Comparative Protein Structure Prediction

Marc A. Marti-Renom
http://bioinfo.cipf.es/sgu/

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

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

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

Summary

- INTRO
- Structural Space
- Profile-Profile & MOULDER
- Function from models

- MODELLER tutorial/example

Nomenclature

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

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

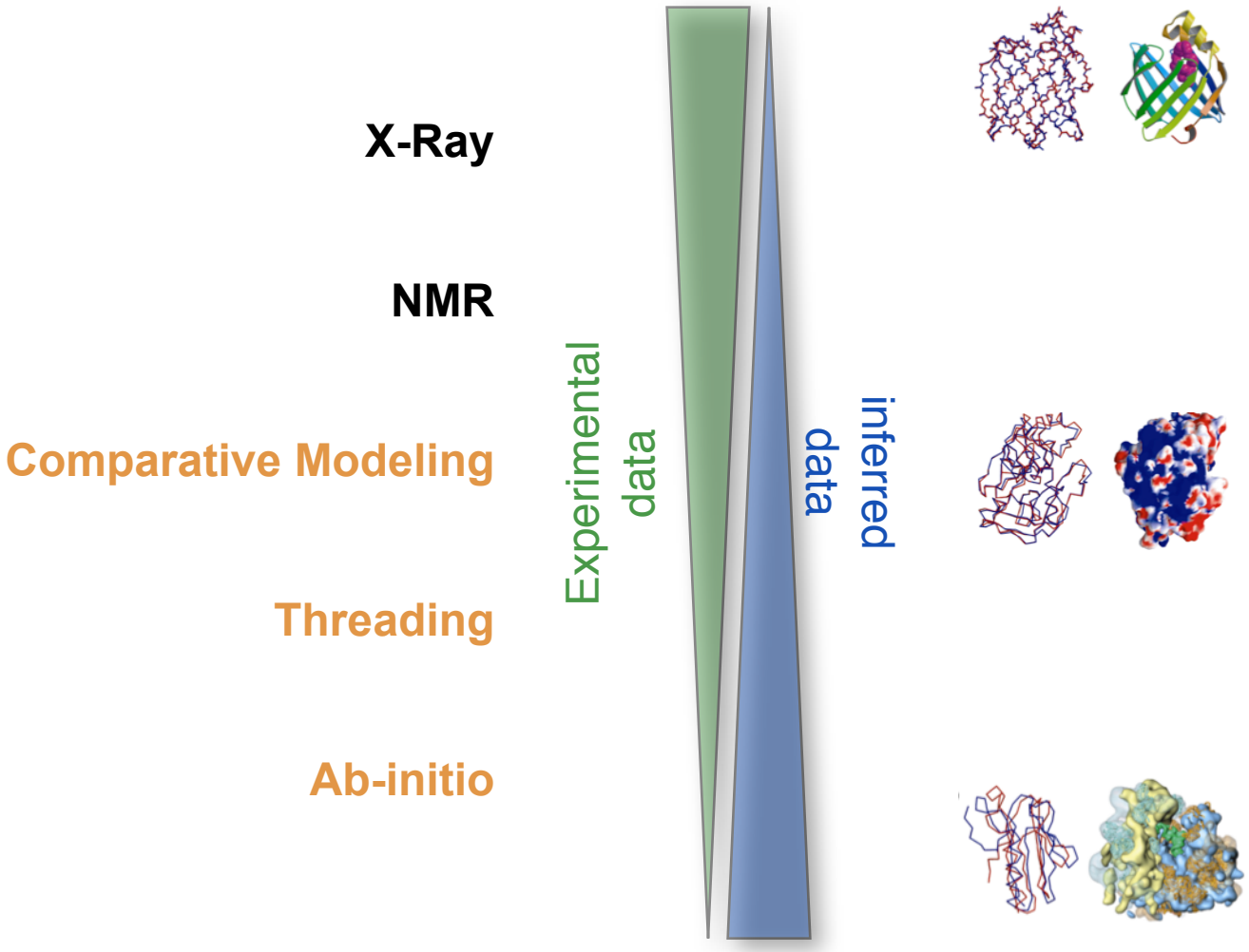
Identity: Fraction of identical aminoacids between two aligned sequences (case of similarity).

Target: Sequence corresponding to the protein to be modeled.

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

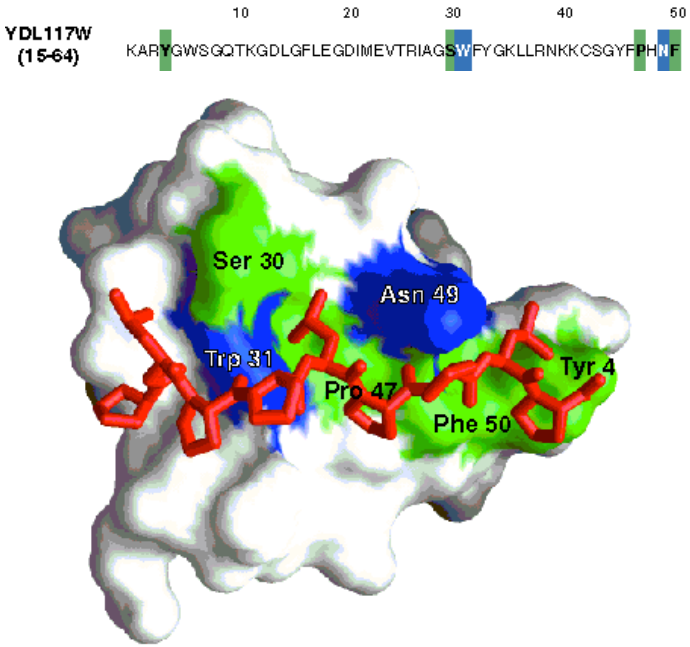
Model: Predicted 3D structure of the target sequence.

protein prediction .vs. protein determination



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

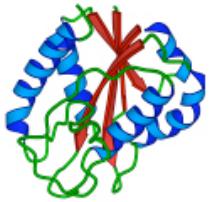
- The biochemical function (activity) of a protein is defined by its interactions with other molecules.
- The biological function is in large part a consequence of these interactions.
- The 3D structure is more informative than sequence because interactions are determined by residues that are close in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



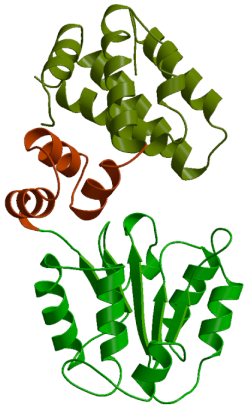
In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence**.

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

From domains to assemblies

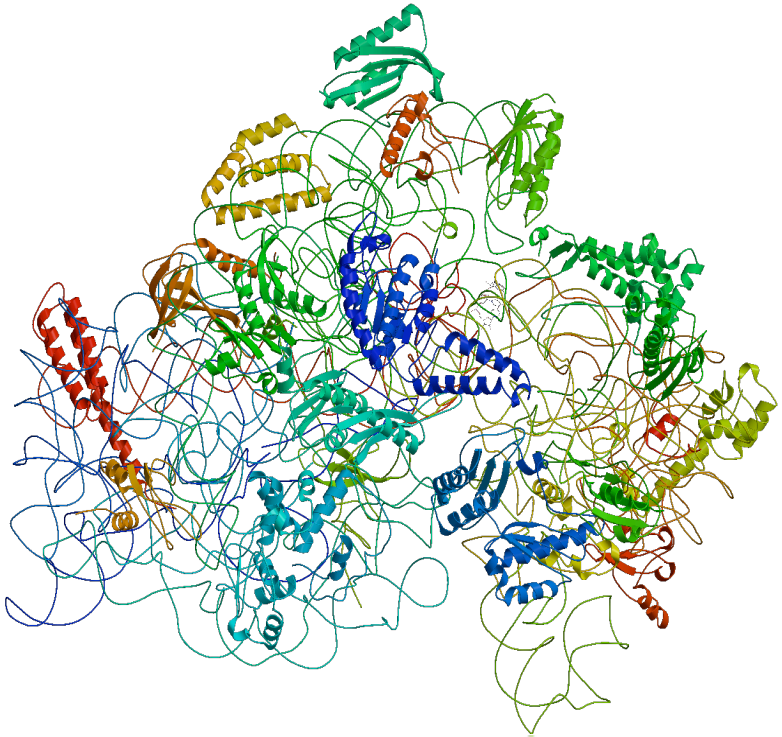


domains



proteins

assemblies

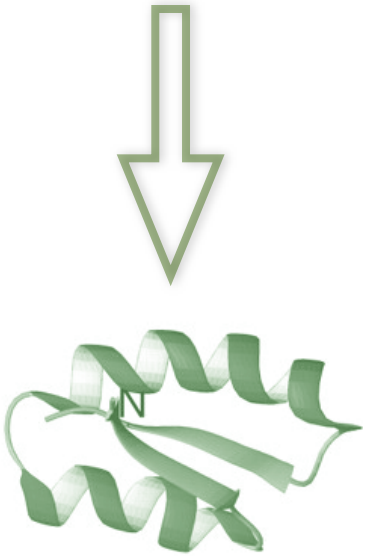


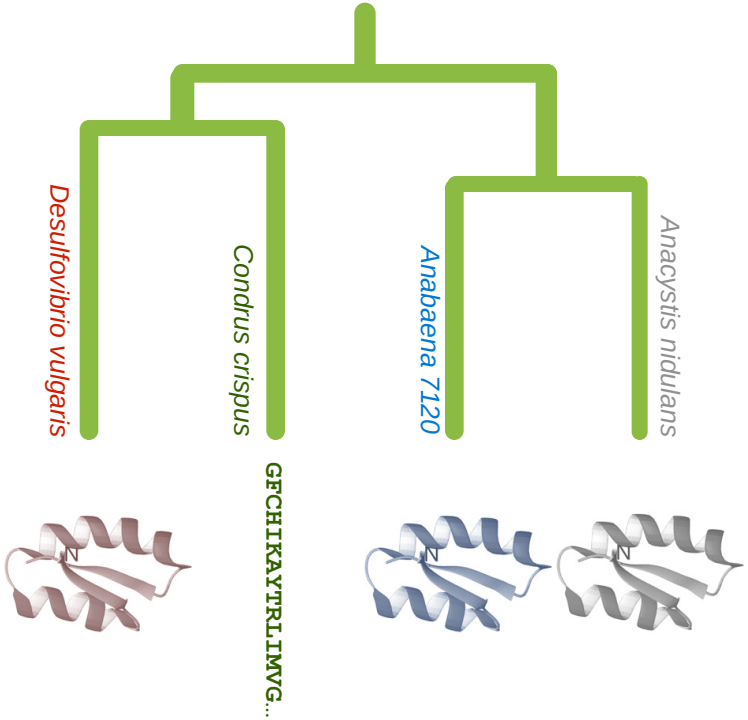
~2.5 domains in a protein
a few domain partners per domain

Russell etal. Curr Opin Struct Biol 14, 313, 2004.

Principles of protein structure

GFCHIKAYTRLIMVG...





Folding (physics)

Ab initio prediction

Evolution (rules)

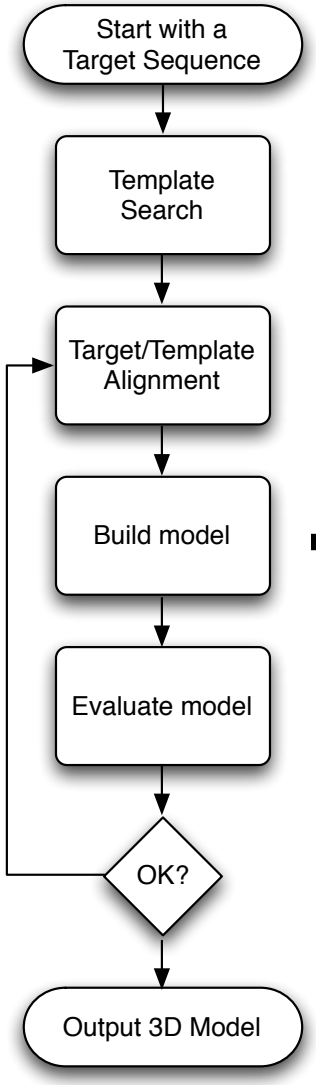
Threading

Comparative Modeling

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

Comparative modeling by satisfaction of spatial restraints

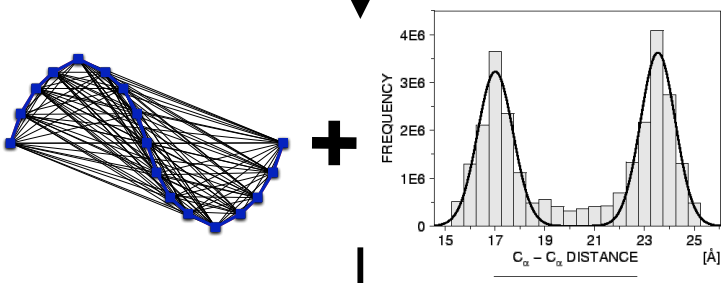
MODELLER



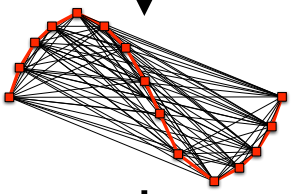
Given an alignment...

MSVIPKR--GNCEQTSE
ASILPKRLFGNCEQTSD

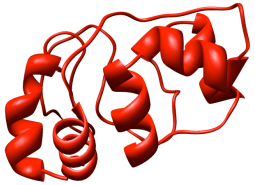
extract spatial features from the template(s) and statistics from known structures



apply these features as restraints on your target sequence



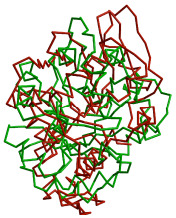
optimize to find the best solution for the restraints to produce your 3D model



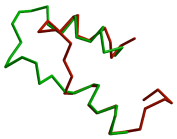
A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.

Comparative modeling by satisfaction of spatial restraints

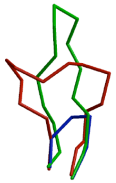
Types of errors and their impact



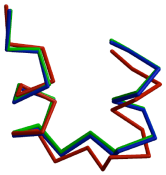
Wrong fold



Miss alignments

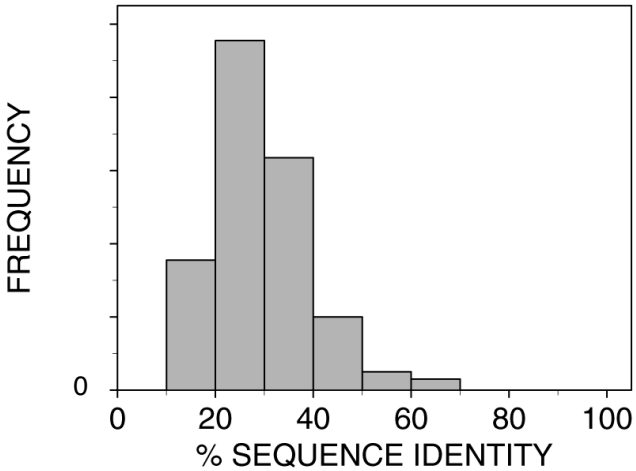


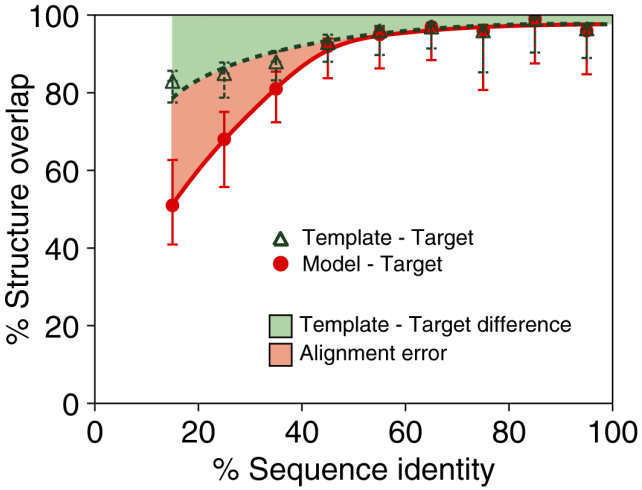
Loop regions



Rigid body distortions

Side-chain packing





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

Model Accuracy

HIGH ACCURACY

NM23
Seq id 77%

Cα equiv 147/148
RMSD 0.41Å

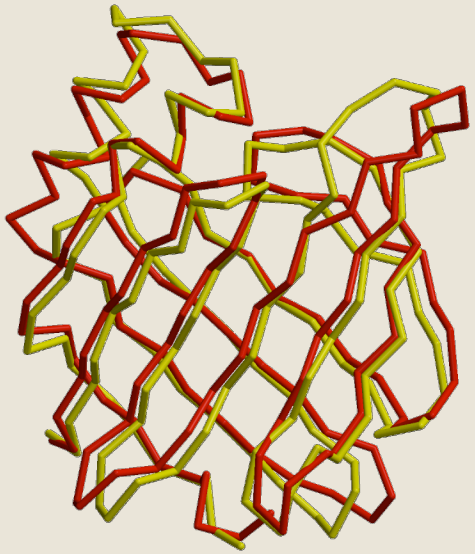
Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

