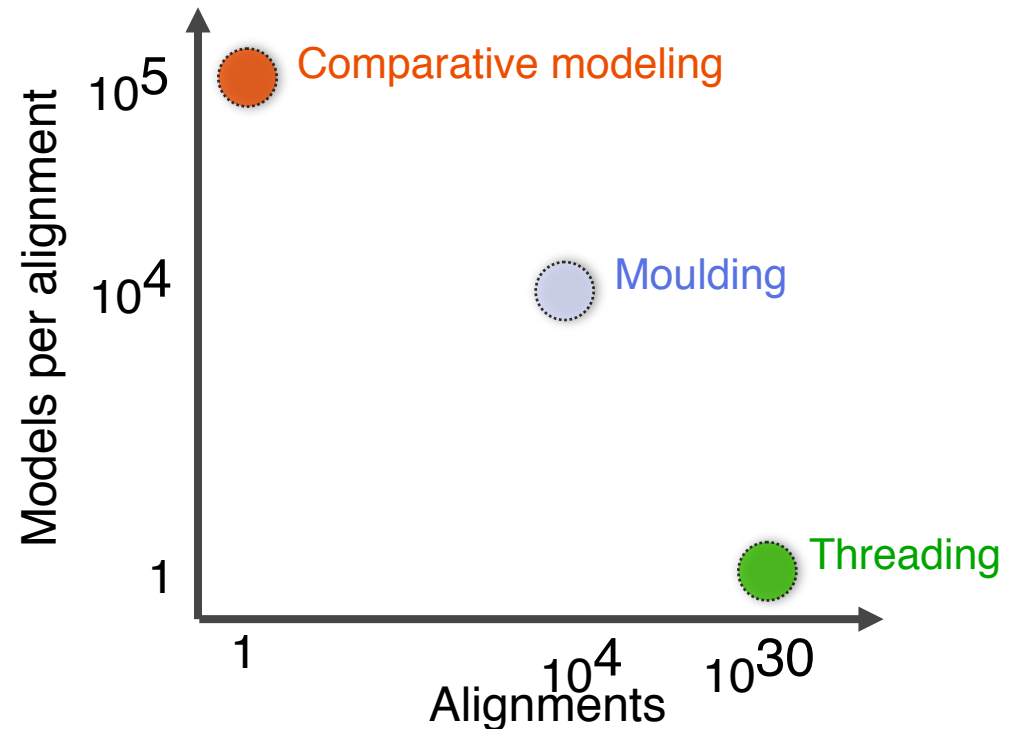
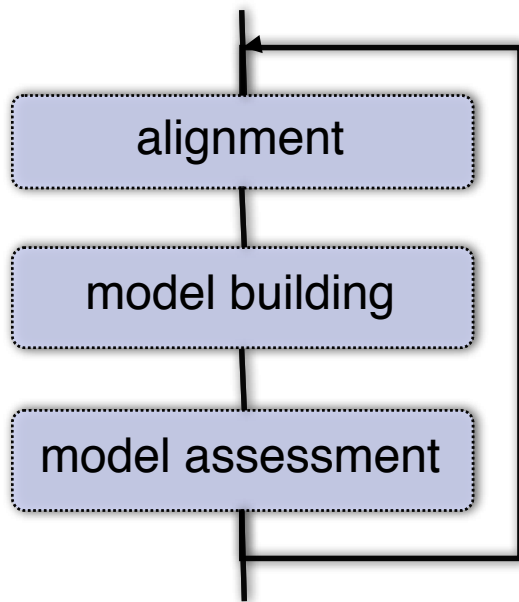




*John, Sali (2003). NAR pp31 3982*



# Moulding: iterative alignment, model building, model assessment



# Genetic algorithm operators

## Single point cross-over

...TSSQ—NMKLG VFWGY—...  
...V—SSCN—GDLHMKVGV—...



...TSSQN MK—LGVFWGY...  
...VSSCN GDLHMKV—GV...

...TSSQ—NMK—LGVFWGY...  
...V—SSCN GDLHMKV—GV...

...TSSQN MKLG VFWGY—...  
...VSSCN—GDLHMKVGV—...

## Gap insertion

...TSSQN MKLG VFWGY...  
...VSSCN GDLHMKVGV...



...TSSQN—MKLG VFWGY...  
...VSSCN GDLHMKVG—V...

## Gap shift

...T—S S QNMKLG VFWGY...  
...VSSC N GDLHMKVGV—...



...—T—S S QNMKLG VFWGY...  
...VSSC N GDLHMKVGV—...

...T—S—S QNMKLG VFWGY...  
...VSSC N GDLHMKVGV—...

...—T S S QNMKLG VFWGY...  
...VSSC N GDLHMKVGV—...

...T S—S QNMKLG VFWGY...  
...VSSC N GDLHMKVGV—...

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 ( $A_s$ ).

$$Z = 0.17 Z(P_p) + 0.02 Z(P_s) + 0.10 Z(S_c) + 0.26 Z(H_a) + 0.45 (A_s)$$

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

$\mu$  ... average score of all models

$\sigma$  ... standard deviation of the scores

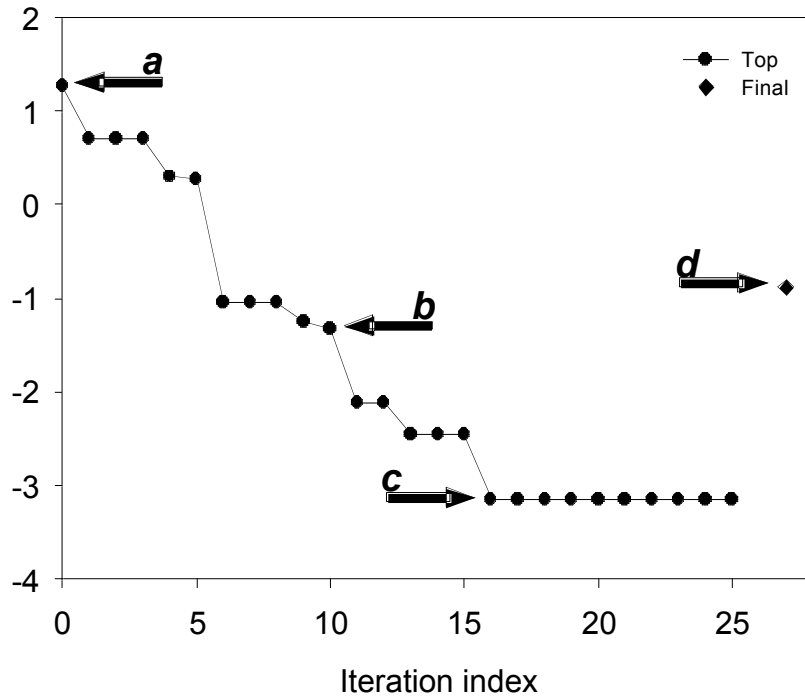
# Benchmark with the “very difficult” test set

## 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	
			C $\alpha$ RMSD [Å]	CE overlap [%]	C $\alpha$ RMSD [Å]	CE overlap [%]	C $\alpha$ 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

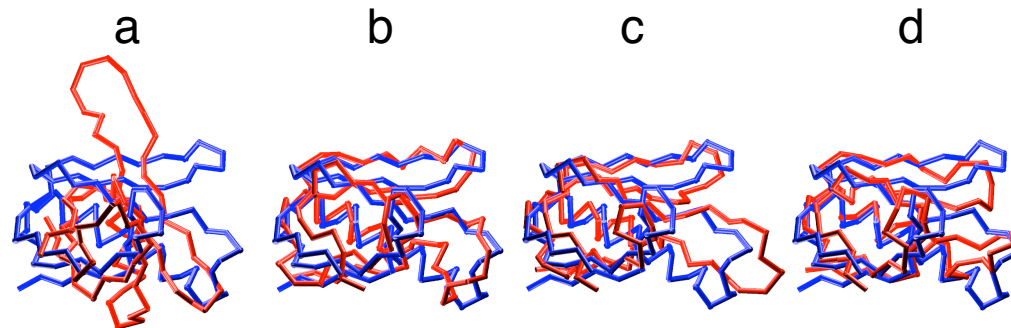
## 1BOV-1LTS



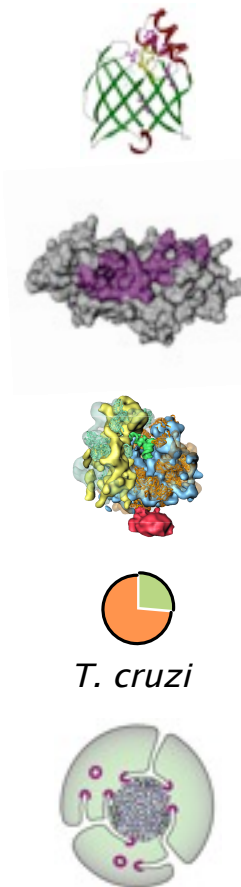
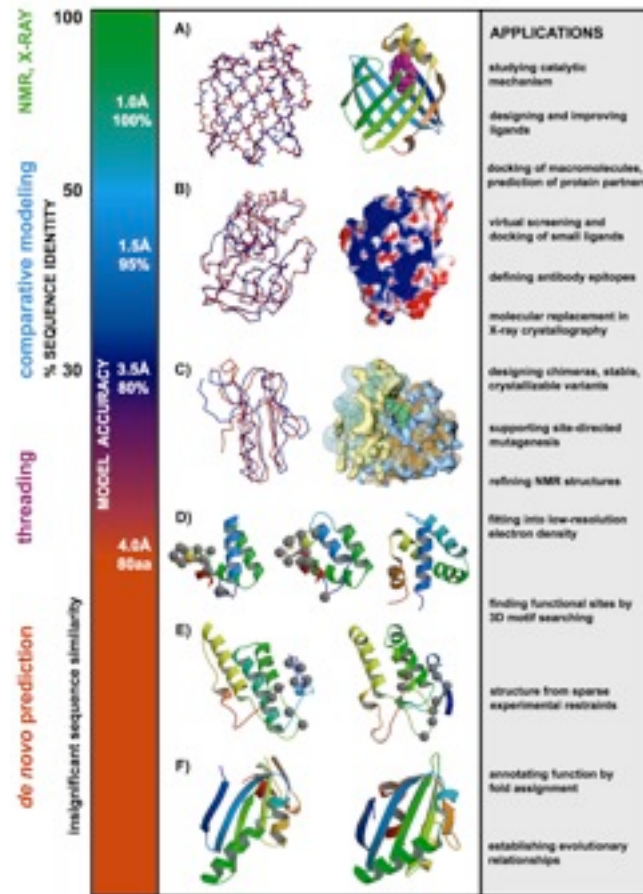
Sequence identity 4.4%

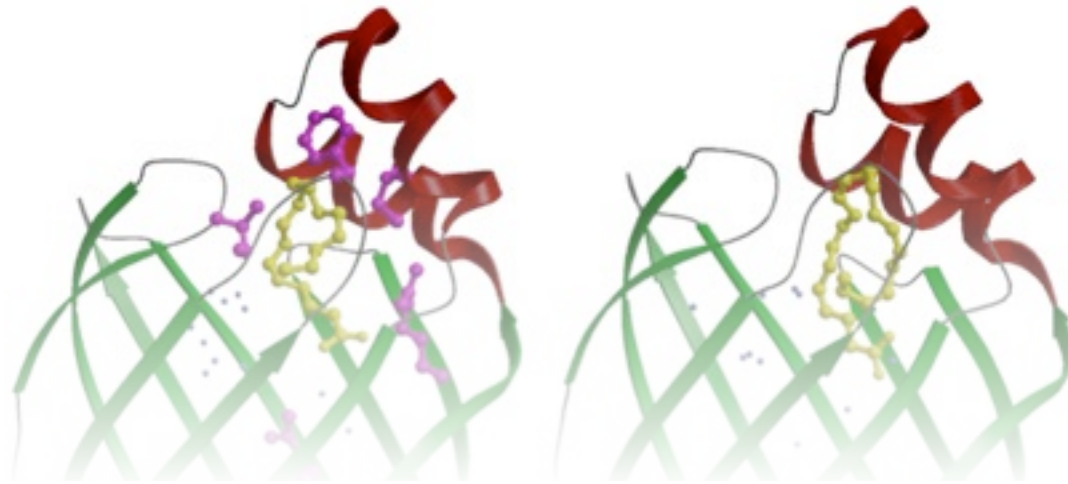
Initial model C $\alpha$  RMSD 10.1Å

Final model C $\alpha$  RMSD 3.6Å



# Can we use models to infer function?





# Modeling genes

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

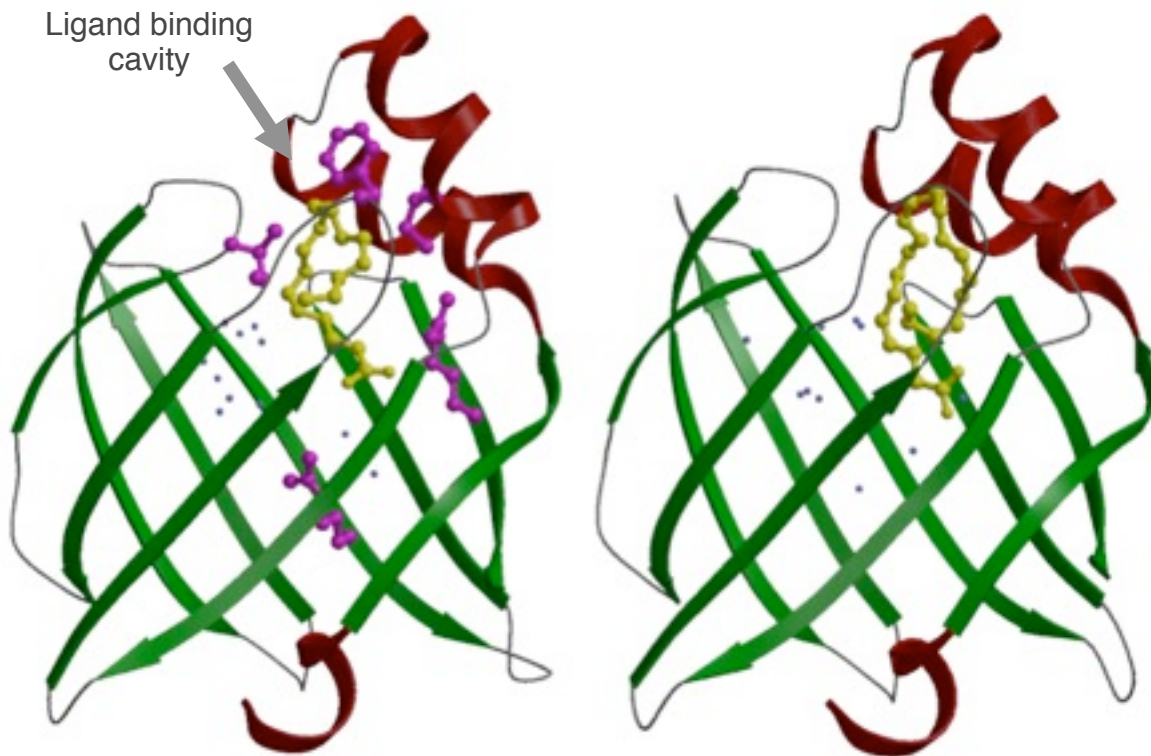
Predicting features of a model that are not present in the template

BLBP/oleic acid

Cavity is **not** filled

BLBP/docosahexaenoic acid

Cavity **is** filled



1. BLBP binds fatty acids.

2. Build a 3D model.

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

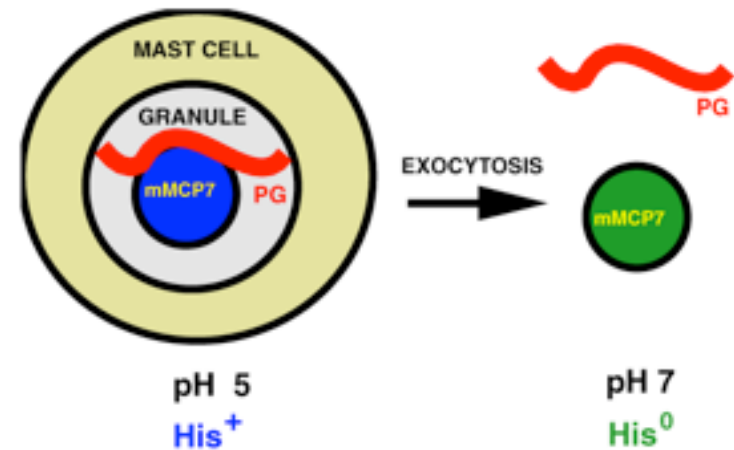
L. Xu, R. Sánchez, A. Šali, N. Heintz, J. Biol. Chem. 271, 24711, 1996.



Do mast cell proteases bind proteoglycans? Where? When?

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

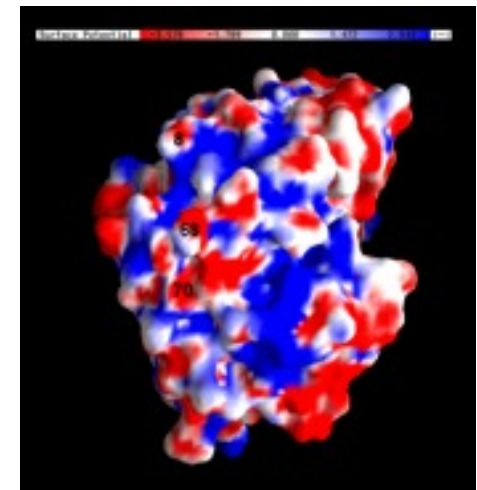
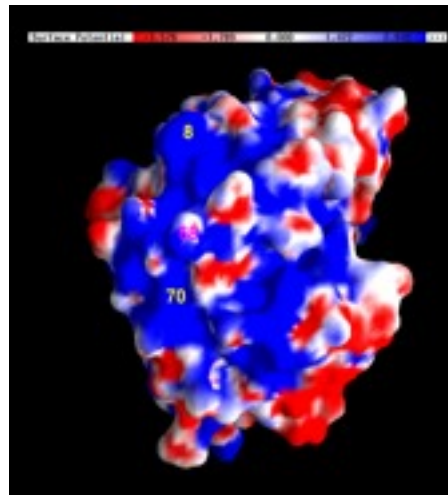
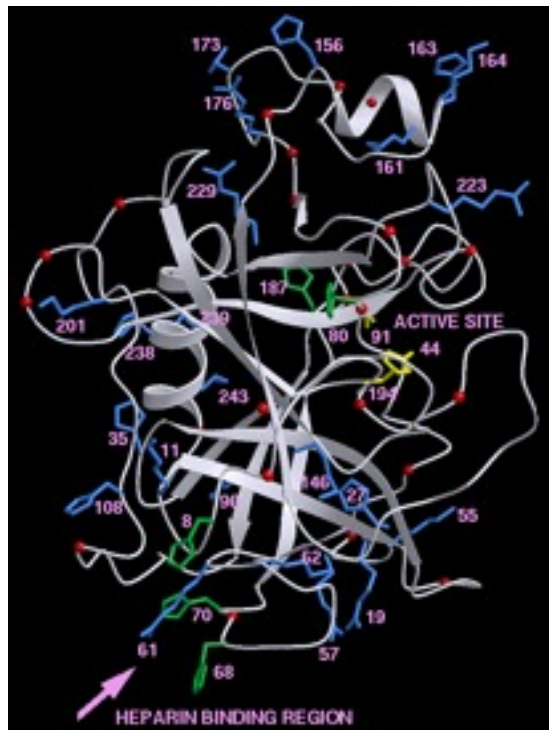
1. mMCPs bind negatively charged proteoglycans through electrostatic interactions
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.