CRABP
Seq id 41%

Cα equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

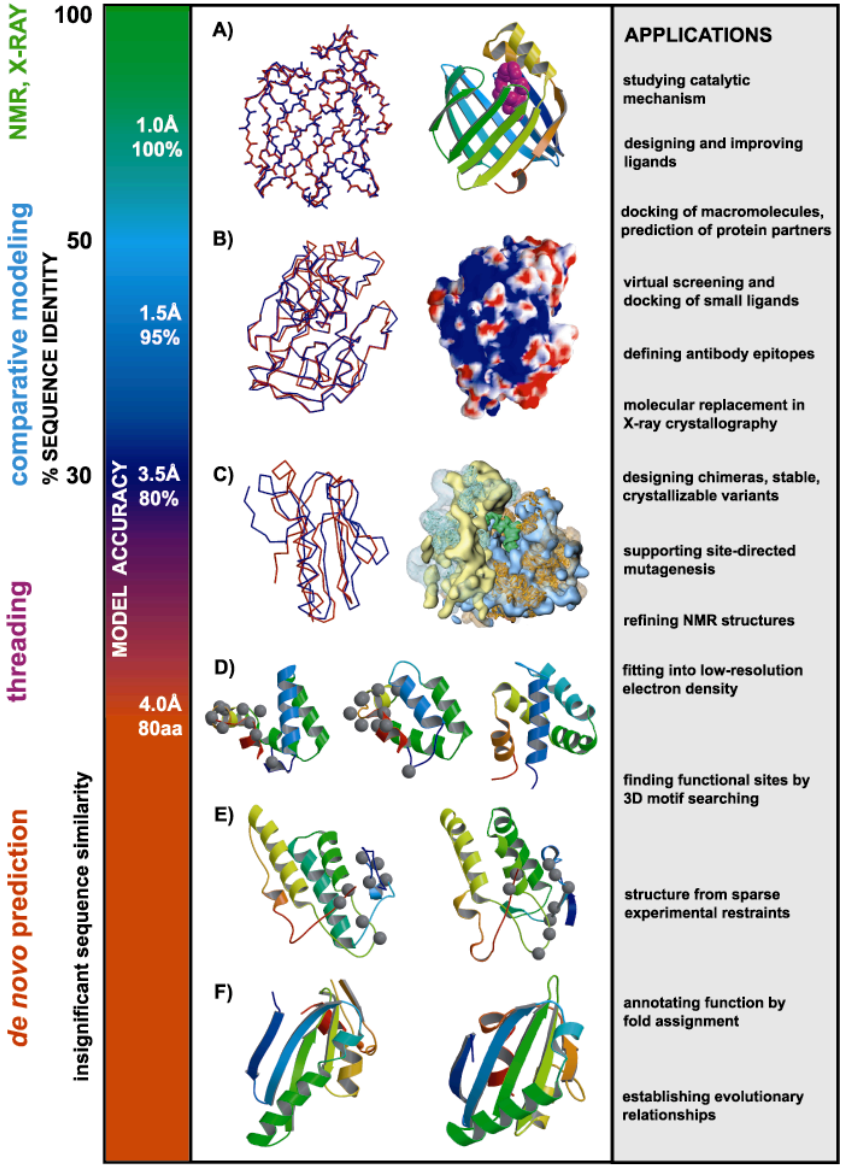
EDN
Seq id 33%

Cα equiv 90/134
RMSD 1.17Å

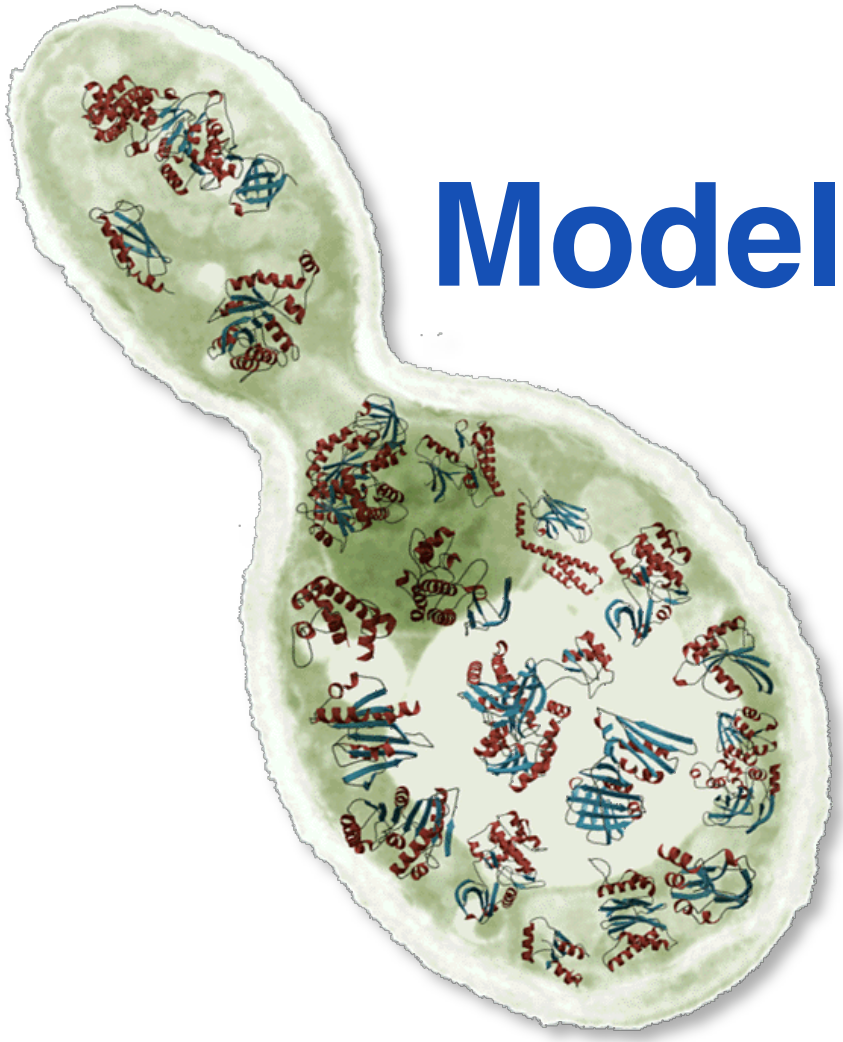
Sidechains
Core backbone
Loops
Alignment
Fold assignment

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.

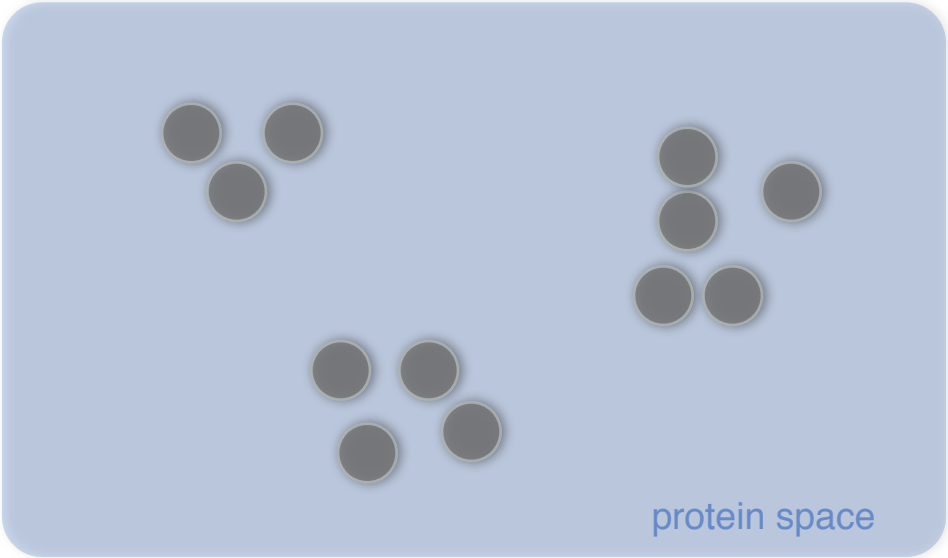


Modeling genomes

Structural Genomics

Characterize most protein sequences based on related known structures

1. The number of "**families**" is much **smaller** than the number of proteins.
2. **Any one** of the members of a family is **fine**.



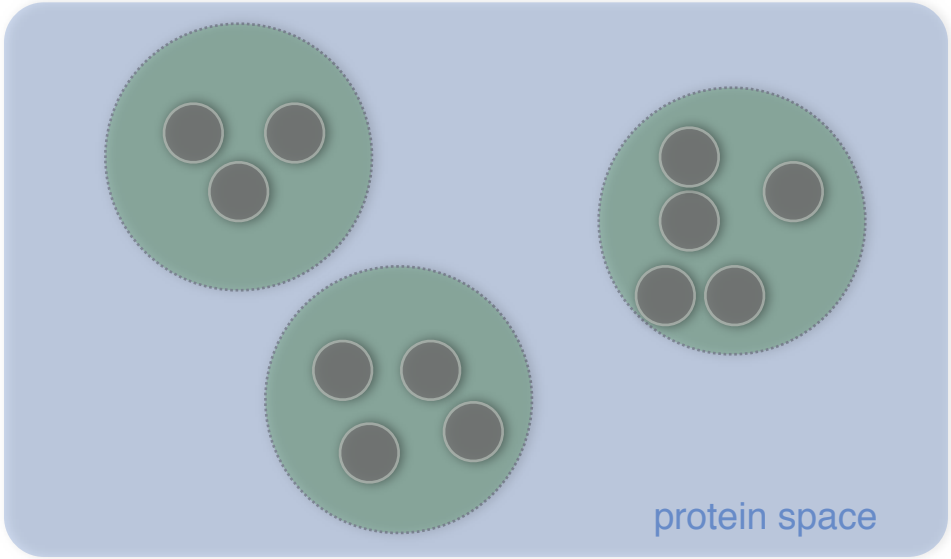
There are **~16,000** families (90%)
@ 30% sequence identity cutoff

Sali. Nat. Struct. Biol. ***5****, 1029, 1998.*
Sali et al. Nat. Struct. Biol., ***7****, 986, 2000.*
Sali. Nat. Struct. Biol. ***7****, 484, 2001.*
Baker & Sali. Science ***294,*** *93, 2001.*
Vitkup et al. Nat. Struct. Biol. ***8****, 559, 2001*

Structural Genomics

Characterize most protein sequences based on related known structures

1. The number of "**families**" is much **smaller** than the number of proteins.
2. **Any one** of the members of a family is **fine**.



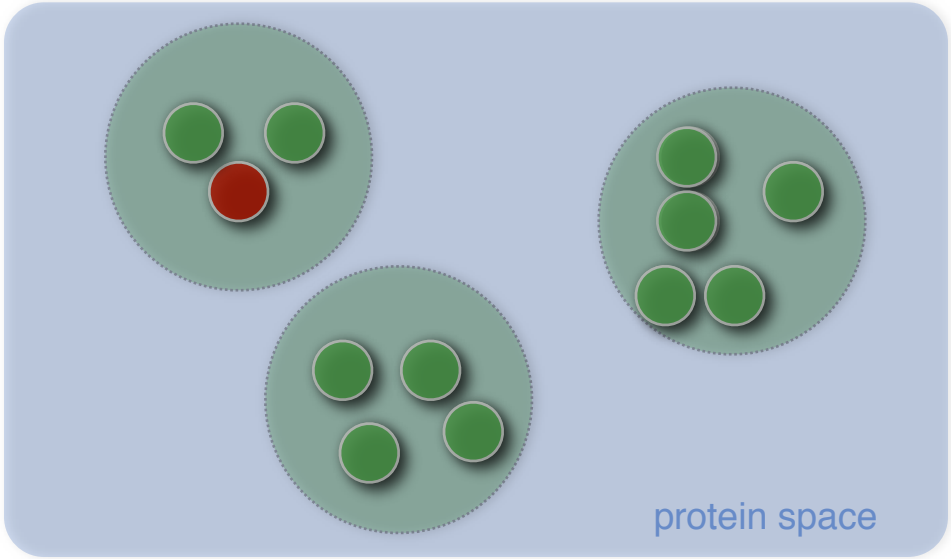
There are **~16,000** families (90%)
@ 30% sequence identity cutoff

Sali. Nat. Struct. Biol. ***5****, 1029, 1998.*
Sali et al. Nat. Struct. Biol., ***7****, 986, 2000.*
Sali. Nat. Struct. Biol. ***7****, 484, 2001.*
Baker & Sali. Science ***294,*** *93, 2001.*
Vitkup et al. Nat. Struct. Biol. ***8****, 559, 2001*

Structural Genomics

Characterize most protein sequences based on related known structures

1. The number of "**families**" is much **smaller** than the number of proteins.
2. **Any one** of the members of a family is **fine**.



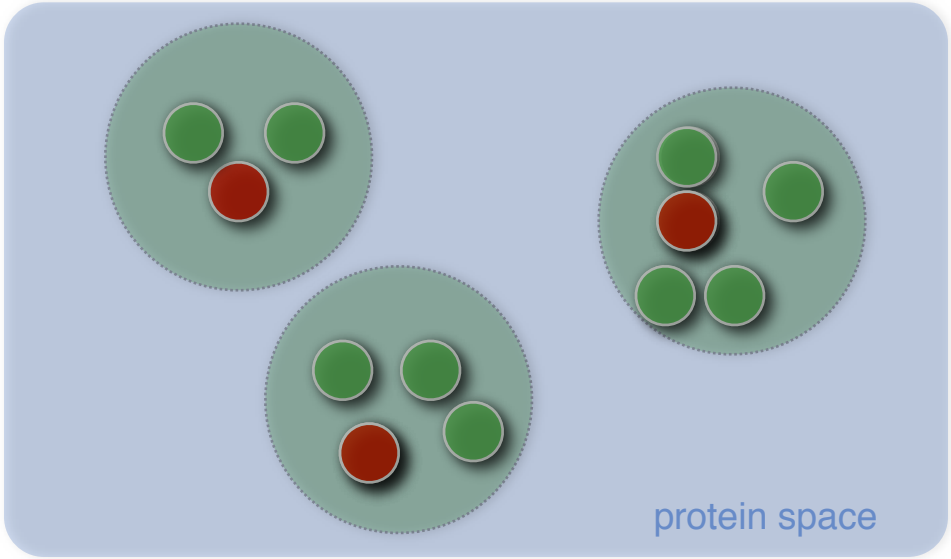
There are **~16,000** families (90%)
@ 30% sequence identity cutoff

Sali. Nat. Struct. Biol. ***5****, 1029, 1998.*
Sali et al. Nat. Struct. Biol., ***7****, 986, 2000.*
Sali. Nat. Struct. Biol. ***7****, 484, 2001.*
Baker & Sali. Science ***294,*** *93, 2001.*
Vitkup et al. Nat. Struct. Biol. ***8****, 559, 2001*

Structural Genomics

Characterize most protein sequences based on related known structures

1. The number of "**families**" is much **smaller** than the number of proteins.
2. **Any one** of the members of a family is **fine**.

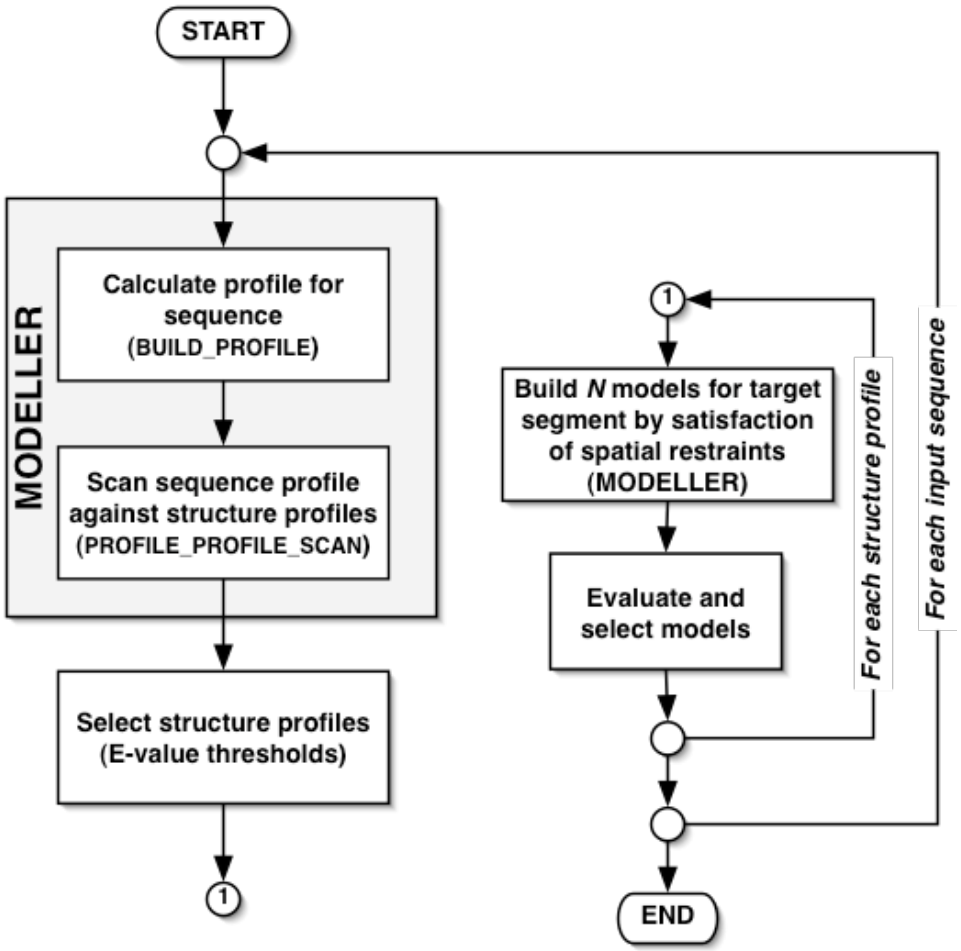


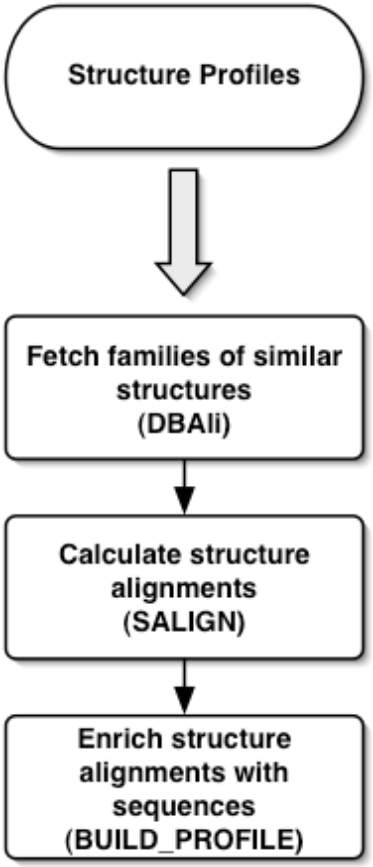
There are **~16,000** families (90%)
@ 30% sequence identity cutoff

Sali. Nat. Struct. Biol. ***5****, 1029, 1998.*
Sali et al. Nat. Struct. Biol., ***7****, 986, 2000.*
Sali. Nat. Struct. Biol. ***7****, 484, 2001.*
Baker & Sali. Science ***294,*** *93, 2001.*
Vitkup et al. Nat. Struct. Biol. ***8****, 559, 2001*

MODPIPE2.0

Large-Scale Protein Structure Modeling





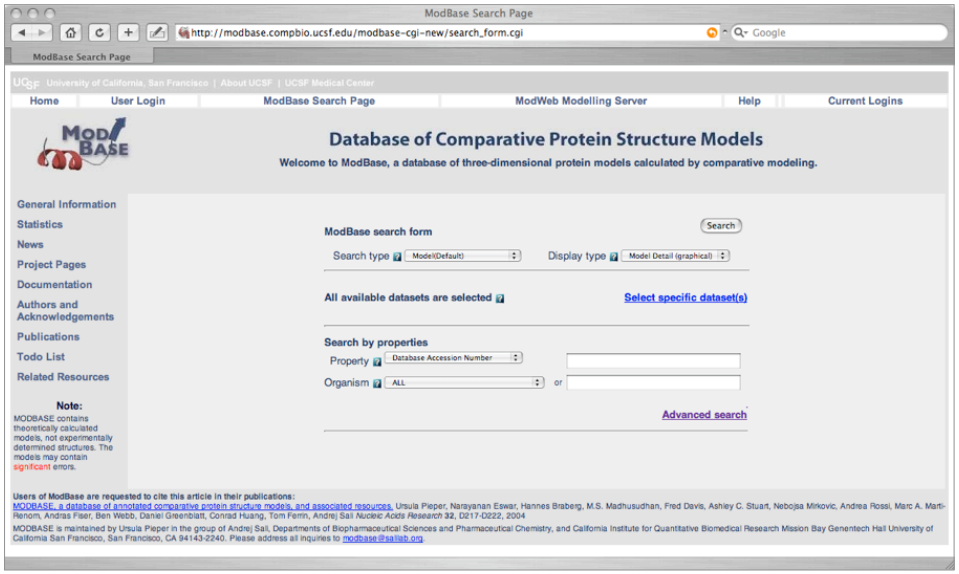
Eswar et.al., (2003) Nucl.Acids.Res. 31(13)

ModBase Statistics

Large-scale modeling of the TrEMBL-SWISSPROT databases

http://www.salilab.org/modbase/

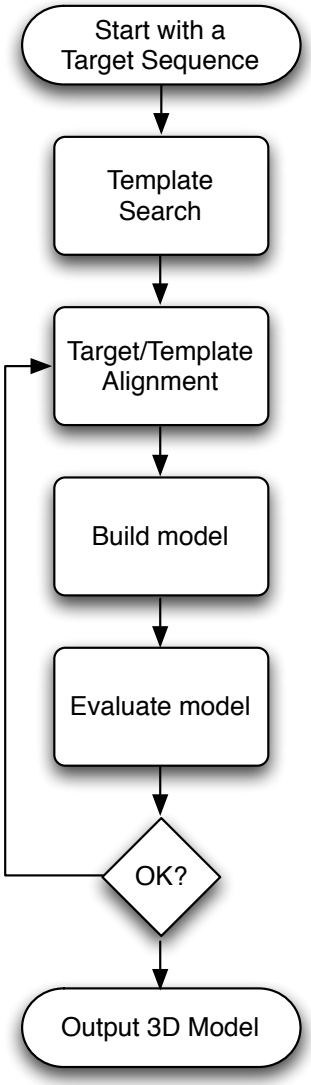
Sequences (total)	2,186,210
Sequences (modeled)	1,340,687
Models	4,580,270



UCSF
University of California
San Francisco

Pieper et al. NAR 34, D291 (2006)

Structure-Structure alignments

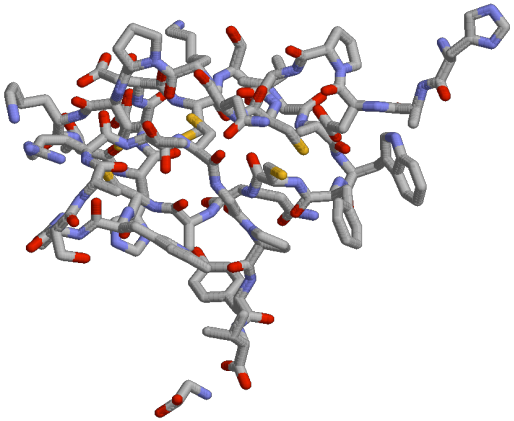


As any other bioinformatics problem…

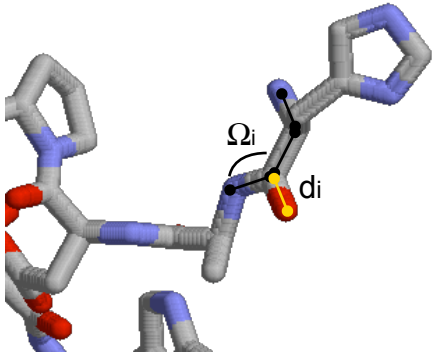
- **Representation**
- **Scoring**
- **Optimizer**

Representation

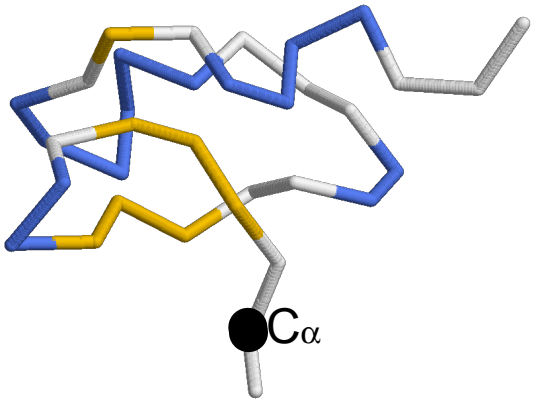
Structures



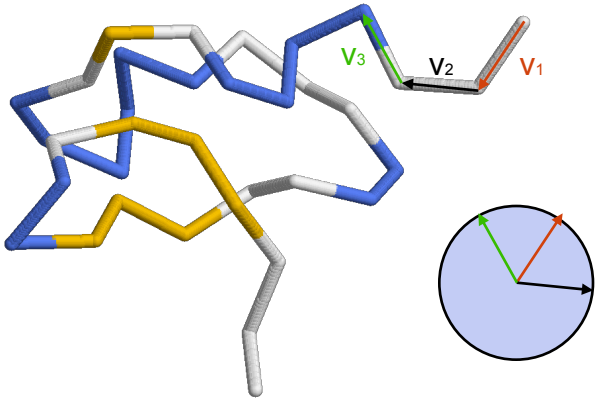
All atoms and coordinates



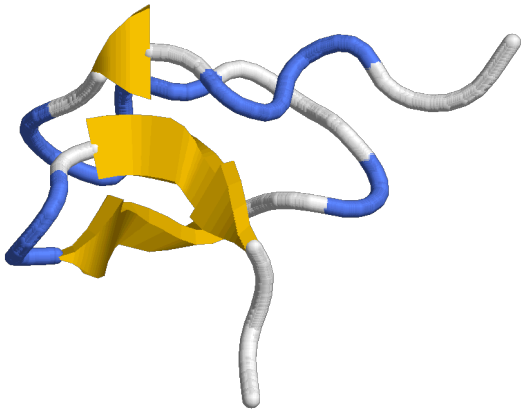
Dihedral space or distance space



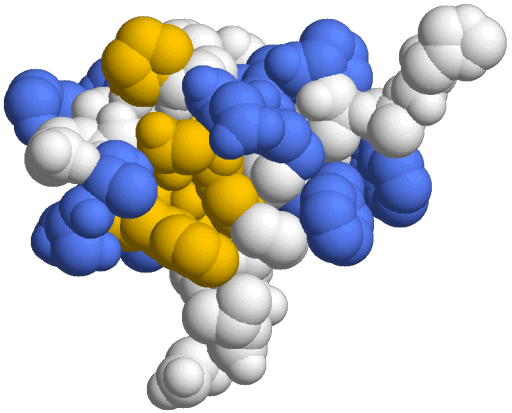
Reduced atom representation



Vector representation



Secondary Structure



Accessible surface (and others)

Scoring

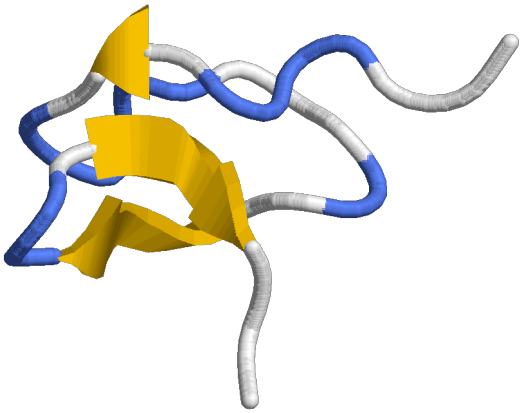
Raw scores

	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**

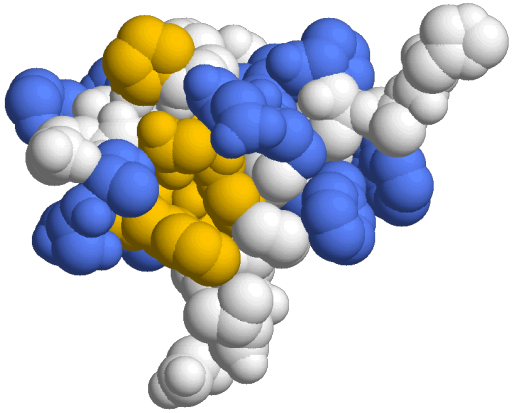
Aminoacid substitutions

$$\text{RMSD} = \sqrt{\sum \left(x_i - \bar{x} \right)^2}$$

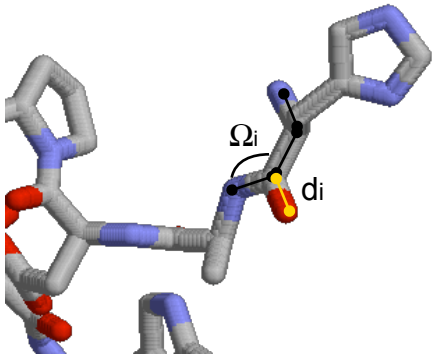
Root Mean Square Deviation



Secondary Structure (H,B,C)



Accessible surface (B,A [%])

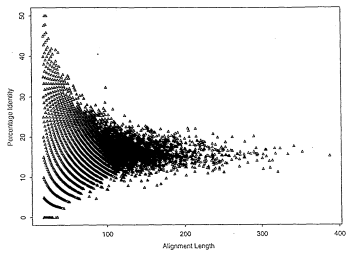


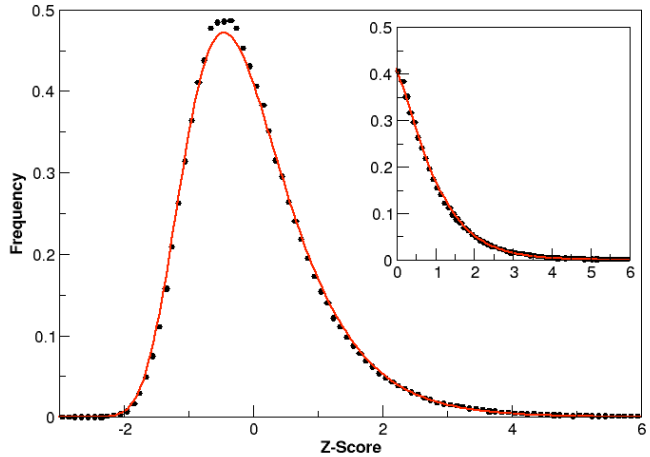
Angles or distances

Significance of an alignment (score)

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

Empirical





Sometimes approximated by Z-score (normal distribution).