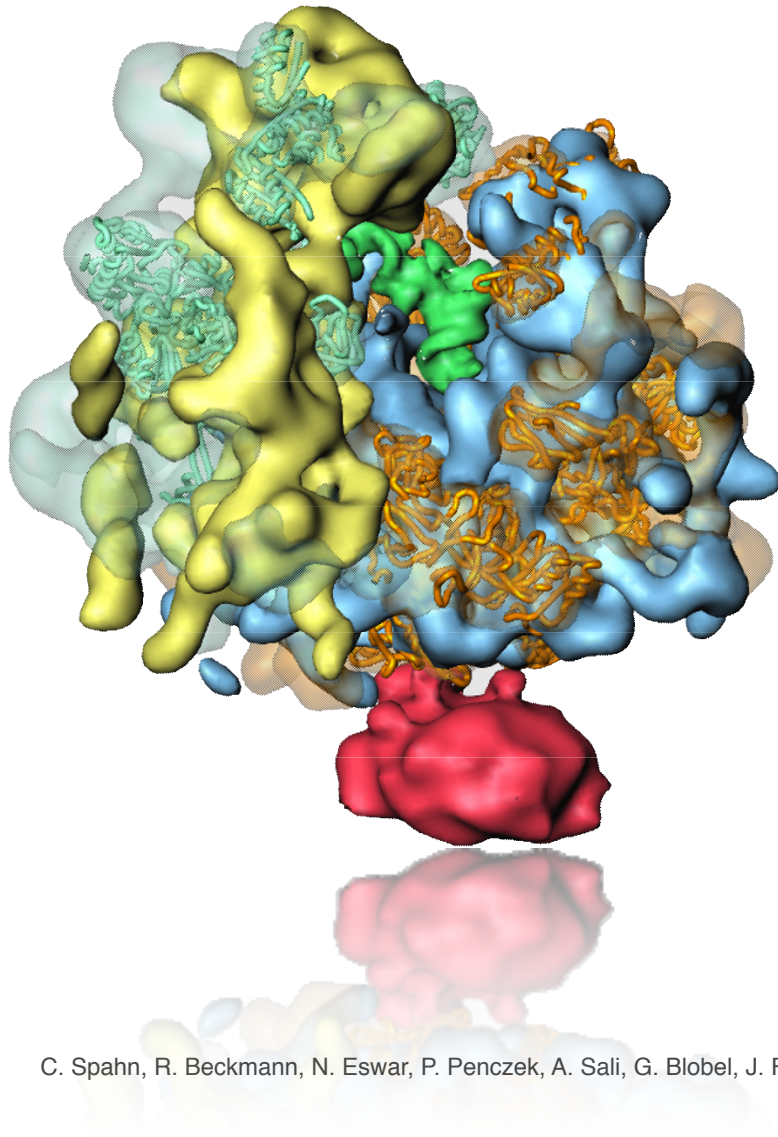
Huang et al. *J. Clin. Immunol.* **18**,169,1998.

Matsumoto et al. *J. Biol. Chem.* **270**,19524,1995.

Šali et al. *J. Biol. Chem.* **268**, 9023, 1993.



# *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

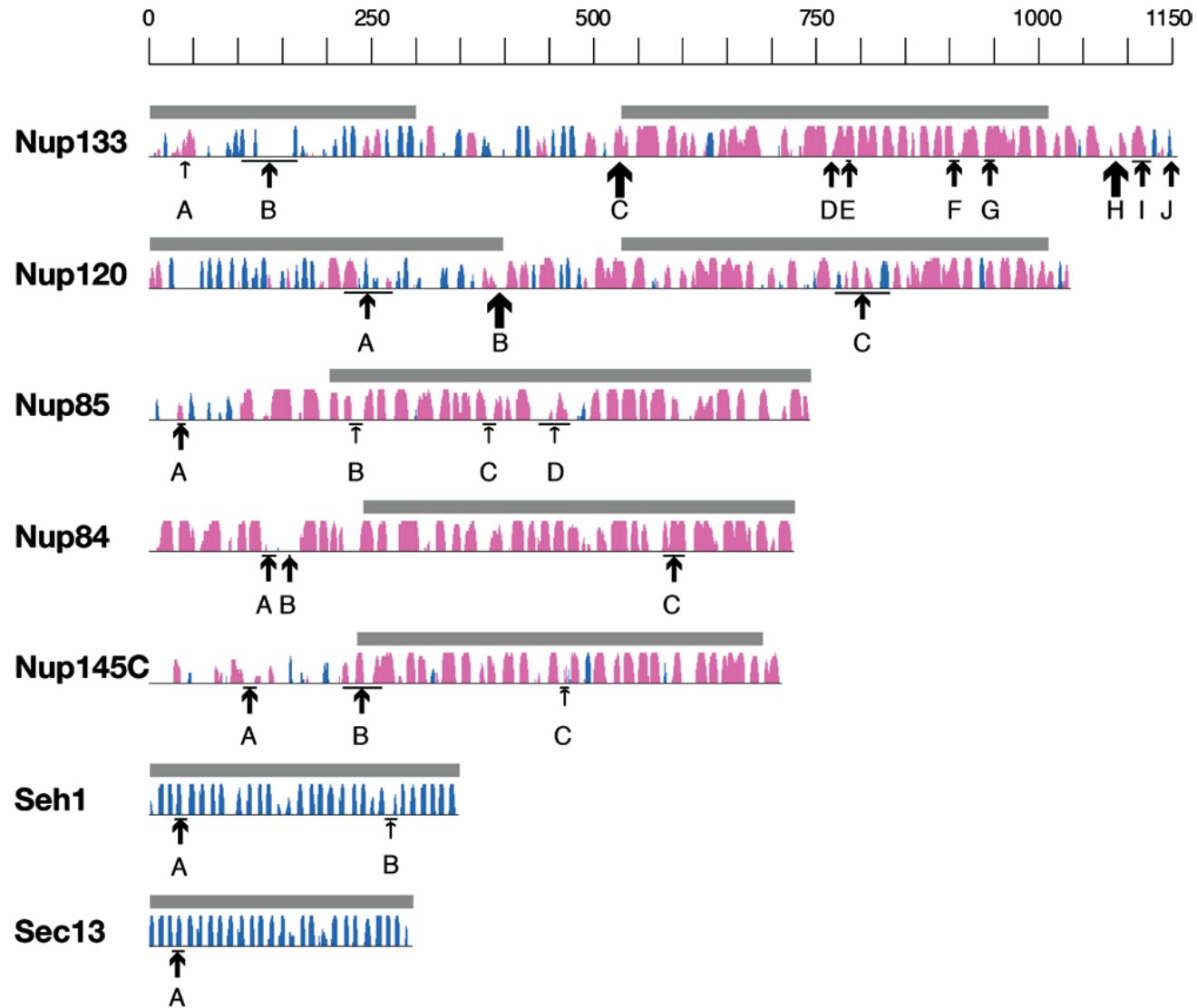
*mGenThreader + SALIGN + MOULDER*

D. Devos, S. Dokudovskaya, F. Alber, R. Williams, B.T. Chait, A. Sali, M.P. Rout.

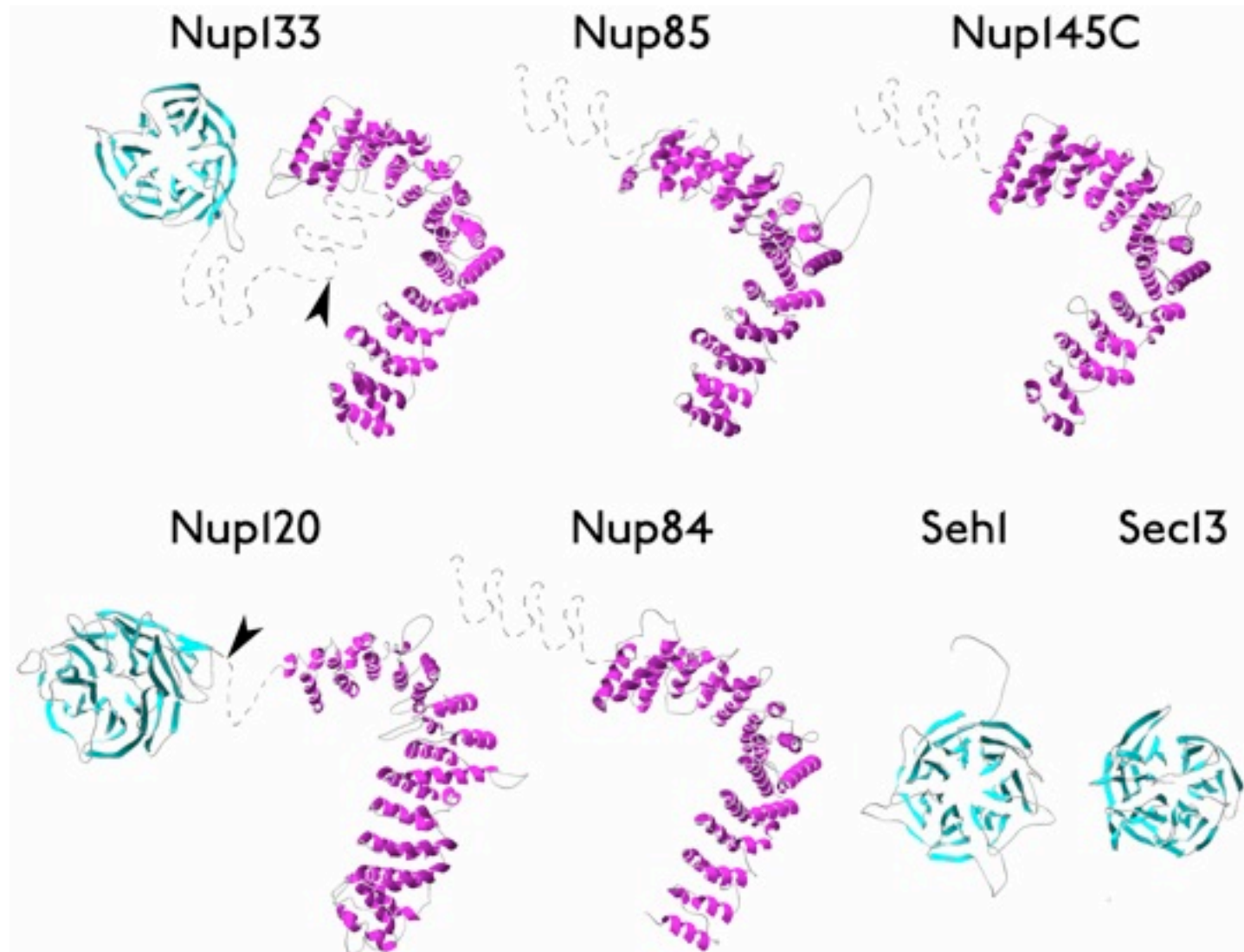
Components of Coated Vesicles and Nuclear Pore Complexes Share a Common Molecular Architecture.