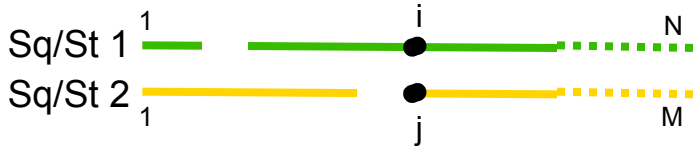
Analytic

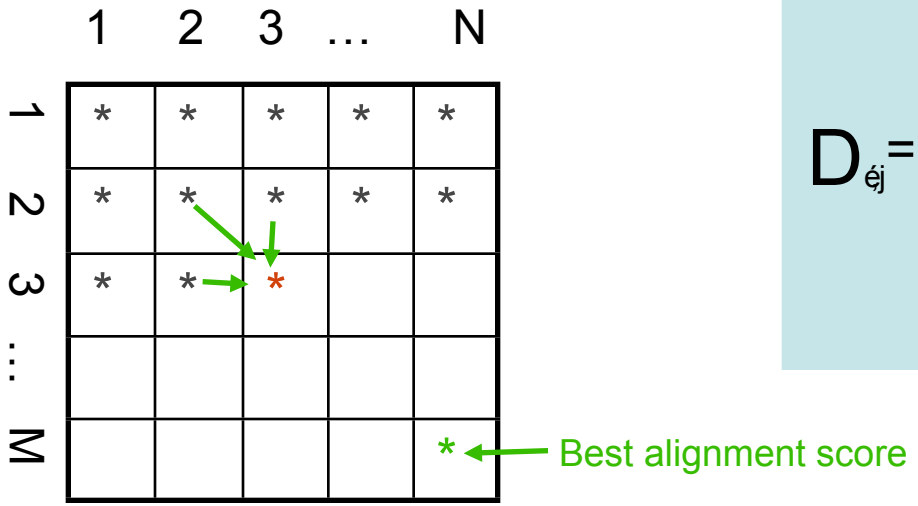
$$P(s) = e^{-\lambda\,(s-\mu)}$$

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

Karlin and Altschul, 1990 *PNAS* **87**, pp2264

Global dynamic programming alignment



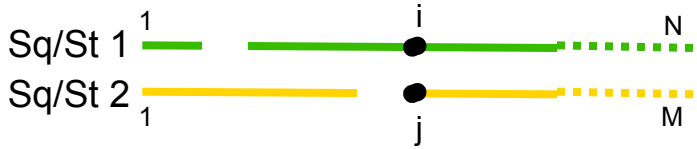


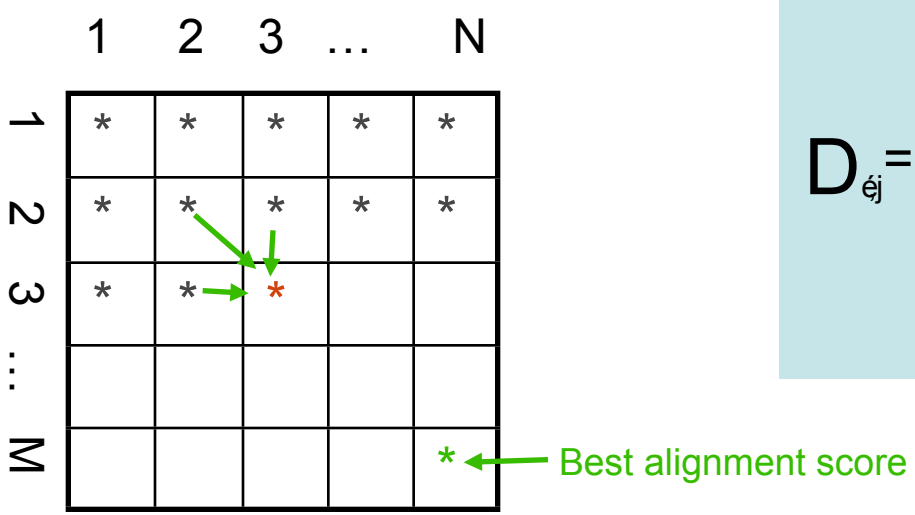
$$D_{éj} = \min \begin{cases} D_{i,j-1} + Score_{(Ä,rj)} \\ D_{i-1,j-1} + Score_{(ri,rj)} \\ D_{i-1,j} + Score_{(ri,Ä)} \end{cases}$$

Backtracking to get the best alignment

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

Global dynamic programming alignment



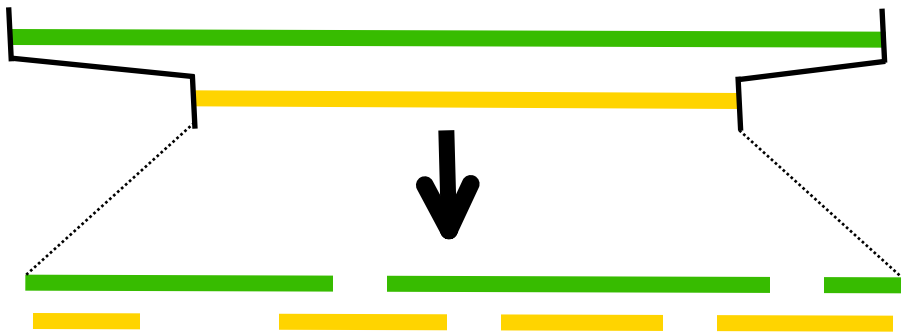


$$D_{éj} = \min \begin{cases} D_{i,j-1} + Score_{(Ä,rj)} \\ D_{i-1,j-1} + Score_{(ri,rj)} \\ D_{i-1,j} + Score_{(ri,Ä)} \\ 0 \end{cases}$$

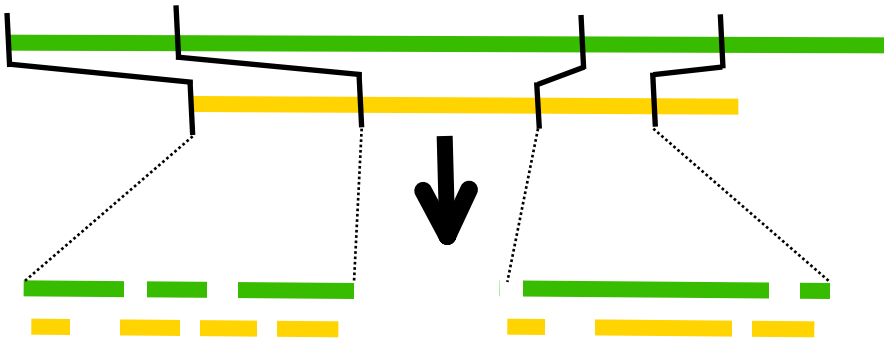
Backtracking to get the best alignment

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

Global *.vs.* local alignment



Global alignment



Local alignment

Optimizer

Multiple alignment

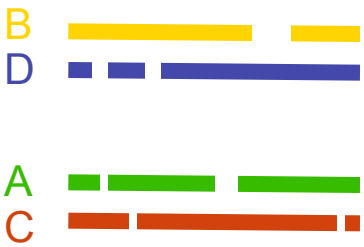
Pairwise alignments

Example – 4 sequences A, B, C, D.

6 pairwise comparisons
then cluster analysis

Multiple alignments

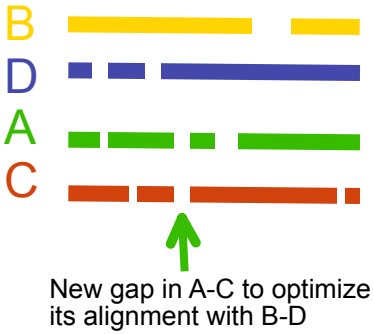
Following the tree from step 1



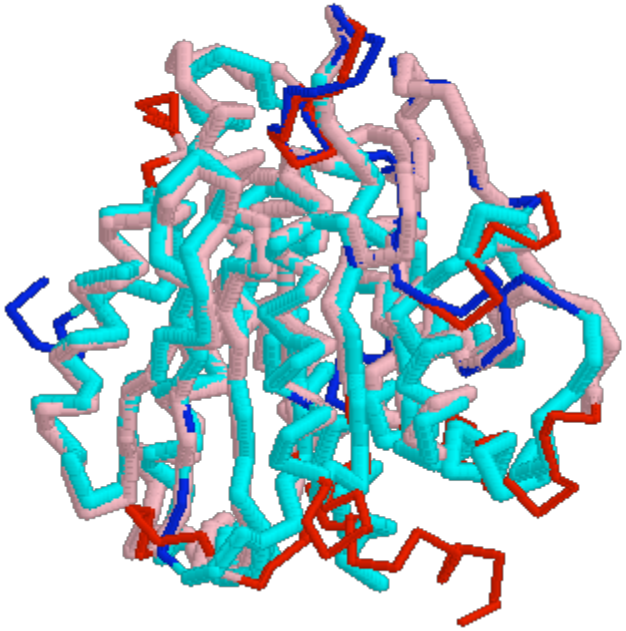
Align the most similar pair

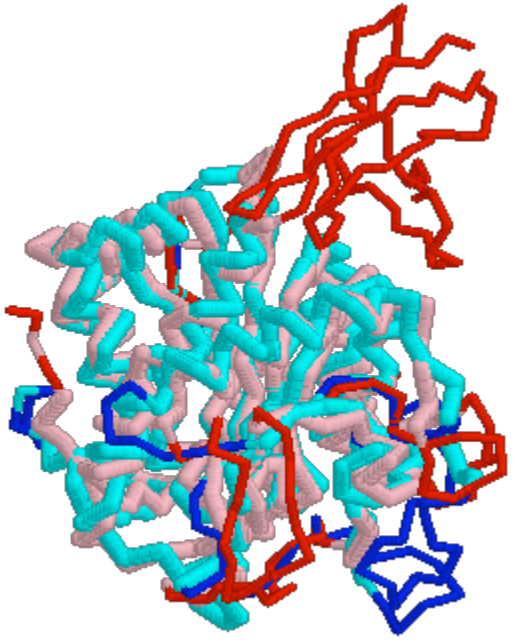
Align next most similar pair

Align B-D with A-C



Coverage .vs. Accuracy



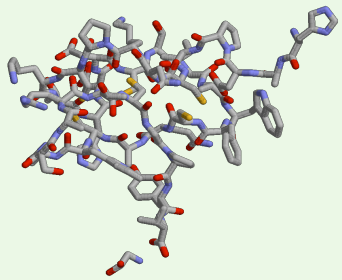


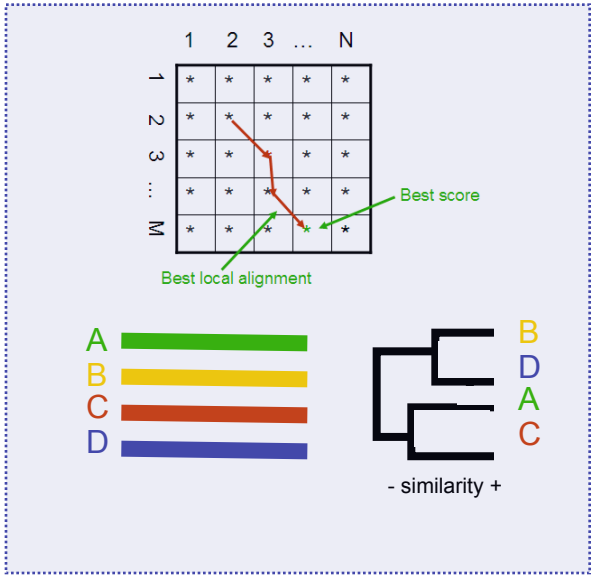
Same RMSD ~ 2.5Å

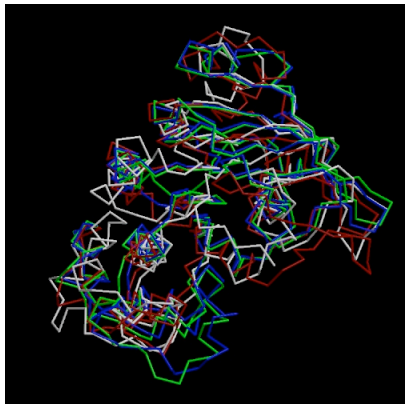
Coverage ~90% Cα

Coverage ~75% Cα

Sequence-Structure alignment by properties conservation (SALIGN-MODELLER)

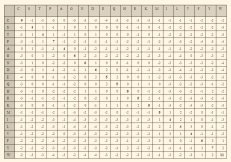




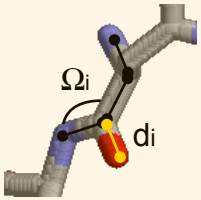


- ✓ Uses all available structural information
- ✓ Provides the optimal alignment

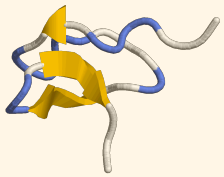
Computationally expensive



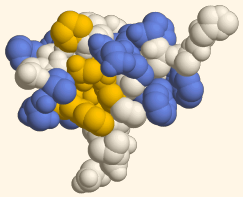
$R_{i,j}$



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



$S_{i,j}$



$B_{i,j}$

$$\text{RMSD} = \sqrt{\sum (x_i - \bar{x})^2}$$

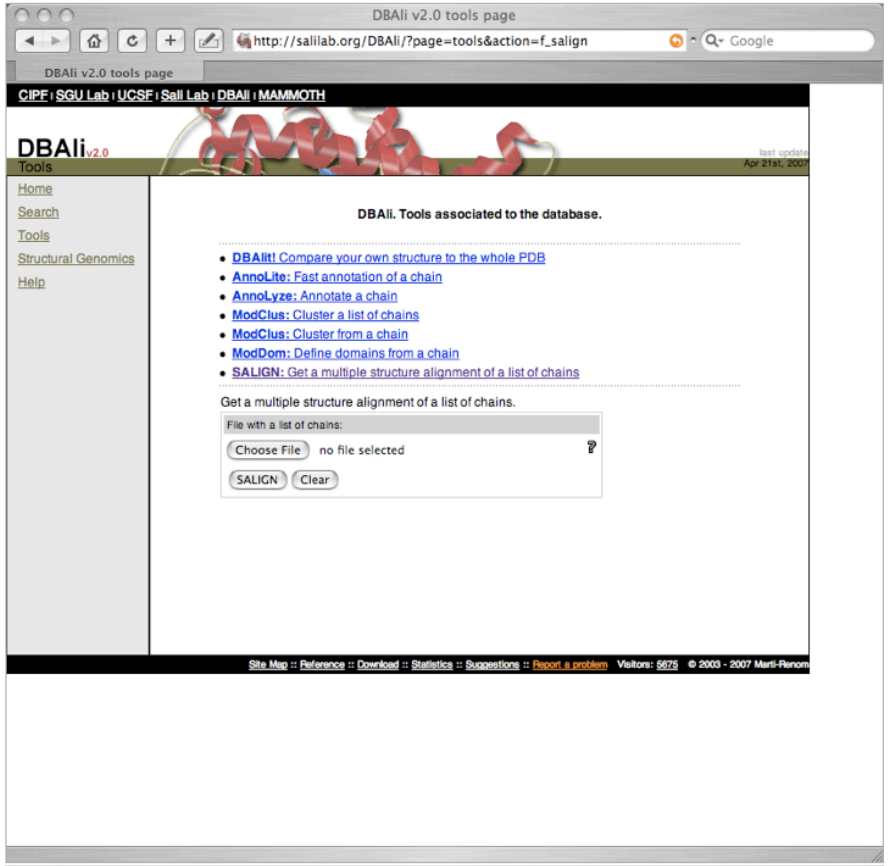
$I_{i,j}$

$$Score_{i,j} = w_1 * R_{i,j} + w_2 * D_{i(a),j(a)} + w_3 * S_{i,j} + w_4 * B_{i,j} + w_5 * I_{i,j} + w_6 * X_{i,j}$$

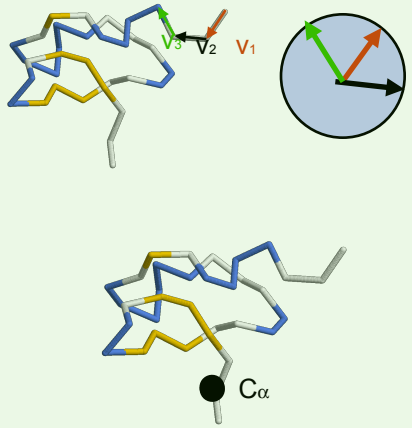
Madhusudhan et al. *in preparation*

Structural alignment by properties conservation (SALIGN-MODELLER)

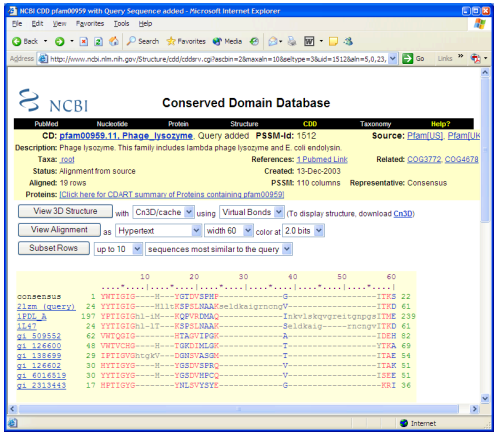
http://www.salilab.org/DBAli



Vector Alignment Search Tool (VAST)

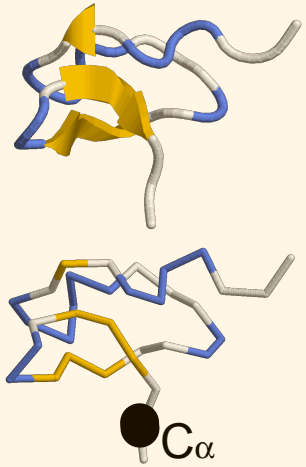


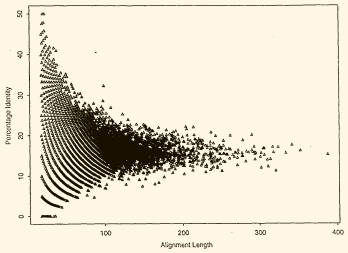
- Graph theory search of similar SSE
- Refining by Monte Carlo at all atom resolution