*PLOS Biology* **2(12)**:e380, 2004

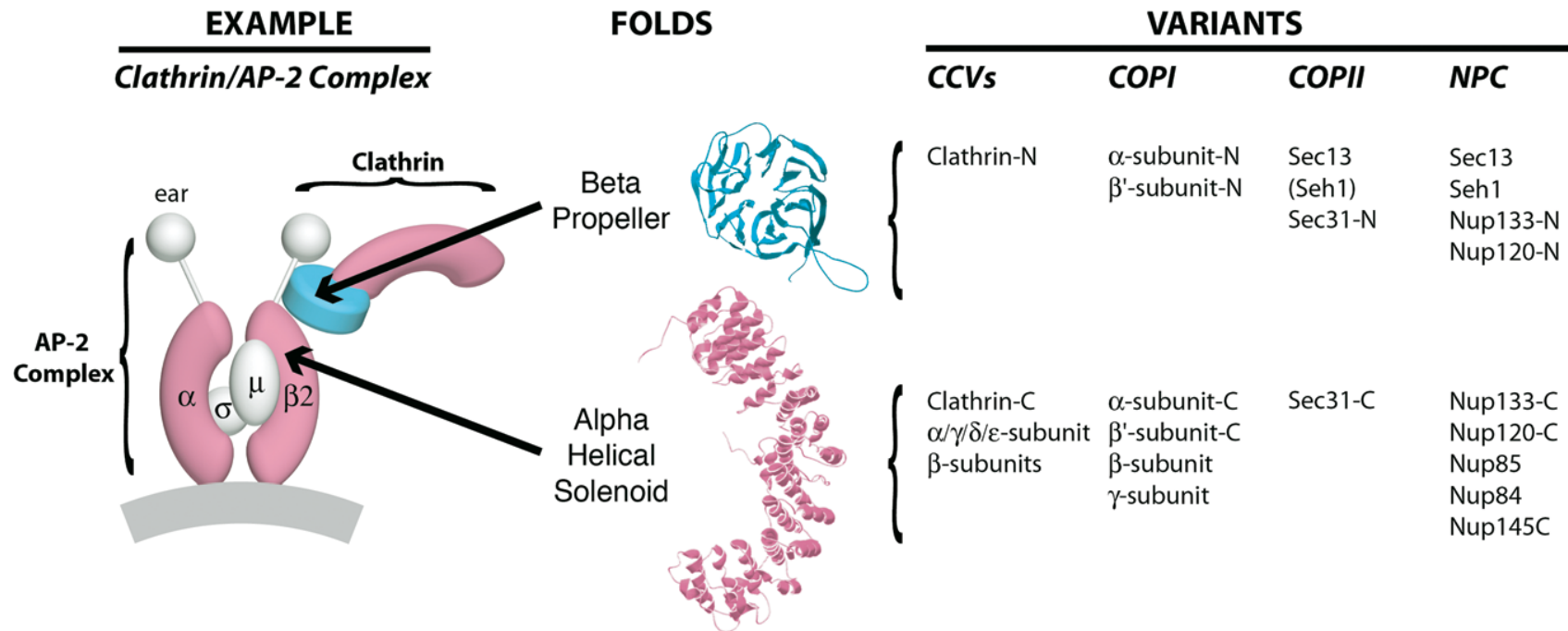
# yNup84 complex proteins



# All Nucleoporins in the Nup84 Complex are Predicted to Contain $\beta$ -Propeller and/or $\alpha$ -Solenoid Folds



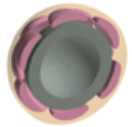
# NPC and Coated Vesicles Share the $\beta$ -Propeller and $\alpha$ -Solenoid Folds and Associate with Membranes



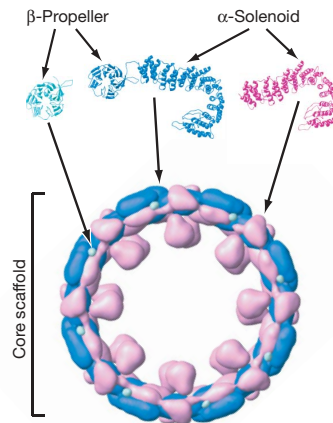
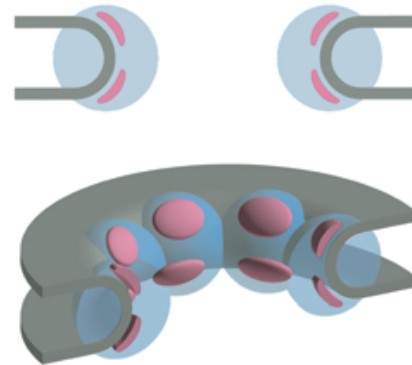


# NPC and Coated Vesicles Both Associate with Membranes

Coated Vesicle

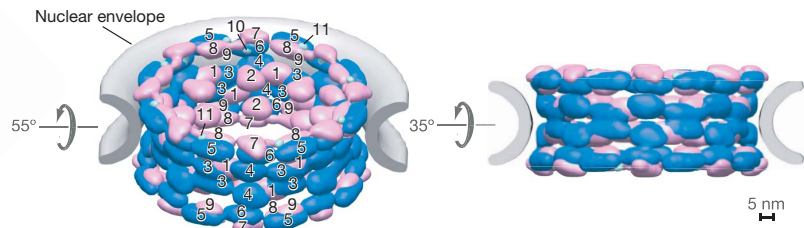


NPC model



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

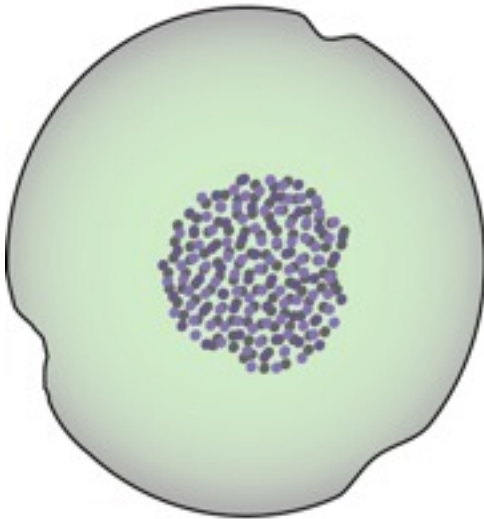


Alber et al. The molecular architecture of the nuclear pore complex. Nature (2007) vol. 450 (7170) pp. 695-701

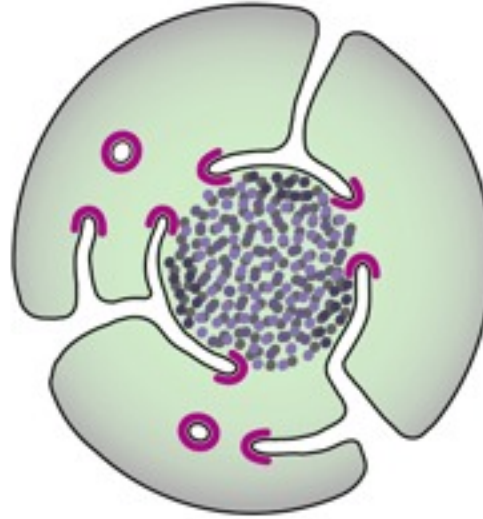
# A Common Evolutionary Origin for Nuclear Pore Complexes and Coated Vesicles?

## The proto-coatomer hypothesis

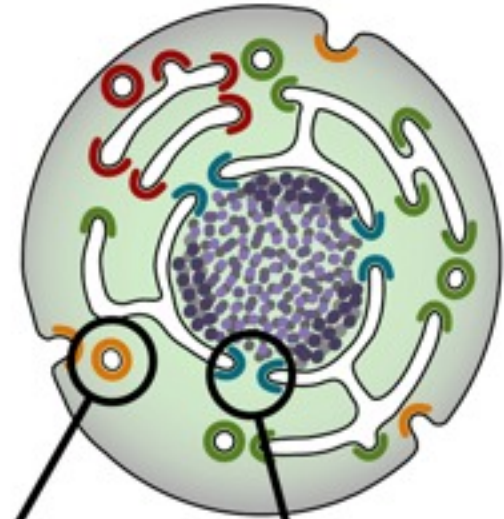
Prokaryote



Early Eukaryote

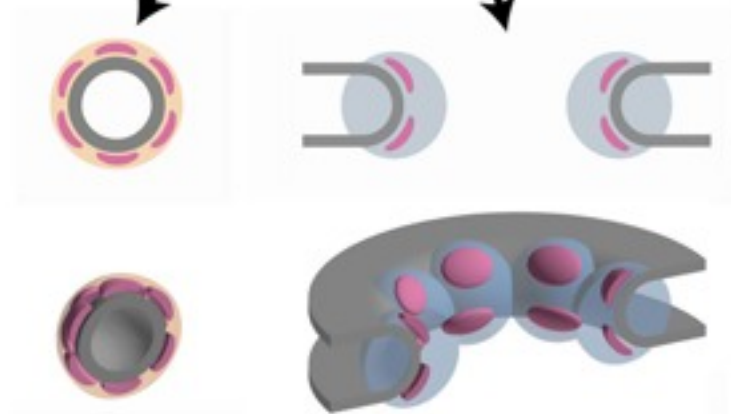


Modern Eukaryote



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

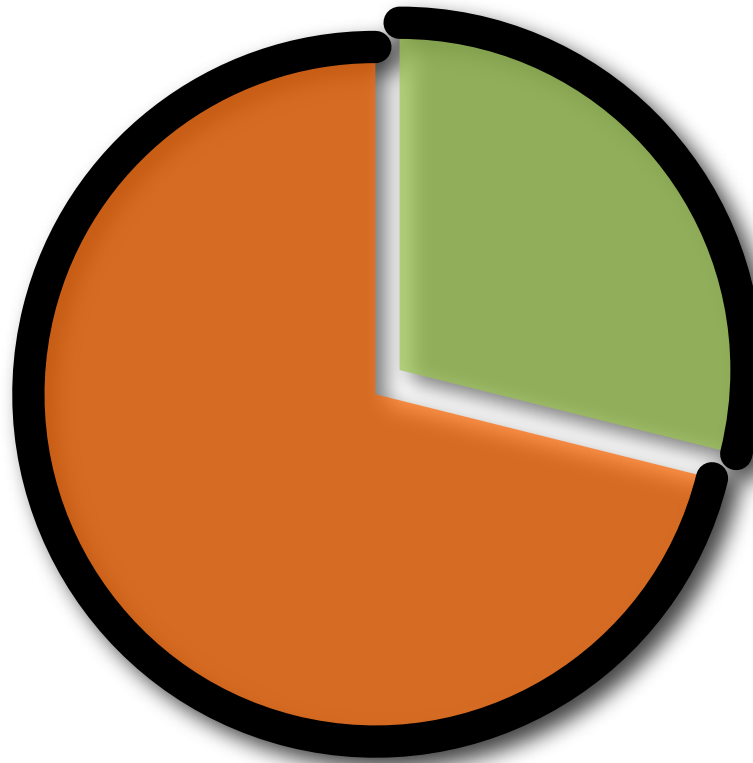
The progenitor of the NPC arose from a membrane-coating module that wrapped extensions of an early ER around the cell's chromatin.





# Tropical Disease Initiative (TDI)

*Predicting binding sites in protein structure models.*

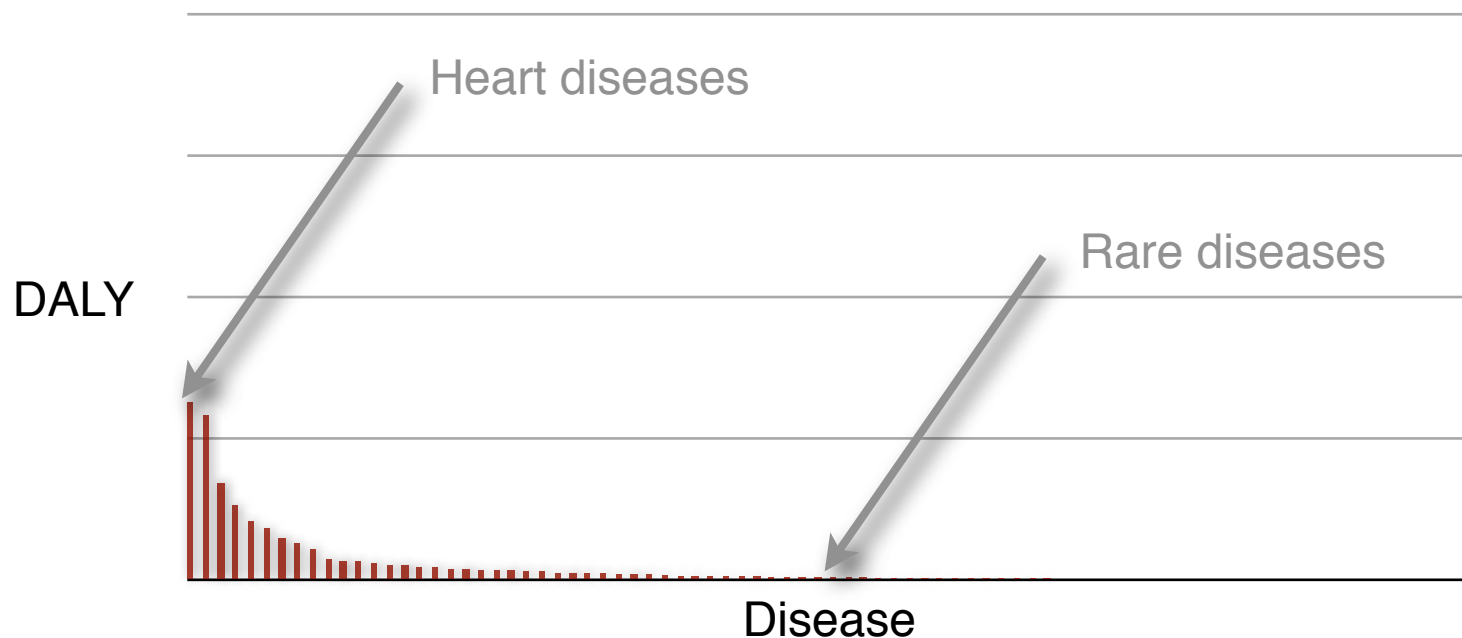


<http://www.tropicaldisease.org>



# 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*

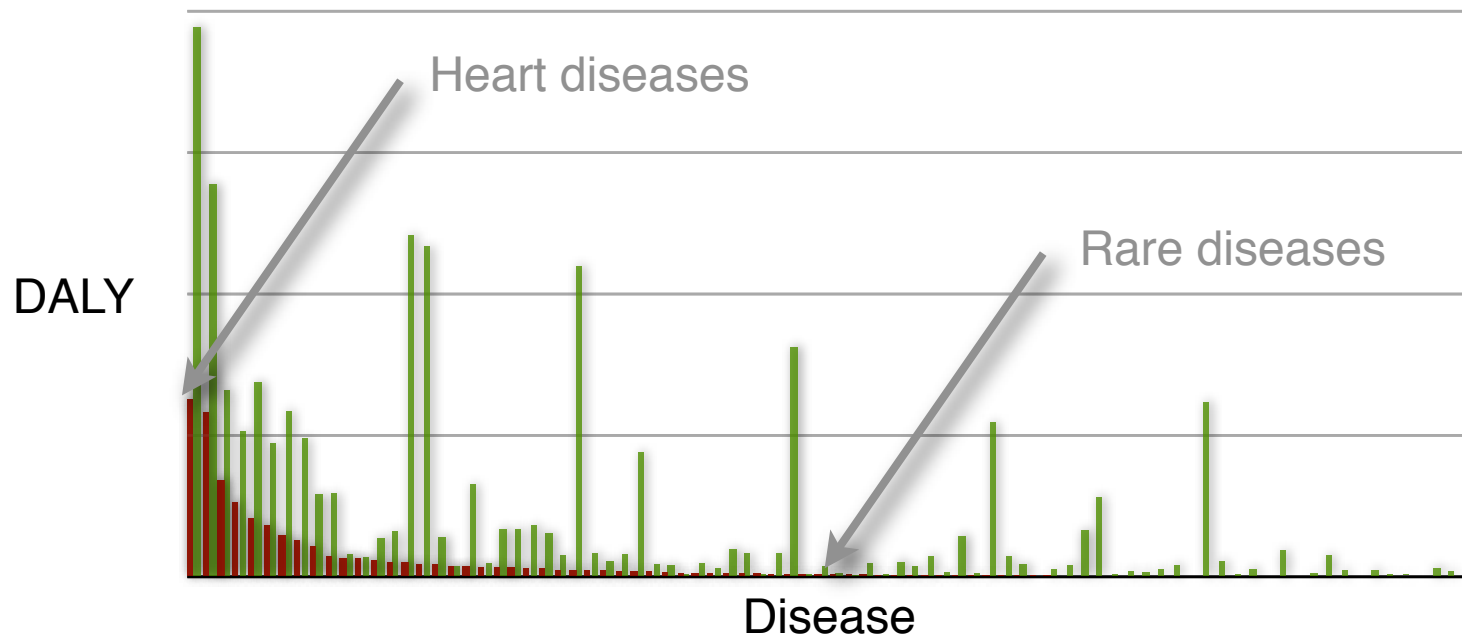
DALY - Disability adjusted life years

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

- 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.*

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

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

Trichuriasis	1,006
Japanese encephalitis	709
<b>Chagas Disease*</b>	<b>667</b>
<b>Dengue*</b>	<b>616</b>
<b>Onchocerciasis*</b>	<b>484</b>
<b>Leprosy*</b>	<b>199</b>
Diphtheria	185
Poliomyelitis	151
Hookworm disease	59

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

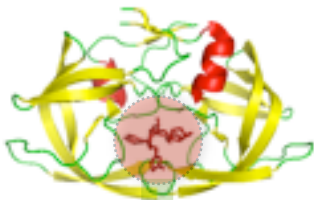
DALY - Disability adjusted life year in 1000’s.

\* Officially listed in the WHO Tropical Disease Research [disease portfolio](#).