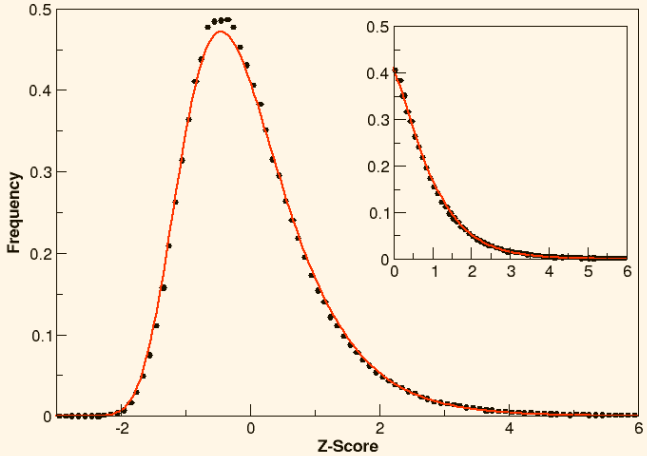
✓ **Good scoring system with significance**

Reduces the protein representation





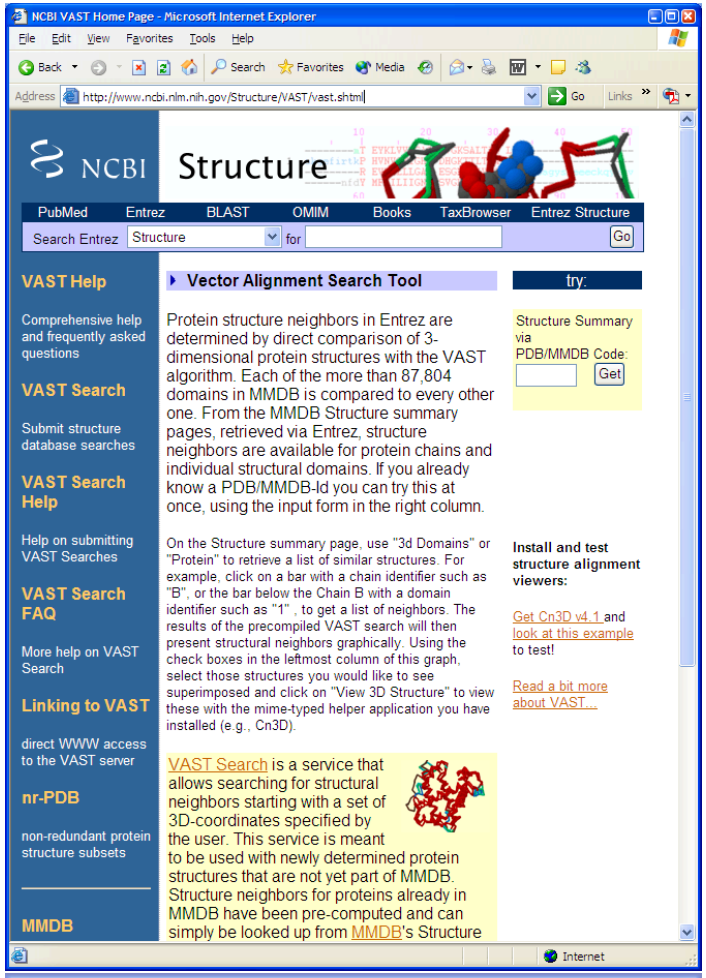
$$\text{RMSD} = \sqrt{\sum\left(x_i - \bar{x}\right)^2}$$



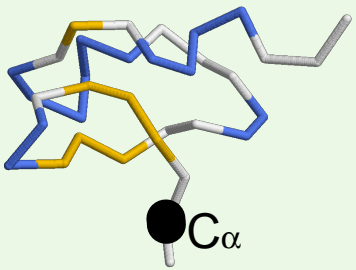
Gibrat JF et al. (1996) *Curr Opin Struct Biol* **3** pp377

Vector Alignment Search Tool (VAST)

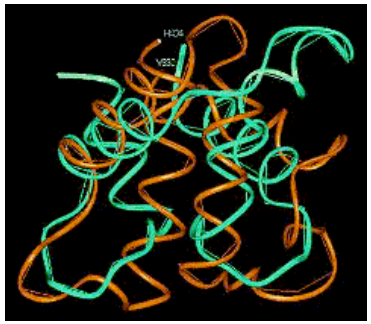
http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml



Incremental combinatorial extension (CE)

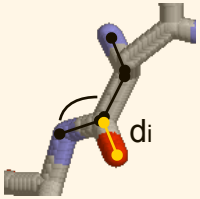


- Exhaustive combination of fragments
- Longest combination of AFPs
- Heuristic similar to PSI-BLAST



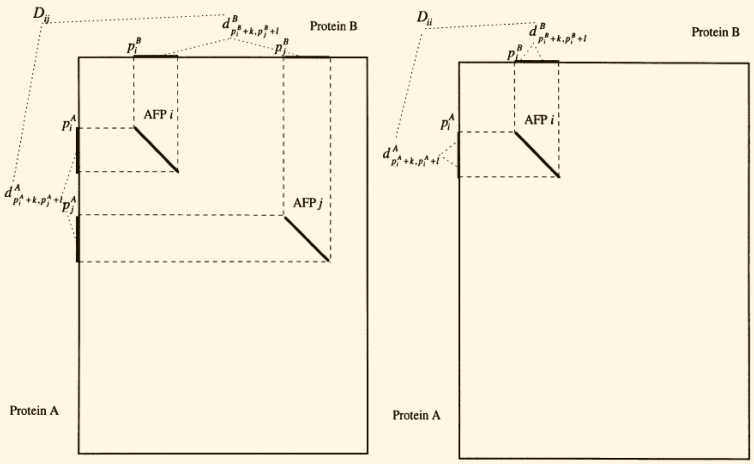
✓ **FAST!**
✓ **Good quality of local alignments**

Complicated scoring and heuristics



8 residues peptides

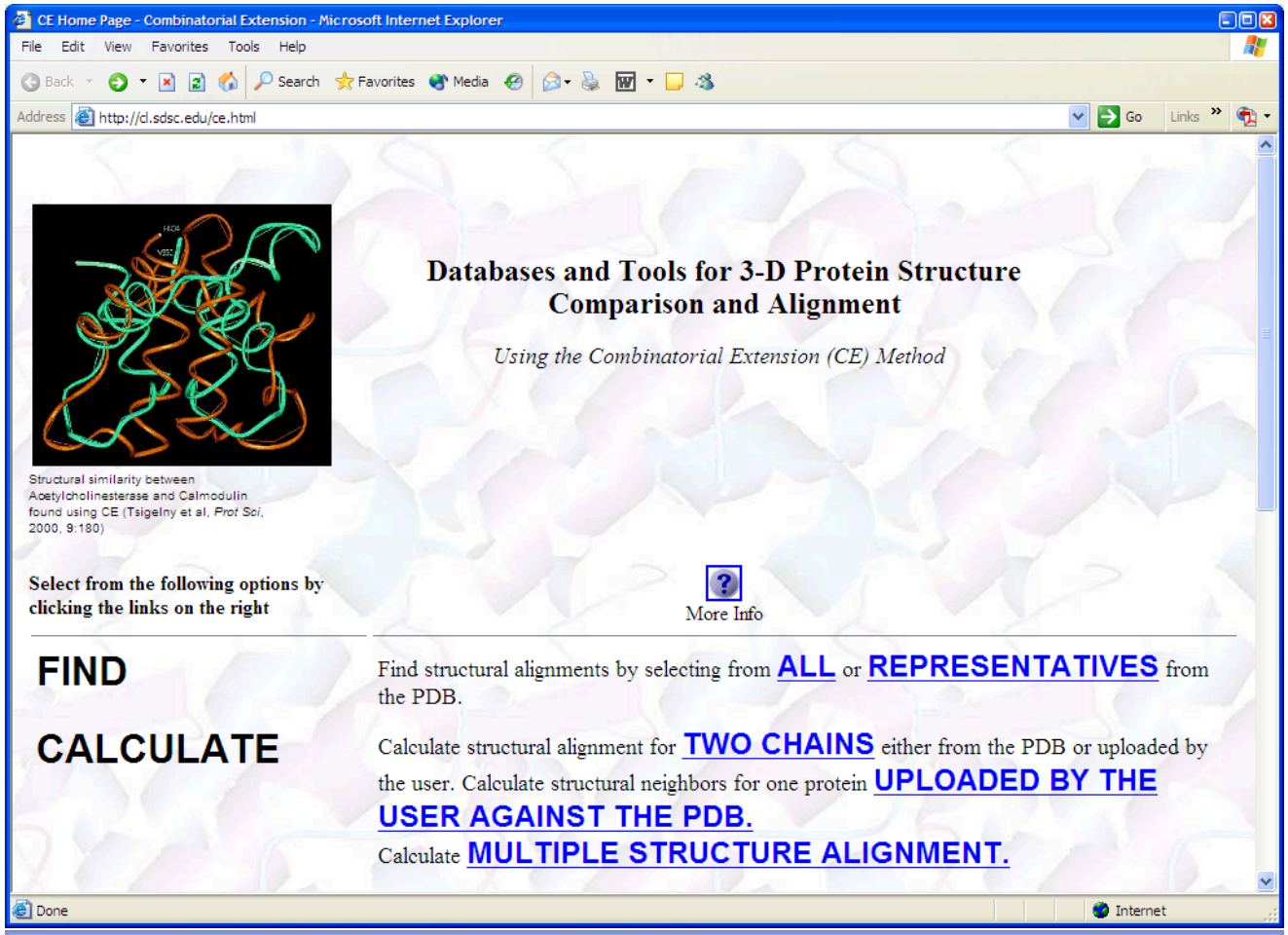
$$\text{RMSD} = \sqrt{\sum (x_i - \bar{x})^2}$$



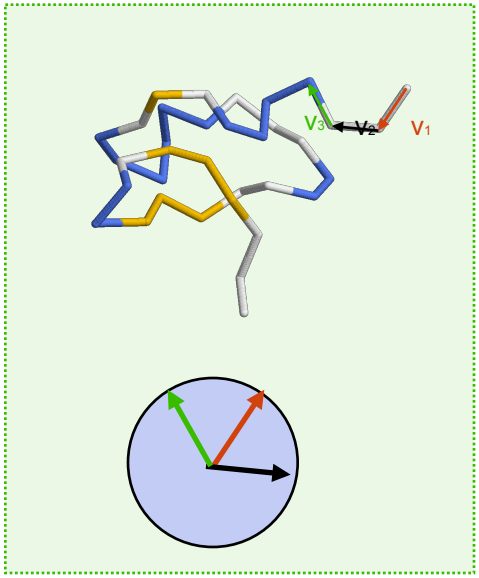
Shindyalov IN, amd Bourne PE. (1998) *Protein Eng.* **9** pp739

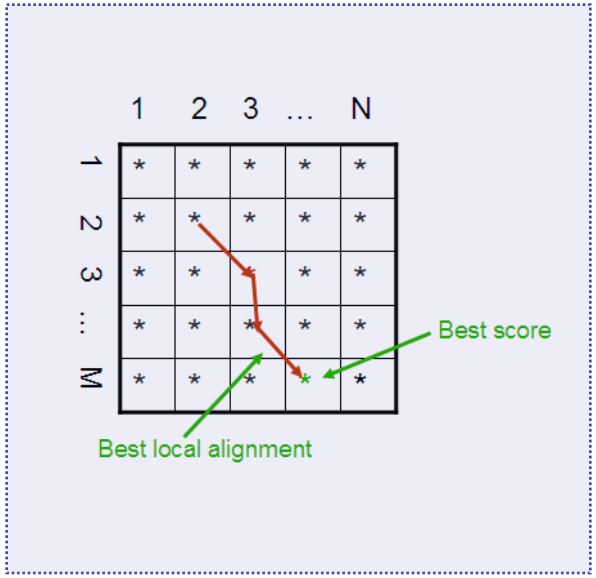
Incremental combinatorial extension (CE)

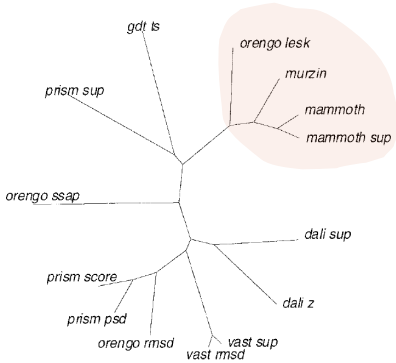
http://cl.sdsc.edu/ce.html



Matching molecular models obtained from theory (MAMMOTH)





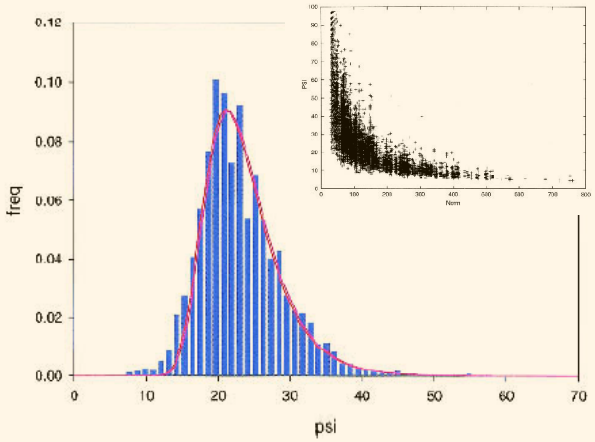


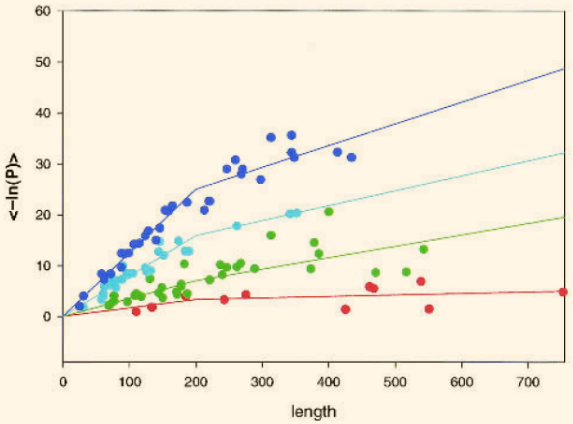
- ✓ **VERY FAST!**
- ✓ **Good scoring system with significance**