# Comparative docking

## Expansion

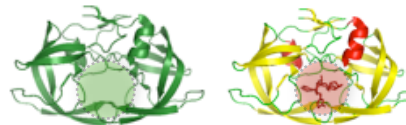
co-crystallized protein/ligand



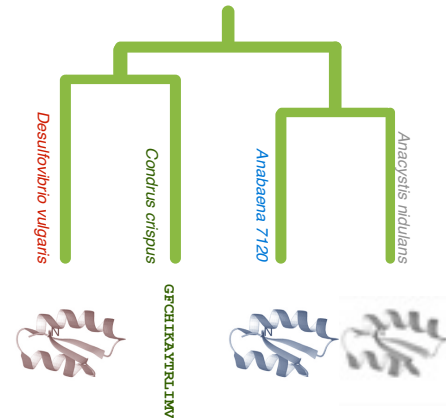
crystallized  
protein

## 2. Inheritance

model



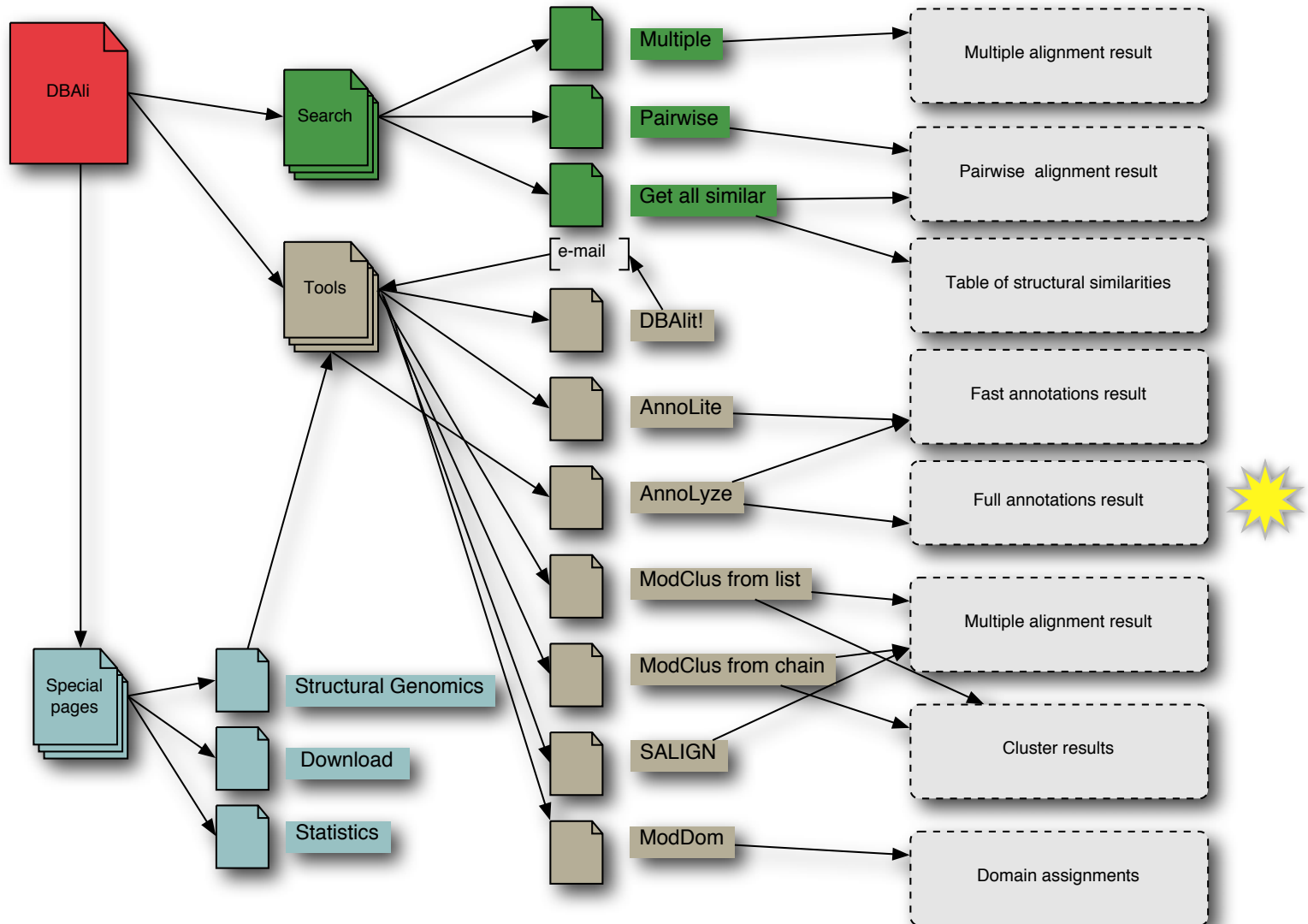
template



## 1. Modeling

# DBAli<sub>v2.0</sub> database

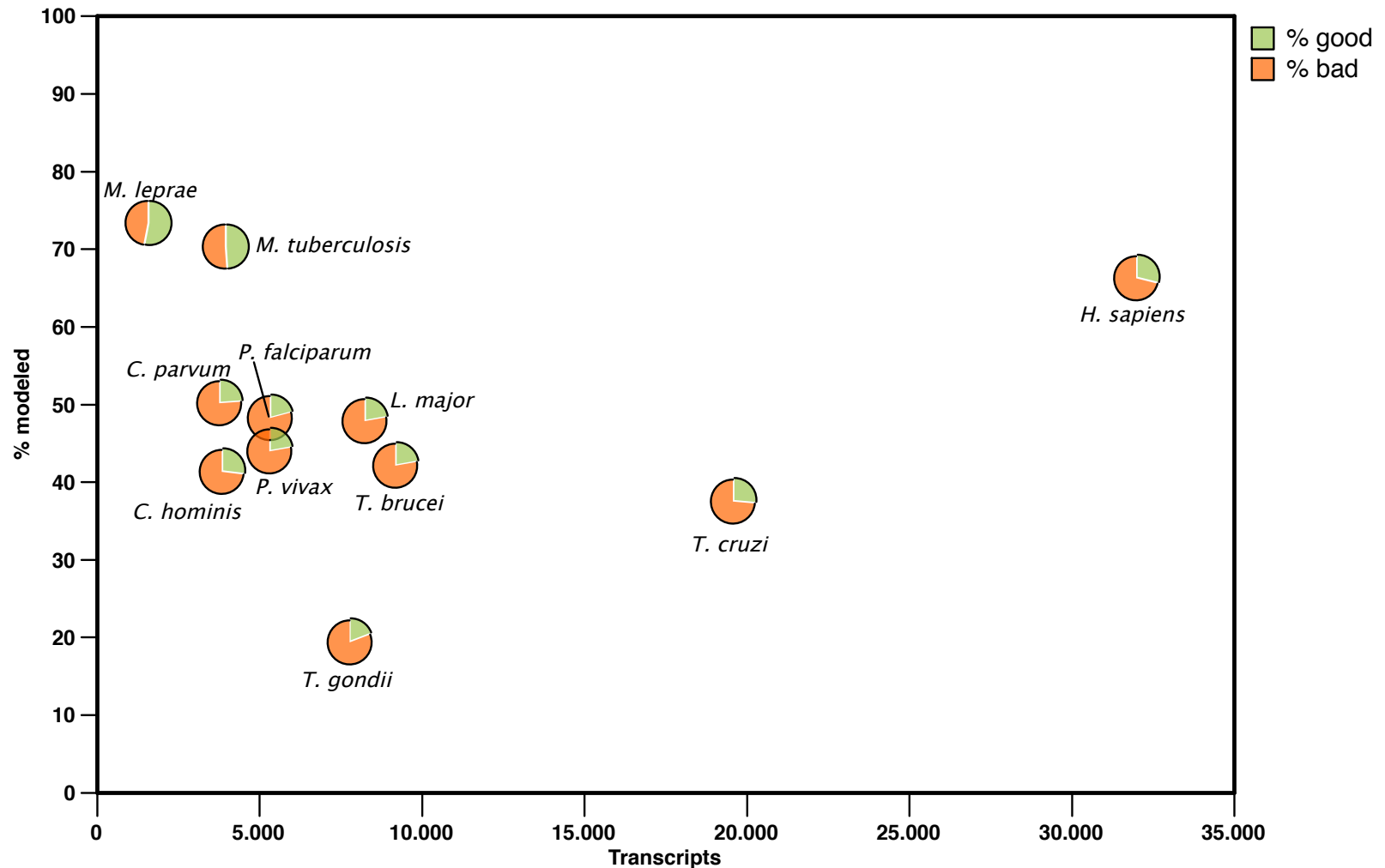
<http://www.dbali.org>



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

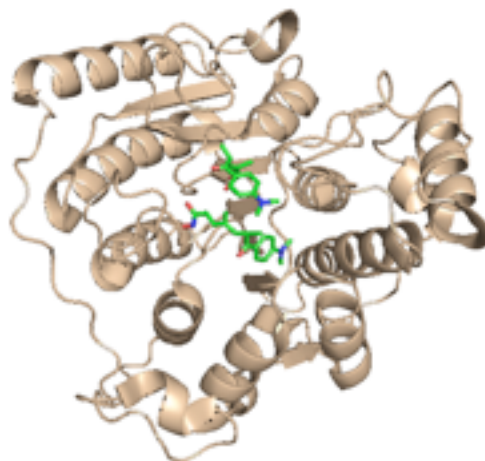
29,271 targets with good models, 297 inherited a ligand/substance similar to a known drug in DrugBank

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

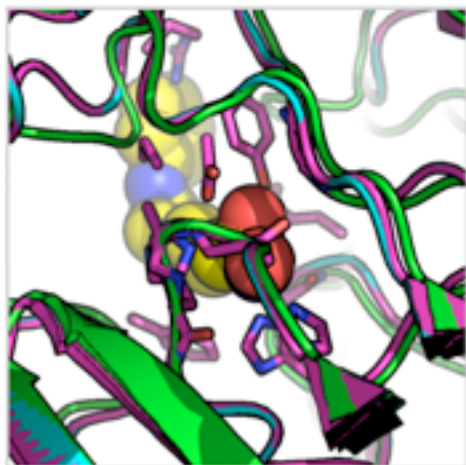


# *L. major* Histone deacetylase 2 + Vorinostat

*Template 1t64A a human HDAC8 protein.*



PDB	EO	Template	Seq	Model		Ligand	Exact	SupStr	SubStr	Similar
<a href="#">1c3sA</a>	83.33/80.00	<a href="#">1t64A</a>	36.00/1.47	<a href="#">LmjF21.0680.1.pdb</a>	90.91/100.00	<a href="#">SHH</a>	<a href="#">DB02546</a>	<a href="#">DB02546</a>	<a href="#">DB02546</a>	<a href="#">DB02546</a>



## [DB02546](#) Vorinostat

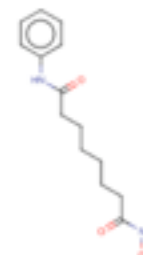
Small Molecule; Approved; Investigational

### 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\*<sup>†</sup>, ANNE M. GURNETT\*, ROBERT W. MYERS\*, PAULA M. DULSKI\*, TAMI M. CRUMLEY\*, JOHN J. ALLOCCO\*, CHRISTINE CANNOVA\*, PETER T. MEINKE<sup>‡</sup>, STEVEN L. COLLETTI<sup>‡</sup>, MARIA A. BEDNAREK<sup>‡</sup>, SHEO B. SINGH<sup>§</sup>, MICHAEL A. GOETZ<sup>§</sup>, ANNE W. DOMBROWSKI<sup>§</sup>, JON D. POLISHOOK<sup>§</sup>, AND DENNIS M. SCHMATZ\*

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

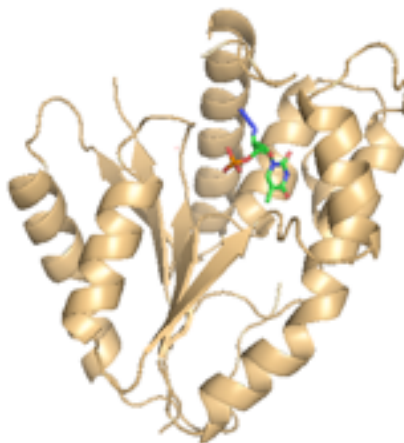
ANTIMICROBIAL AGENTS AND CHEMOTHERAPY, Apr. 2004, p. 1435–1436  
0066-4804/04/\$08.00+0 DOI: 10.1128/AAC.48.4.1435–1436.2004  
Copyright © 2004, American Society for Microbiology. All Rights Reserved.

Vol. 48, No. 4

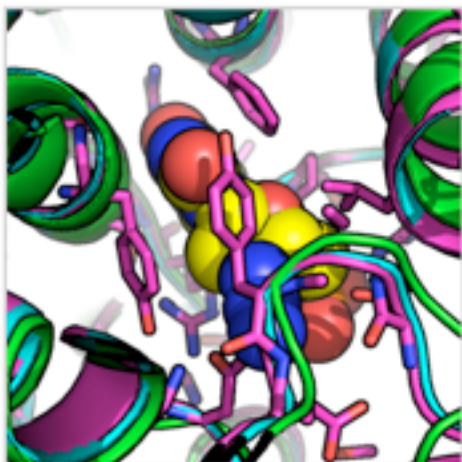
### **Antimalarial and Antileishmanial Activities of Aroyl-Pyrrolyl-Hydroxyamides, a New Class of Histone Deacetylase Inhibitors**

# *P. falciparum* thymidylate kinase + zidovudine

*Template 3tmkA a yeast thymidylate kinase.*



PDB	iQ	Template	iQ	Model	iQ	Ligand	Exact	SupStr	SubStr	Similar
<a href="#">2tmkB</a>	100.00/100.00	<a href="#">3tmkA</a>	41.00/1.49	<a href="#">PFL2465c.2.pdb</a>	82.61/100.00	<a href="#">ATM</a>		<a href="#">DB00495</a>		<a href="#">DB00495</a>



## [DB00495](#) Zidovudine

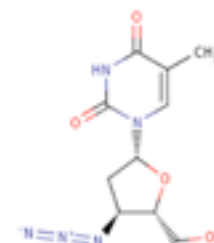
Small Molecule; Approved

### Drug categories:

Anti-HIV Agents  
Antimetabolites  
Nucleoside and Nucleotide Reverse Transcriptase Inhibitors

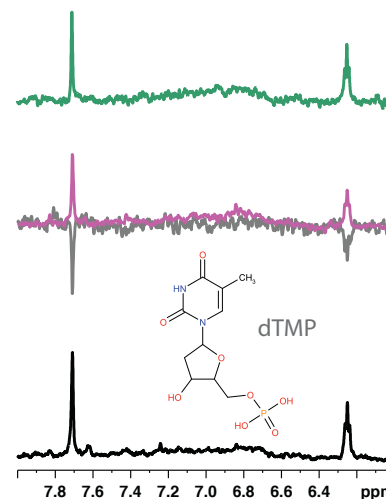
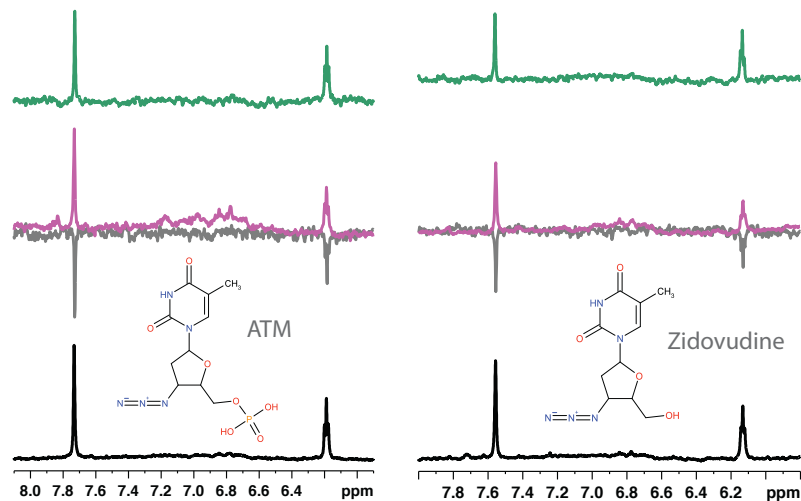
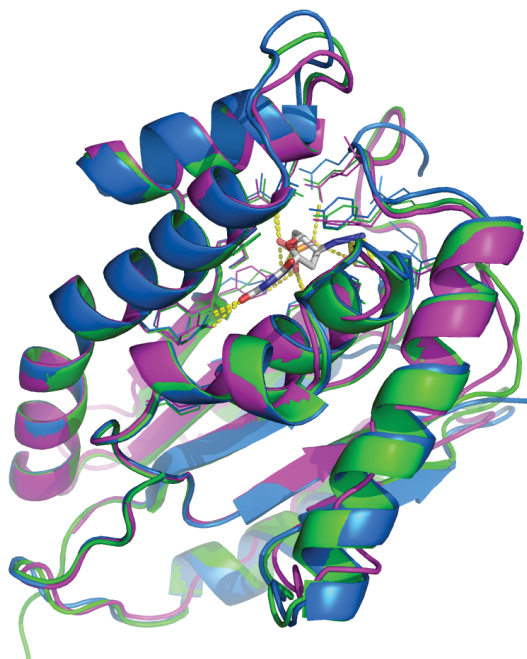
### Drug indication:

*For the treatment of human immunovirus (HIV) infections.*



# *P. falciparum* thymidylate kinase + zidovudine

NMR Water-LOGSY and STD experiments



Leticia Ortí, Rodrigo J. Carbajo, and Antonio Pineda-Lucena

# TDI's kernel

<http://tropicaldisease.org/kernel>



the **Tropical Disease Initiative**

*an open source drug discovery project*

Kernel 1.0

You are browsing version 1.0 (2008/05/01) of the TDI Kernel.

Posted on 05.07.08 to Target. 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, Q9GU59, 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	id	Template	as	Model	Ligand	Exact	SupStr	SubStr	Similar
1c3aA	85.33/90.00	1f64A	37.20/1.47	q9gu_1390.1.pdb	SHH	DB02346	DB02346	DB02346	DB02346



**DB02346** 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.*



Shown ligand [SHH](#)

OCTAMETHOXYACETAMIDE PHENYLAMIDE

expanded from [SHH](#) to template [SHH](#) used for building a 3D model of [q9gu\\_1390.1.pdb](#). Download the coordinates [q9gu\\_1390.1.pdb](#)

Search the kernel

Advanced Search

Browse the kernel

Download Q9GU59

Login / Register

Batch downloads

Help

Methods

Highest rated target:

- ATUB1 (5 out of 5)

2008 - Open Access.

Powered by WordPress.

Theme by Updat Blogger.

# TDI's kernel

## <http://tropicaldisease.org/kernel>

L. Orti *et al.*, *Nat Biotechnol* 27, 320 (Apr, 2009).

L. Orti *et al.*, *PLoS Negl Trop Dis* 3, e418 (2009).

### CORRESPONDENCE

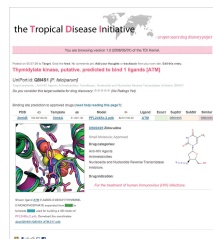
## A kernel for the Tropical Disease Initiative

### To the Editor:

Identifying proteins that are good drug targets and finding drug leads that bind to them is generally a challenging problem. It is particularly difficult for neglected tropical diseases, such as malaria and tuberculosis, where research resources are relatively scarce<sup>1</sup>. Fortunately, several developments improve our ability to deal with drug discovery for neglected diseases: first, the sequencing of many complete genomes of organisms that cause tropical diseases; second, the determination of a large number of protein structures; third, the creation of compound libraries, including already-approved drugs; and fourth, the availability of improved bioinformatics analysis, including methods for ligand identification, virtual ligand screening and drug design. Therefore, we are now in a position to increase the odds of identifying high-quality drug targets and drug leads for neglected tropical diseases. Here we encourage a collaboration among scientists to engage in drug discovery for tropical diseases by providing a kernel for the Tropical Disease Initiative (TDI, <http://www.tropicaldisease.org/>). As the Linux kernel did for open source code development, we suggest that the TDI kernel may help overcome a major stumbling block, in this case, for open source drug discovery: the absence of a critical mass of preexisting work that volunteers can build on incrementally. This kernel complements several other initiatives on neglected tropical diseases<sup>2–5</sup>, including collaborative web portals (e.g., <http://www.thetropicaldisease.org/>), public-

private partnerships (e.g., <http://www.mmm.org/>) and private foundations (e.g., <http://www.gatesfoundation.org/>); for an updated list of initiatives, see the TDI website above. The TDI kernel was developed with our software pipeline<sup>6,7</sup> for predicting structures of protein sequences by comparative modeling, localizing small-molecule binding sites on the surfaces of the models and predicting ligands that bind to them. Specifically, the pipeline linked 297 proteins from ten pathogen genomes with already approved drugs that were developed for treating other diseases (Table 1). Such links, if proven experimentally, may significantly increase the efficiency of target identification, target validation, lead discovery, lead optimization and clinical trials. Two of the kernel targets were tested for their binding to a known drug by NMR spectroscopy, validating one of our predictions (Fig. 1 and Supplementary Data online).

It is difficult to assess the accuracy of our computational predictions based on this limited experimental testing. Thus, we encourage other investigators to donate their expertise and facilities to test additional predictions. We hope the testing will occur within the



**Figure 1.** TDI kernel snapshot of the web page for the *Plasmodium falciparum* (P. falciparum) kinase target (<http://tropicaldisease.org/kernel/pfk/>). Our computational pipeline predicted that imatinib binds to P. falciparum kinase (ATM (3'-azido-3'-deoxythymidine 5'-monophosphate), a superstructure of the adenosine drug approved for the treatment of HIV infection. The binding of this ligand to a site on the kinase was experimentally validated by one-dimensional Water-LOGSY<sup>8</sup> and saturation transfer difference<sup>9,10</sup> NMR experiments.

open source context, where results are made available with limited or no restrictions.