Reduces the protein representation

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

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

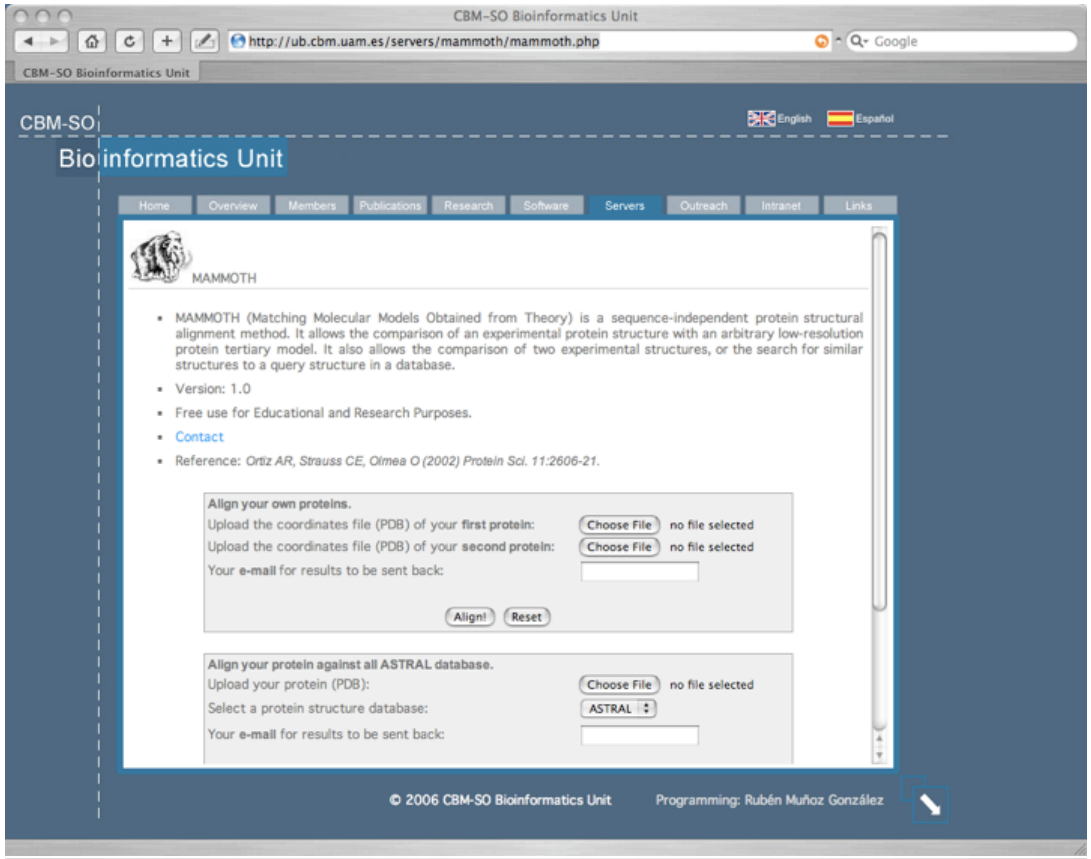




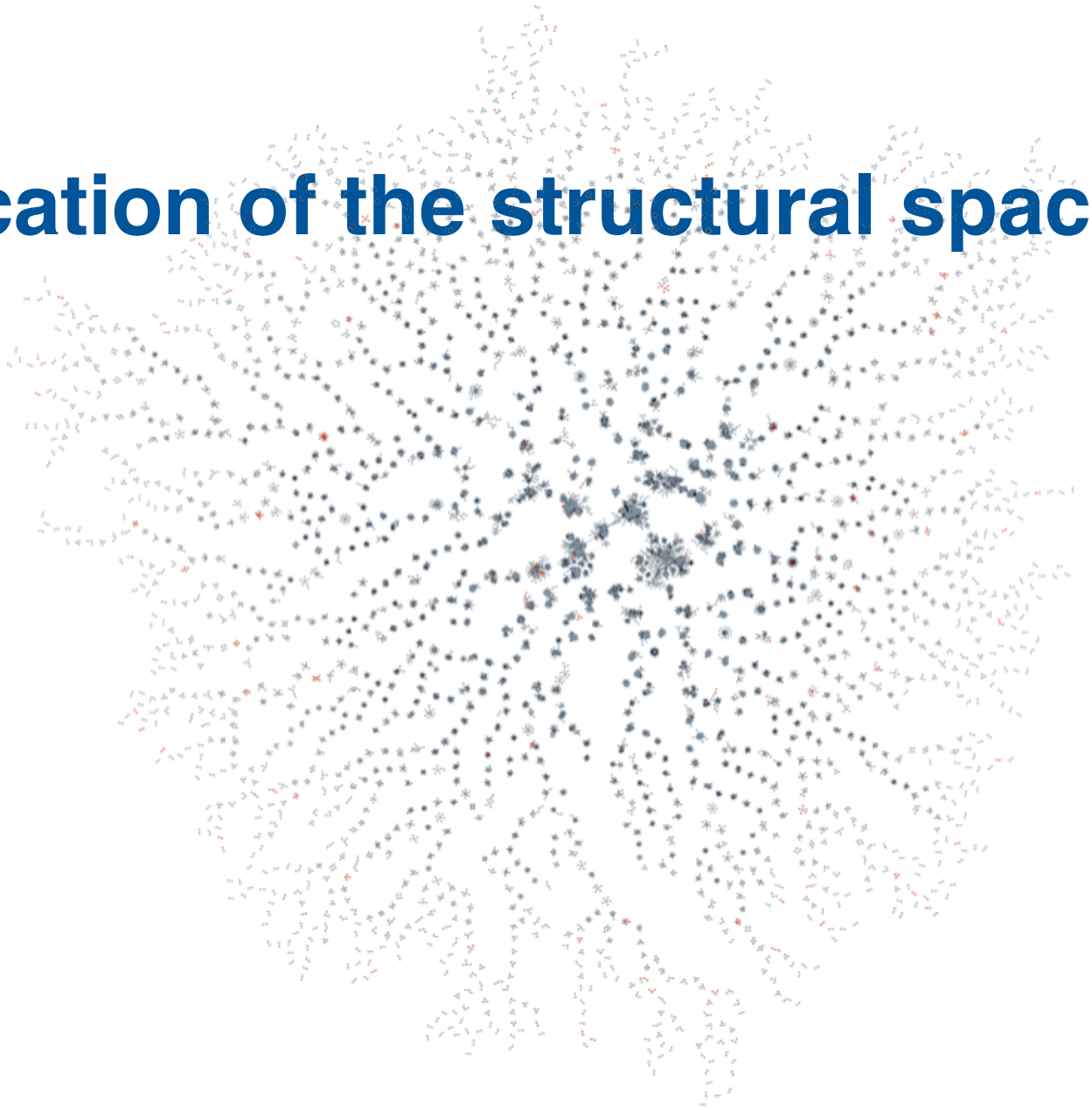
Ortiz AR, (2002) *Protein Sci.* **11** pp2606

Matching molecular models obtained from theory (MAMMOTH)

http://ub.cbm.uam.es/servers/mammoth/

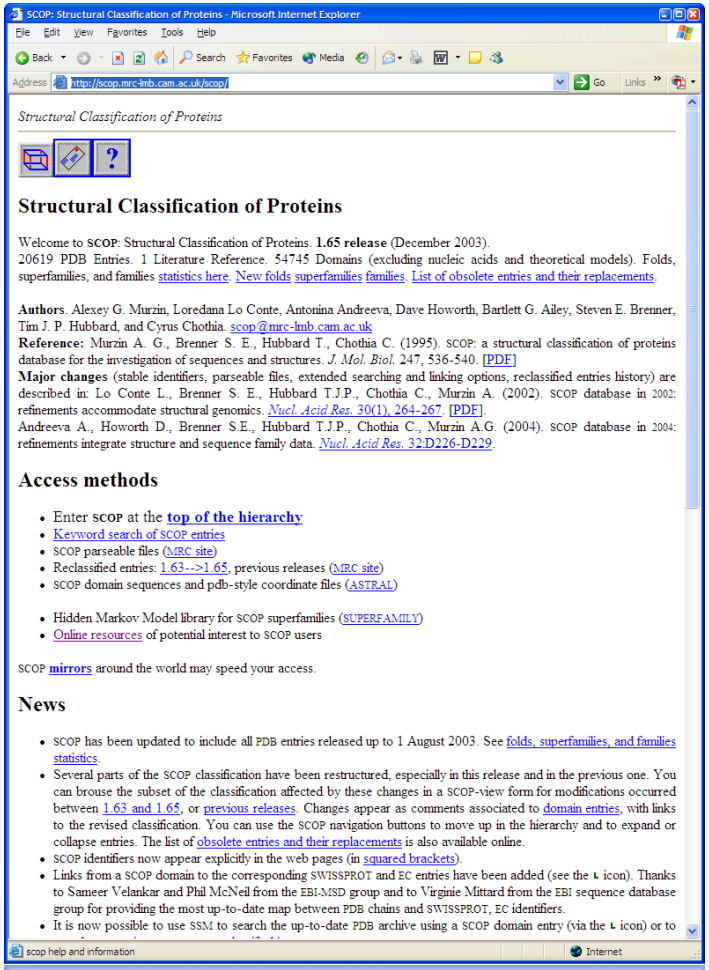


Classification of the structural space



SCOP$_{1.71}$ database

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



SCOP: Structural Classification of Proteins - Microsoft Internet Explorer

File Edit View Favorites Tools Help

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

Structural Classification of Proteins

Structural Classification of Proteins

Welcome to SCOP: Structural Classification of Proteins. **1.65 release** (December 2003).
20619 PDB Entries. 1 Literature Reference. 54745 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, Loredana Lo Conte, Antonina Andreeva, Dave Howorth, 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]
Major changes (stable identifiers, parseable files, extended searching and linking options, reclassified entries history) 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.

Access methods

- Enter SCOP at the top of the hierarchy
- Keyword search of SCOP entries
- SCOP parseable files (MRC site)
- Reclassified entries: 1.63-->1.65, previous releases (MRC site)
- SCOP domain sequences and pdb-style coordinate files (ASTRAL)

- Hidden Markov Model library for SCOP superfamilies (SUPERFAMILY)
- Online resources of potential interest to SCOP users

SCOP mirrors around the world may speed your access.

News

- SCOP has been updated to include all PDB entries released up to 1 August 2003. See folds, superfamilies, and families statistics.
- Several parts of the SCOP classification have been restructured, especially in this release and in the previous one. You can browse the subset of the classification affected by these changes in a SCOP-view form for modifications occurred between 1.63 and 1.65, or previous releases. Changes appear as comments associated to domain entries, with links to the revised classification. You can use the SCOP navigation buttons to move up in the hierarchy and to expand or collapse entries. The list of obsolete entries and their replacements is also available online.
- SCOP identifiers now appear explicitly in the web pages (in squared brackets).
- Links from a SCOP domain to the corresponding SWISSPROT and EC entries have been added (see the L icon). Thanks to Sameer Velankar and Phil McNeil from the EBI-MSD group and to Virginie Mittard from the EBI sequence database group for providing the most up-to-date map between PDB chains and SWISSPROT, EC identifiers.
- It is now possible to use SSM to search the up-to-date PDB archive using a SCOP domain entry (via the L icon) or to

scop help and information | Internet

- ✓ 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

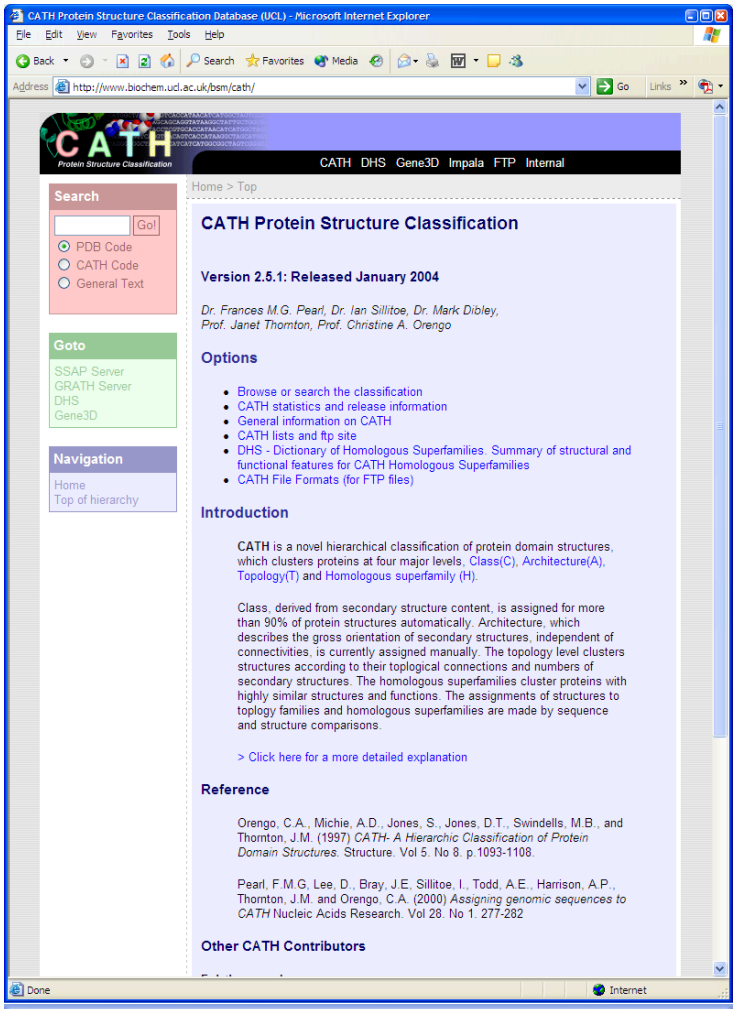
Manually classification
Not 100% up-to-date
Domain boundaries definition

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	226	392	645
All beta proteins	149	300	594
Alpha and beta proteins (a/b)	134	221	661
Alpha and beta proteins (a+b)	286	424	753
Multi-domain proteins	48	48	64
Membrane and cell surface proteins	49	90	101
Small proteins	79	114	186
Total	971	1589	3004

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

CATH3.1.0 database

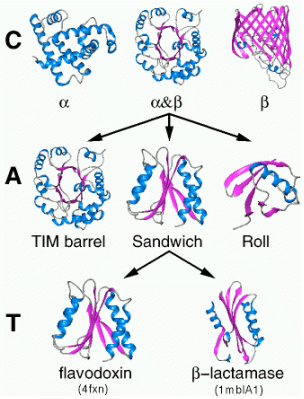
http://www.cathdb.info



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

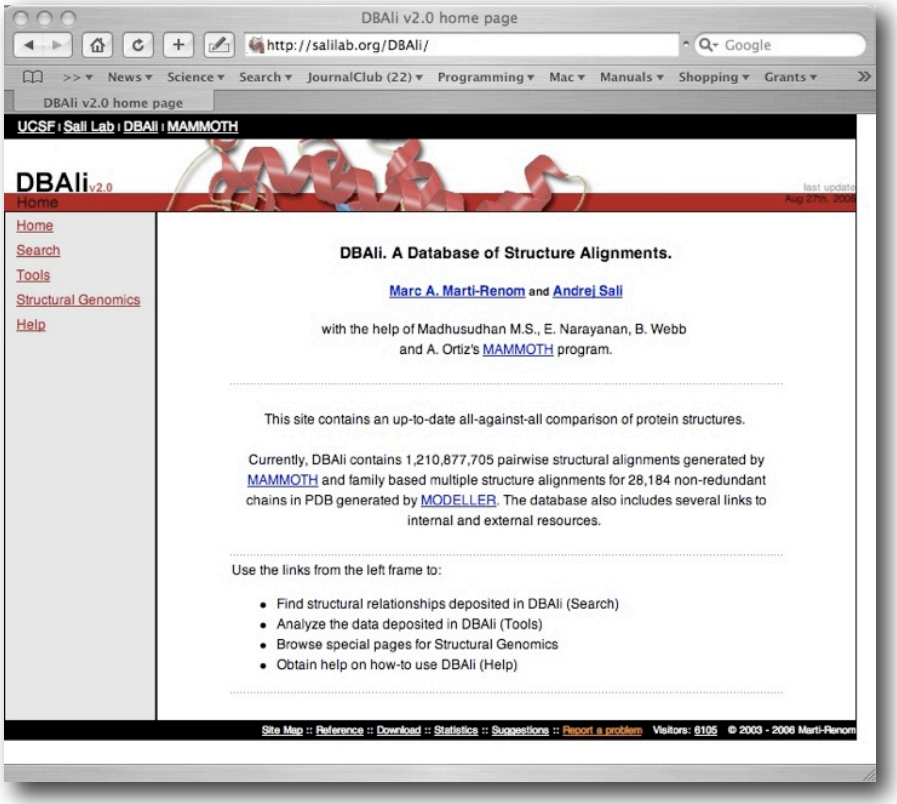


C	A	T	H	S	O	L	I	D
Mainly Alpha	5	305	652	1850	2329	3001	5587	19729
Mainly Beta	20	191	415	1860	2531	3846	6503	25537
Alpha Beta	14	496	922	3922	5303	6659	12998	47193
Few Secondary Structures	1	92	102	162	200	275	403	1426
Total	**40**	**1084**	**2091**	**7794**	**10363**	**13781**	**25491**	**93885**

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

DBAli$_{v2.0}$ database

http://bioinfo.cipf.es/squ/services/DBAli/
http://www.salilab.org/DBAli/



- ✓ Fully-automatic
- ✓ Data is kept up-to-date with PDB releases
- ✓ Tools for "on the fly" classification of families.
- ✓ Easy to navigate
- ✓ Provides tools for structure analysis

Does not provide a stable classification similar to that of CATH or SCOP

Pairwise structure alignments	
Last update:	February 15th, 2007
Number of chains:	88,276
Number of structure-structure comparisons:*	1,425,479,365
Multiple structure alignments	
Last update:	January 23rd, 2007
Number of representative chains:	30,900
Number of families:	11,615