A freely downloadable version of the TDI kernel is available in accordance with the Science Commons protocol for implementing open access data (<http://sciencecommons.org/projects/publishing/open-access-data-protocol/>), which prescribes standard academic attribution and facilitates tracking of work but imposes no other restrictions. We do not seek intellectual property rights in the actual discoveries based on the TDI kernel, in the hope of reigniting drug discovery for neglected tropical diseases<sup>11</sup>. By minimizing restrictions on the data, including vital terms that would be inherited by all derivative works, we hope to attract as many eyeballs as we possibly can to use and improve the kernel. Although many of the drugs in the kernel are proprietary under diverse types of rights, we believe that the existence of public domain parts of targets and compounds will reduce the royalties that patent owners can charge and sponsors must pay. This should decrease the large sums of money governments and

Organism	Transcript <sup>a</sup>	Modelled targets <sup>b</sup>	Similar <sup>c</sup>	Exact <sup>d</sup>
<i>Cryptosporidium hominis</i>	3,886	666	20	13
<i>Cryptosporidium parvum</i>	3,806	742	24	13
<i>Leishmania major</i>	8,274	1,409	43	20
<i>Mycobacterium leprae</i>	1,629	893	25	6
<i>Mycobacterium tuberculosis</i>	3,991	1,608	30	10
<i>Plasmodium falciparum</i>	5,363	818	28	13
<i>Plasmodium vivax</i>	5,342	822	24	13
<i>Trypanosoma brucei</i>	7,783	300	12	6
<i>Trypanosoma cruzi</i>	19,607	3,070	51	28
<i>Trypanosoma brucei</i>	9,210	1,386	39	21
Total	68,877	11,714	297	143

<sup>a</sup>Organisms in bold are included in the World Health Organization's Neglected Tropical Diseases portfolio. <sup>b</sup>Number of transcripts in each genome. <sup>c</sup>Number of targets with at least one domain accurately modeled (that is, BOPPE quality score of at least 1.0). <sup>d</sup>Number of modelled targets with at least one predicted binding site for a molecule with a Tanimoto score<sup>12</sup> of at least 0.75 to a drug in DrugBank<sup>13</sup>. <sup>e</sup>Number of modelled targets with at least one predicted binding site for a molecule in DrugBank.

OPEN ACCESS Freely available online



## A Kernel for Open Source Drug Discovery in Tropical Diseases

Leticia Orti<sup>1,2</sup>, Rodrigo J. Carbajo<sup>3</sup>, Ursula Pieper<sup>4</sup>, Narayanan Eswar<sup>5,6</sup>, Stephen M. Maurer<sup>7</sup>, Arti K. Rai<sup>8</sup>, Ginger Taylor<sup>9</sup>, Matthew H. Todd<sup>1</sup>, Antonio Pineda-Lucena<sup>2</sup>, Andrej Sali<sup>10</sup>, Marc A. Marti-Renom<sup>1</sup>

<sup>1</sup>Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain, <sup>2</sup>Structural Biology Laboratory, Molecular Chemistry Department, Centro de Investigación Príncipe Felipe, Valencia, Spain, <sup>3</sup>Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California San Francisco, San Francisco, California, United States of America, <sup>4</sup>Goldilocks of Law, University of Southern California, Los Angeles, California, United States of America, <sup>5</sup>School of Law, DePaul University, Durham, North Carolina, United States of America, <sup>6</sup>The Synaptic Leap, San Ramon, California, United States of America, <sup>7</sup>School of Chemistry, University of Sydney, Sydney, New South Wales, Australia

### Abstract

**Background:** Conventional patent-based drug development incentives work badly for the developing world, where commercial markets are usually small to non-existent. For this reason, the past decade has seen extensive experimentation with alternative R&D institutions ranging from private-public partnerships to development prizes. Despite extensive discussion, however, one of the most promising avenues—open source drug discovery—has remained elusive. We argue that the stumbling block has been the absence of a critical mass of preexisting work that volunteers can improve through a series of granular contributions. Historically, open source software collaborations have almost never succeeded without such “kernels”.

**Methodology/Principal Findings:** Here, we use a computational pipeline for: (i) comparative structure modeling of target proteins; (ii) predicting the localization of ligand binding sites on their surfaces; and (iii) assessing the similarity of the predicted ligands to known drugs. Our kernel currently contains 143 and 297 protein targets from ten pathogen genomes that are predicted to bind a known drug or a molecule similar to a known drug, respectively. The kernel provides a source of potential drug targets and drug candidates around which an online open source community can nucleate. Using NMR spectroscopy, we have experimentally tested our predictions for two of these targets, confirming one and invalidating the other.

**Conclusions/Significance:** The TDI kernel, which is being offered under the Creative Commons attribution share-alike license for free and unrestricted use, can be accessed on the World Wide Web at <http://www.tropicaldisease.org/>. We hope that the kernel will facilitate collaborative efforts towards the discovery of new drugs against parasites that cause tropical diseases.

**Citation:** Orti L, Carbajo RJ, Pieper U, Eswar N, Maurer SM, et al. (2009) A Kernel for Open Source Drug Discovery in Tropical Diseases. *PLoS Negl Trop Dis* 3(4): e418. doi:10.1371/journal.pntd.0000418

**Editor:** Timothy G. Geary, McGill University, Canada

**Received:** December 29, 2008; **Accepted:** March 23, 2009; **Published:** April 21, 2009

**Copyright:** © 2009 Orti et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** MAM-R acknowledges the support from a Spanish Ministerio de Educación y Ciencia grant (BG2007/66670). AS acknowledges the support from the Sanchez Family Supporting Foundation and the National Institutes of Health (NIH) GM63624 (U.S.), GM63624 (U.S.), GM63624 (U.S.), and GM63624 (U.S.). We also acknowledge the support from a Spanish Ministerio de Ciencia e Innovación grant (SAF2008-01845). RJC acknowledges the support from the Ramon y Cajal Program of the Spanish Ministerio de Educación y Ciencia. We are also grateful for computer hardware gifts to AS from Ron Conway, Mike Hammer, Intel, IBM, Hewlett-Packard, and NetApp. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

<sup>1</sup> E-mail: [carbajo@princefe.org](mailto:carbajo@princefe.org) (RJC), [eswar@synapticleap.com](mailto:eswar@synapticleap.com) (NE).

<sup>2</sup> Current address: Daphne Knowledge Center, Hyderabad, India

### Introduction

There is a lack of high-quality protein drug targets and drug leads for neglected diseases [1–2]. Fortunately, many genomes of organisms that cause tropical diseases have already been sequenced and posted. Therefore, we are now in a position to leverage this information by identifying potential protein targets for drug discovery. Atomic-resolution structures can facilitate this task.

In the absence of an experimentally determined structure, comparative modeling can provide useful models for sequences that are distantly related to known protein structures [3,4]. Approximately half of known protein sequences contain domains that can be currently predicted by comparative modeling [5,6]. This coverage

will increase as the number of experimentally determined structures grows and modeling software improves. A protein model can facilitate at least four important tasks in the early stages of drug discovery [7]: prioritizing protein targets for drug discovery [8], identifying binding sites for small molecules [9,10], suggesting drug leads [11,12], and optimizing these leads [13–15].

Here, we address the first three tasks by assembling our computer programs into a software pipeline that automatically and on large-scale predicts protein structures, their ligand binding sites, and known drugs that interact with them. As a proof of principle, we applied the pipeline to the genomes of ten organisms that cause tropical diseases (“target genomes”). We also experimentally tested two predicted drug-target interactions using Nuclear Magnetic

320

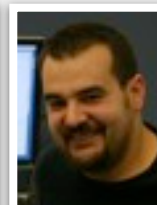
VOLUME 27 | NUMBER 4 | APRIL 2009 | NATURE BIOTECHNOLOGY

# Acknowledgments

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

<http://tropicaldisease.org>

<http://integrativemodeling.org>



## COMPARATIVE MODELING

**Andrej Sali**

M. S. Madhusudhan

**Narayanan Eswar**

Min-Yi Shen

**Ursula Pieper**

Ben Webb

Maya Topf (Birbeck College)

## MODEL ASSESSMENT

Francisco Melo (CU)

Alejandro Panjkovich (CU)

## NMR

**Antonio Pineda-Lucena**

**Leticia Ortí**

**Rodrigo J. Carbajo**

## MAMMOTH

**Angel R. Ortiz**

## FUNCTIONAL ANNOTATION

**Fatima Al-Shahrour**

**Joaquin Dopazo**

## BIOLOGY

Jeff Friedman (RU)

James Hudsped (RU)

Partho Ghosh (UCSD)

Alvaro Monteiro (Cornell U)

Stephen Krilis (St. George H)

## Tropical Disease Initiative

**Stephen Maurer (UC Berkeley)**

**Arti Rai (Duke U)**

**Andrej Sali (UCSF)**

**Ginger Taylor (TSL)**

**Matthew Todd (U Sydney)**

## CCPR Functional Proteomics

Patsy Babbitt (UCSF)

Fred Cohen (UCSF)

Ken Dill (UCSF)

Tom Ferrin (UCSF)

John Irwin (UCSF)

Matt Jacobson (UCSF)

Tack Kuntz (UCSF)

Andrej Sali (UCSF)

Brian Shoichet (UCSF)

Chris Voigt (UCSF)

## EVA

Burkhard Rost (Columbia U)

Alfonso Valencia (CNB/UAM)

## CAMP

Xavier Aviles (UAB)

Hans-Peter Nester (SANOFI)

Ernst Meinjohanns (ARPIDA)

Boris Turk (IJS)

Markus Gruetter (UE)

Matthias Wilmanns (EMBL)

Wolfram Bode (MPG)

## MODEL ASSESSMENT

David Eramian

Min-Yi Shen

Damien Devos

## FUNCTIONAL ANNOTATION

Andrea Rossi (Rinat-Pfizer)

Fred Davis (Janelia Fram)

## FUNDING

Prince Felipe Research Center

**Ministerio de Educación y Ciencia**

STREP UE Grant

Marie Curie Reintegration Grant