Uses MAMMOTH for similarity detection

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

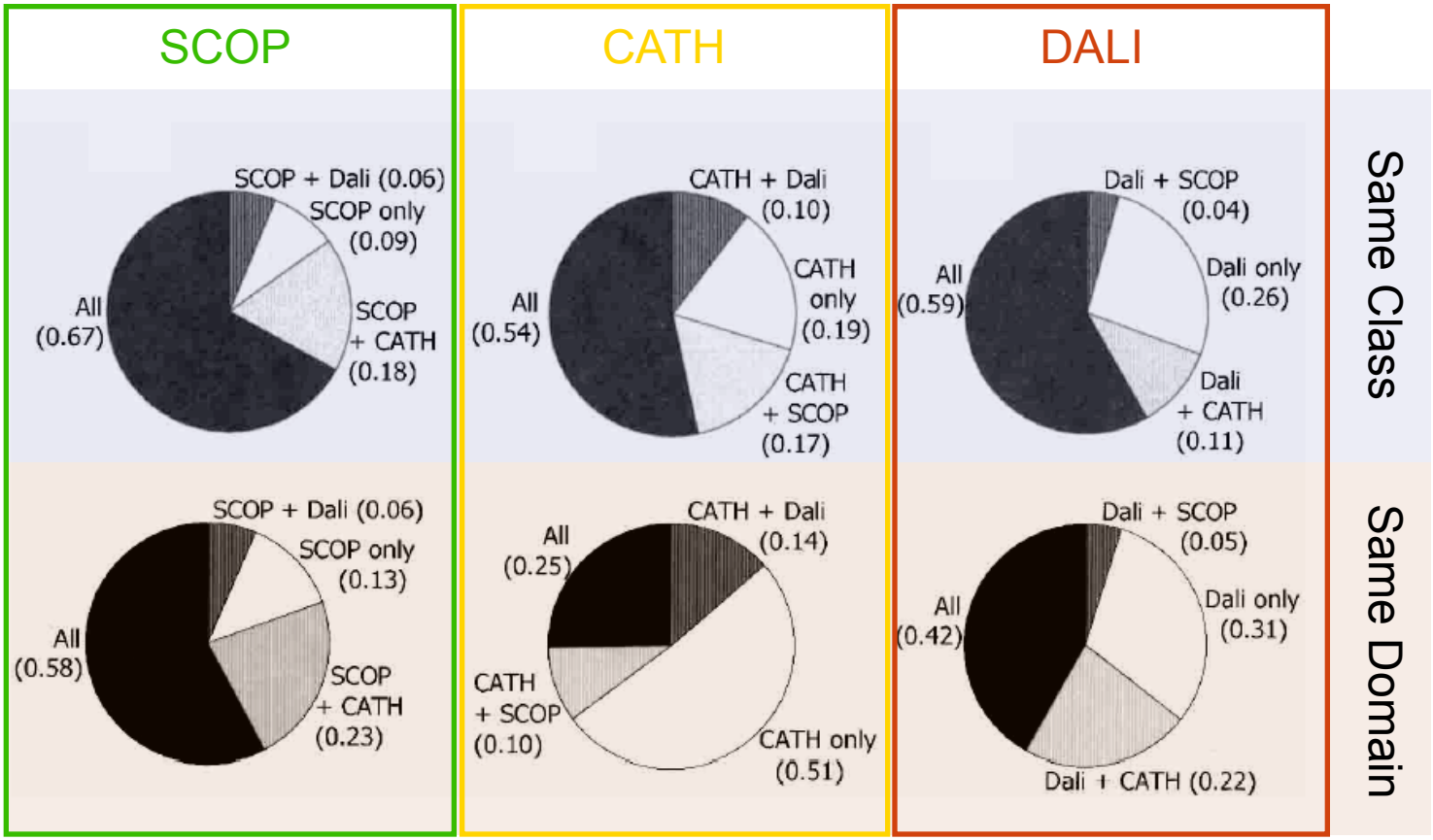
Ortiz AR, (2002) Protein Sci. 11 pp2606

Marti-Renom et al. 2001. Bioinformatics. 17, 746

Classification of the structural space

Not an easy task!

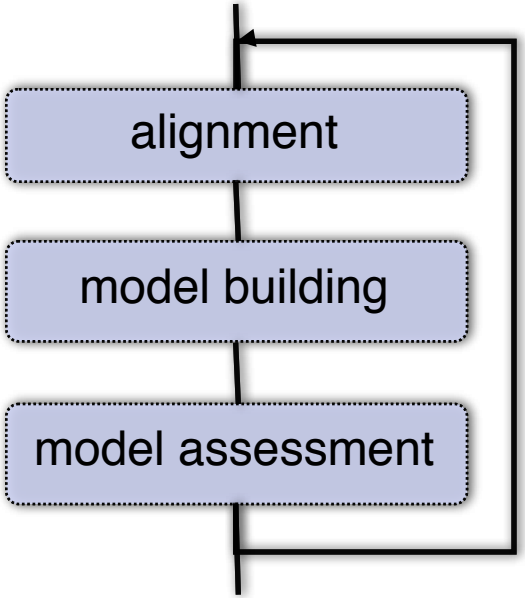
Domain definition AND domain classification

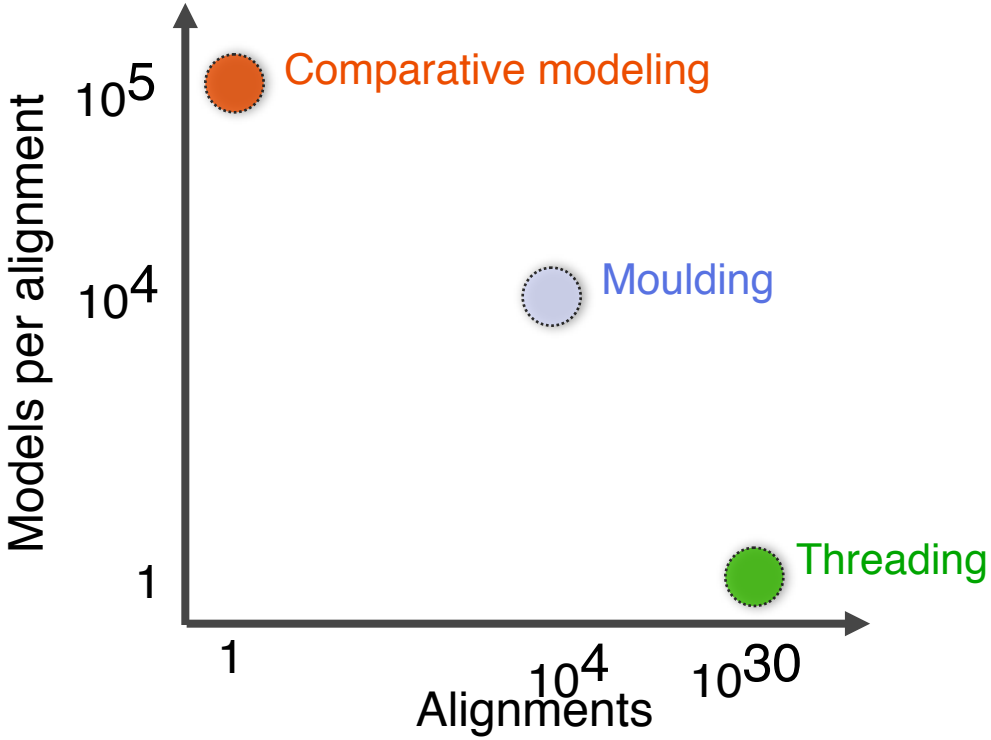


*Day, et al. (2003) Protein Sciences, **12** pp2150*

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

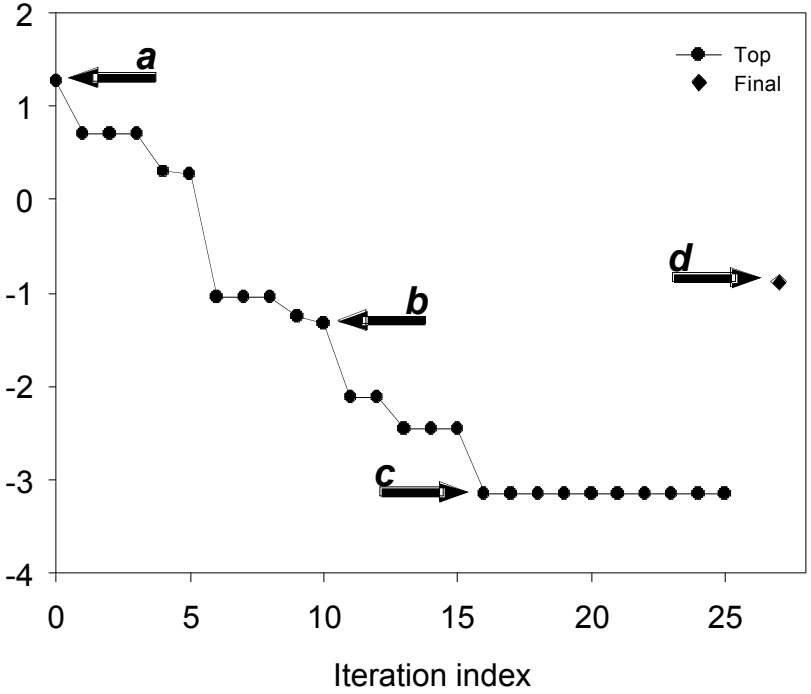
Benchmark with the "very difficult" test set

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

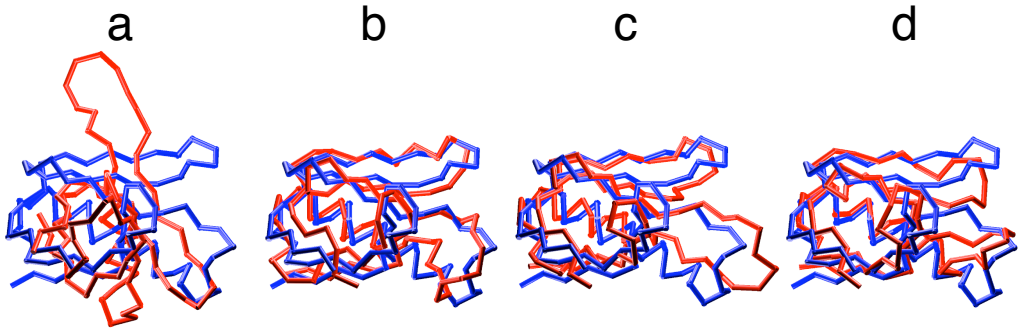
1BOV-1LTS

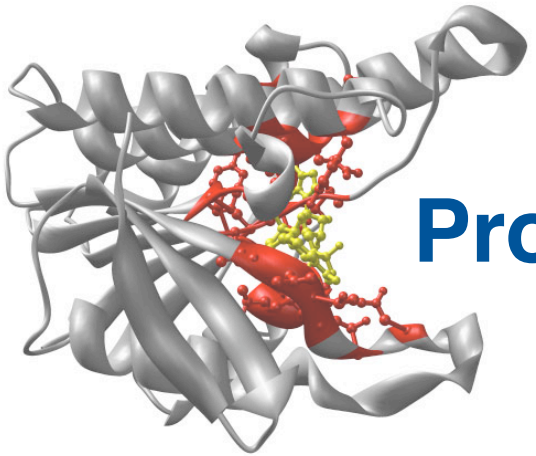


Sequence identity 4.4%

Initial model Cα RMSD 10.1Å

Final model Cα RMSD 3.6Å





Protein function from structure

ab-initio *localization of binding sites*

Rossi, et al. Protein Science, 15, 2366 (2006)

For many protein structures function is *unknown*

	Structural Genomics*	Traditional methods
Annotated**	654	28,342
Not Annotated	506 (43.6%)	6,815 (19,4%)
Total deposited	1,160	35,157

** annotated as STRUCTURAL GENOMICS in the header of the PDB file*
***annotated with either CATH, SCOP, Pfam or GO terms in the MSD database*
36,317 protein structures, as of August 8th, 2006

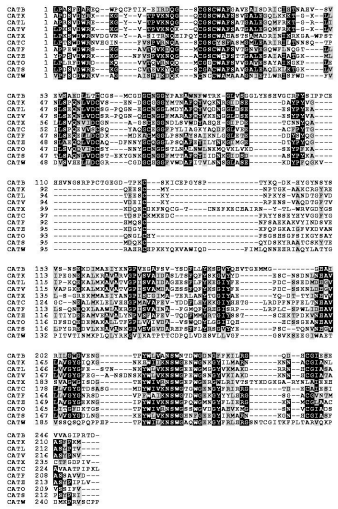
For 20% protein structures function is *unknown*

	Structural Genomics*	Traditional methods
Annotated**	654	28,342
Not Annotated	506 (43.6%)	6,815 (19,4%)
Total deposited	1,160	35,157

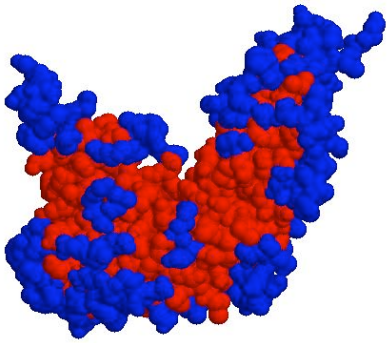
** annotated as STRUCTURAL GENOMICS in the header of the PDB file*
***annotated with either CATH, SCOP, Pfam or GO terms in the MSD database*
36,317 protein structures, as of August 8th, 2006

Representation

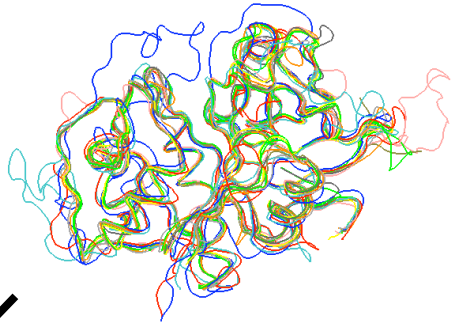
Sequence conservation



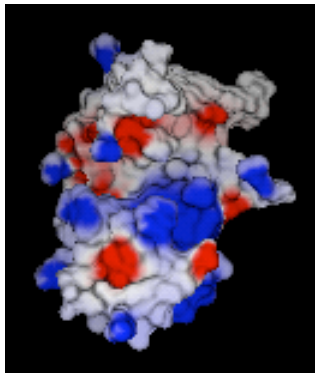
Surface geometry

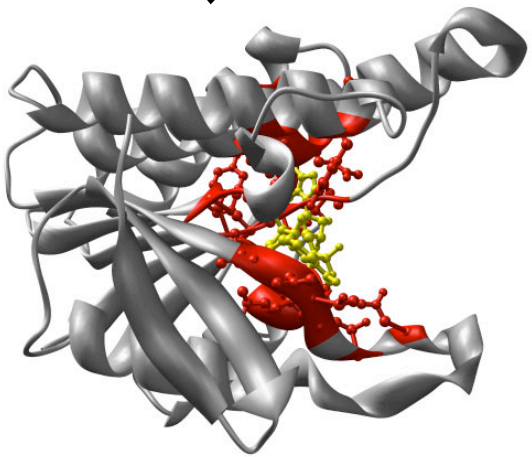


Structure conservation

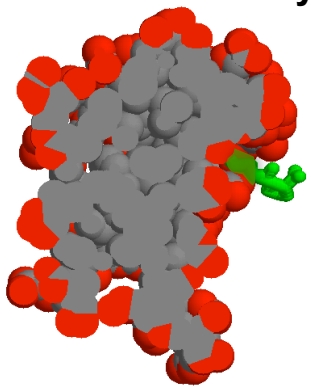


Electrostatics



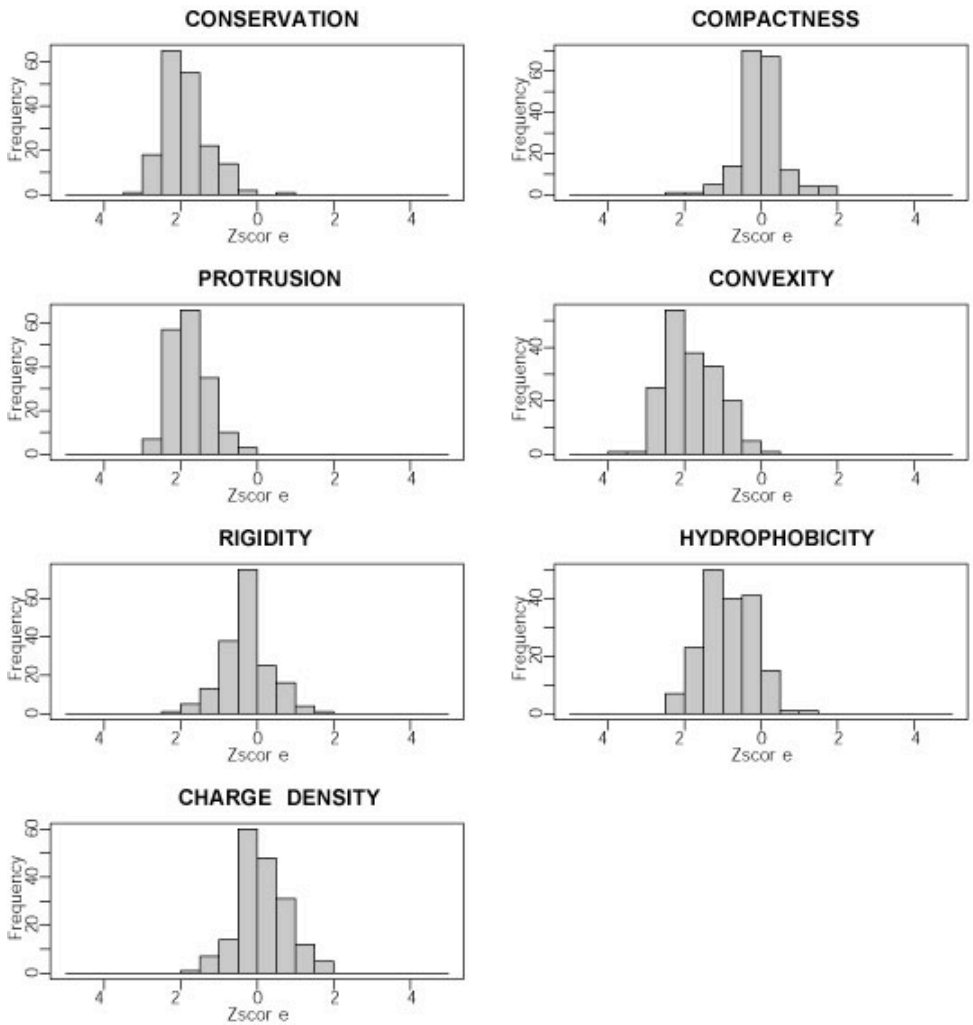


Solvent accessibility



Scoring

NAD



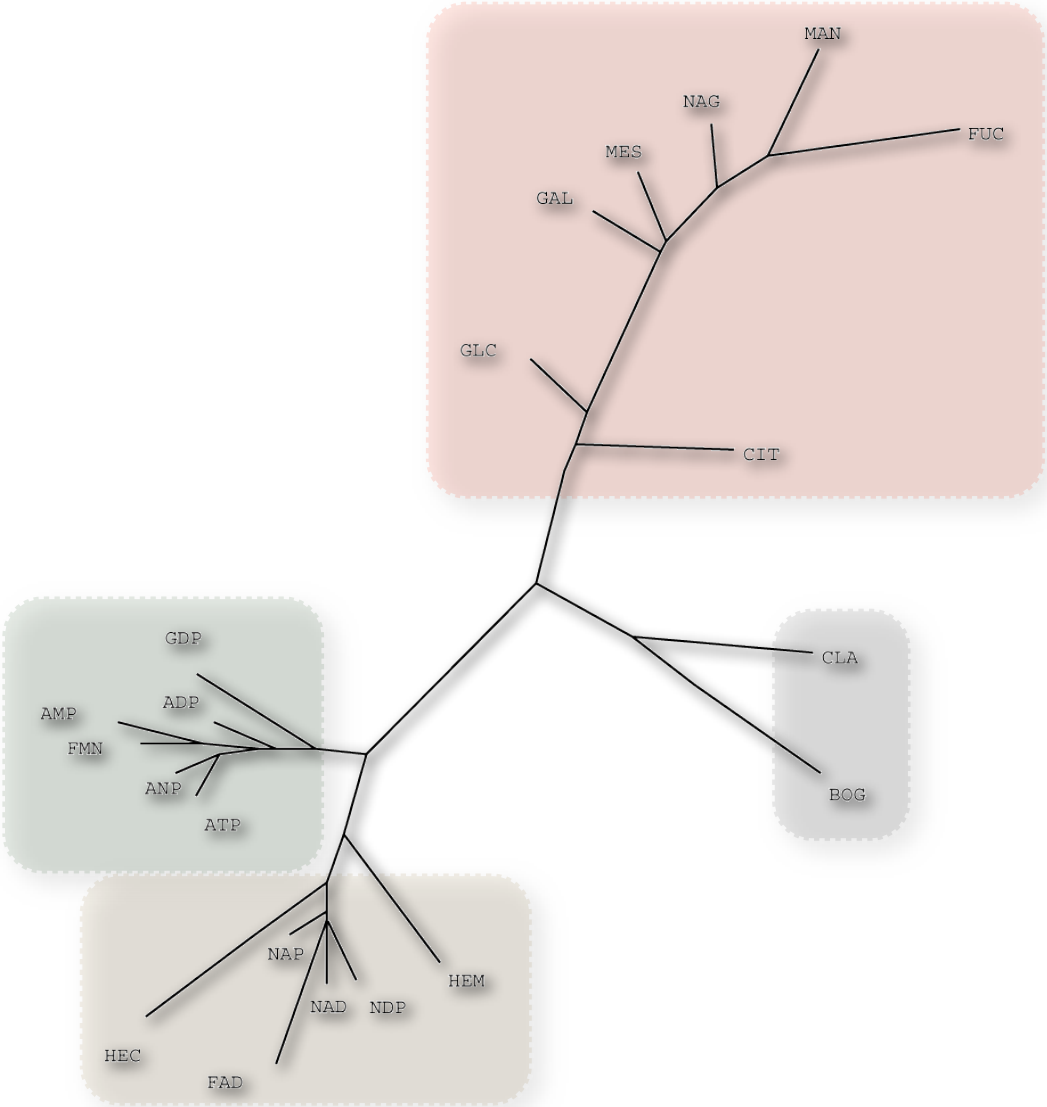
$$w_k = \frac{1}{M} \sum_{\alpha=1}^{M} \tilde{f}_k^{(\alpha)}$$

M = number of proteins in training set

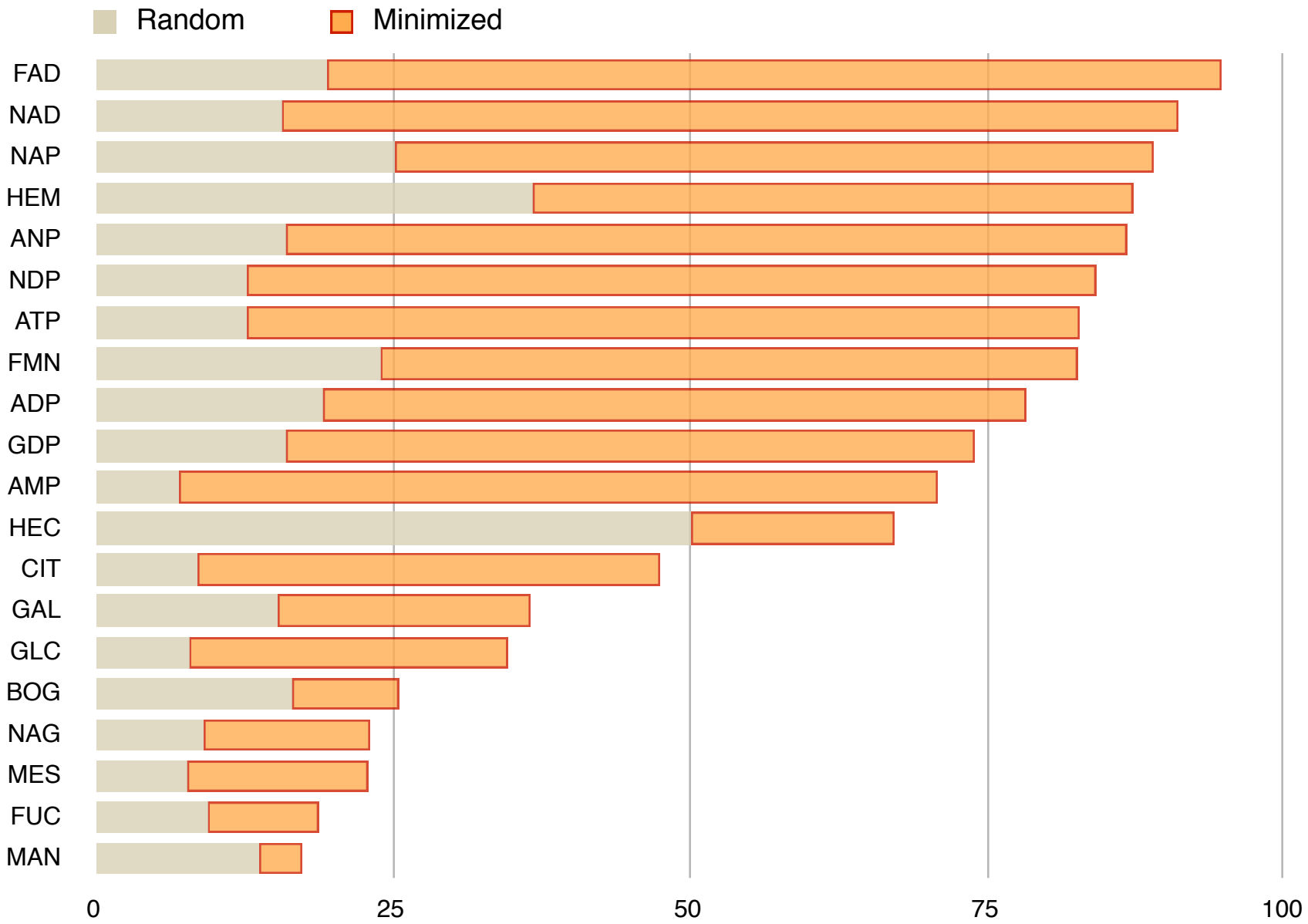
Ligand fingerprints

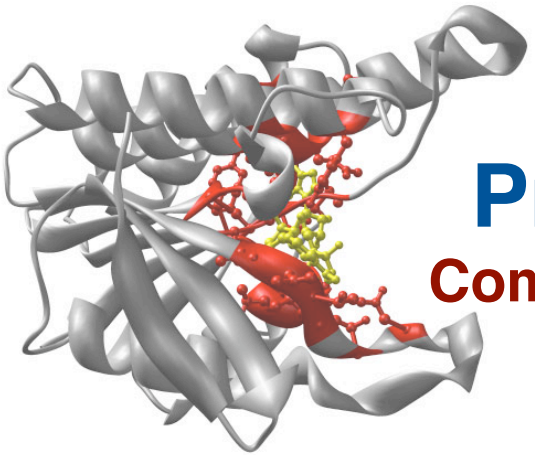
	Compactness	Conservation	Charge density	B-factor	Protrusion coefficient	Convexity score	Hydrophobicity
ADP	-1.266	-2.009	0.447	-0.414	-1.521	-1.388	-0.118
AMP	-1.62	-1.962	0.341	-0.381	-1.909	-1.944	-0.518
ANP	-1.007	-2.227	0.176	-0.392	-1.706	-1.595	-0.14
ATP	-1.122	-2.156	0.228	-0.274	-1.845	-1.768	0.038
BOG	-2.067	-0.012	0.552	-0.465	-0.356	-0.49	-0.781
CIT	-2.948	-1.58	0.563	-0.527	-0.922	-0.838	-0.113
FAD	0.505	-2.108	0.366	-0.702	-1.735	-1.725	-0.75
FMN	-1.132	-1.98	0.382	-0.387	-1.803	-1.886	-0.695
FUC	-3.43	0.016	-0.295	-0.123	0.002	0.132	0.459
GAL	-3.186	-0.538	-0.234	-0.068	-0.906	-0.987	0.298
GDP	-1.061	-1.471	0.409	-0.81	-1.472	-1.423	0.182
GLC	-2.813	-1.247	-0.207	-0.399	-1.247	-1.337	-0.089
HEC	-0.172	-0.912	0.286	-0.325	-1.153	-1.27	-1.282
HEM	-0.651	-1.571	0.683	-0.51	-1.797	-1.937	-1.47
MAN	-3.72	0.131	0.105	-0.52	-0.605	-0.509	0.405
MES	-3.049	-0.24	-0.338	-0.479	-0.714	-0.926	0.296
NAD	-0.005	-1.852	0.156	-0.232	-1.775	-1.804	-0.858
NAG	-3.419	-0.46	-0.126	-0.154	-0.341	-0.523	-0.078
NAP	-0.009	-1.898	0.612	-0.321	-1.587	-1.656	-0.336
NDP	0.217	-1.741	0.535	-0.312	-1.463	-1.562	-0.498

Ligand fingerprints



Prediction accuracy





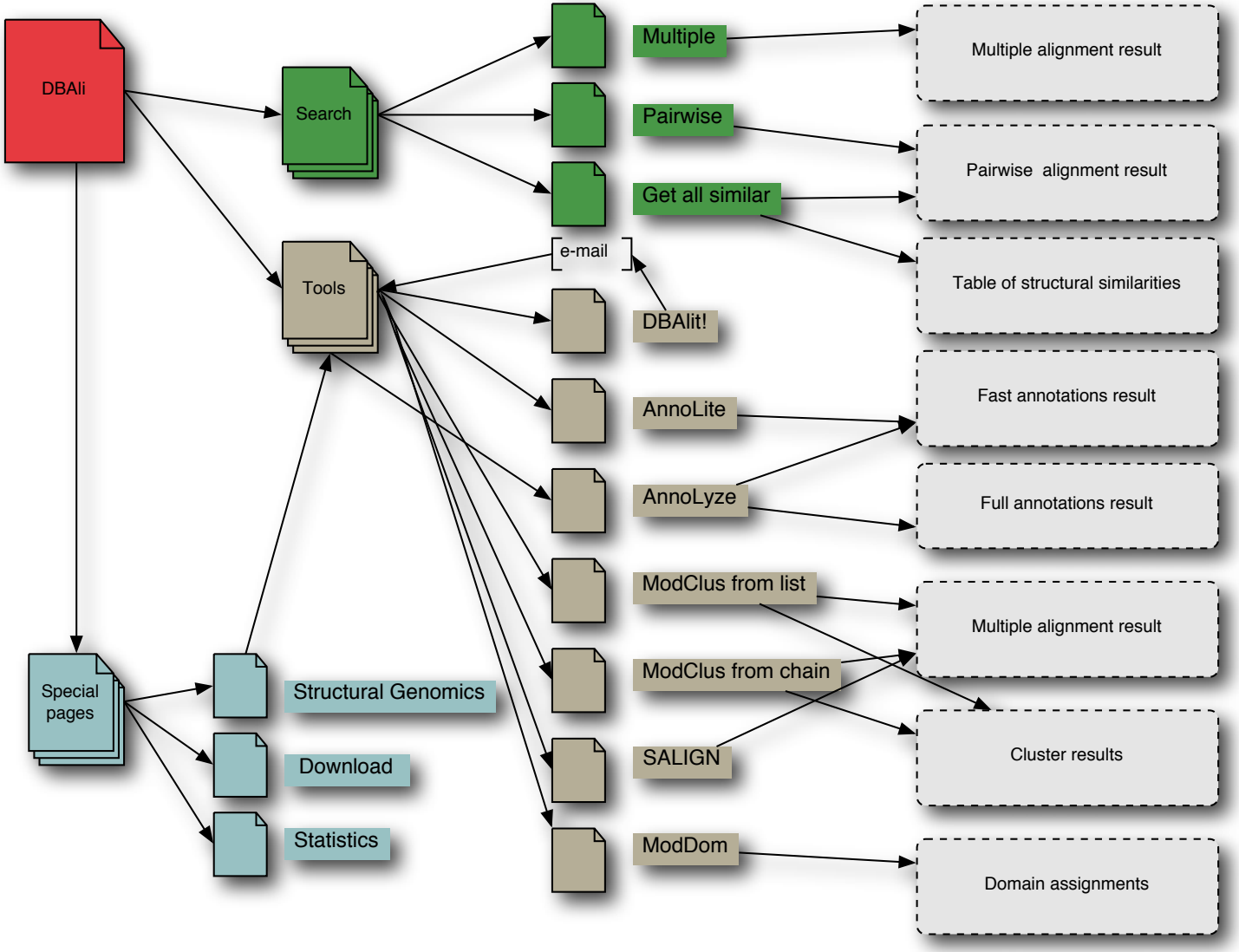
Protein function from structure

Comparative annotation. AnnoLite and AnnoLyze.

Marti-Renom et al. BMC Bioinformatics (2007) in press

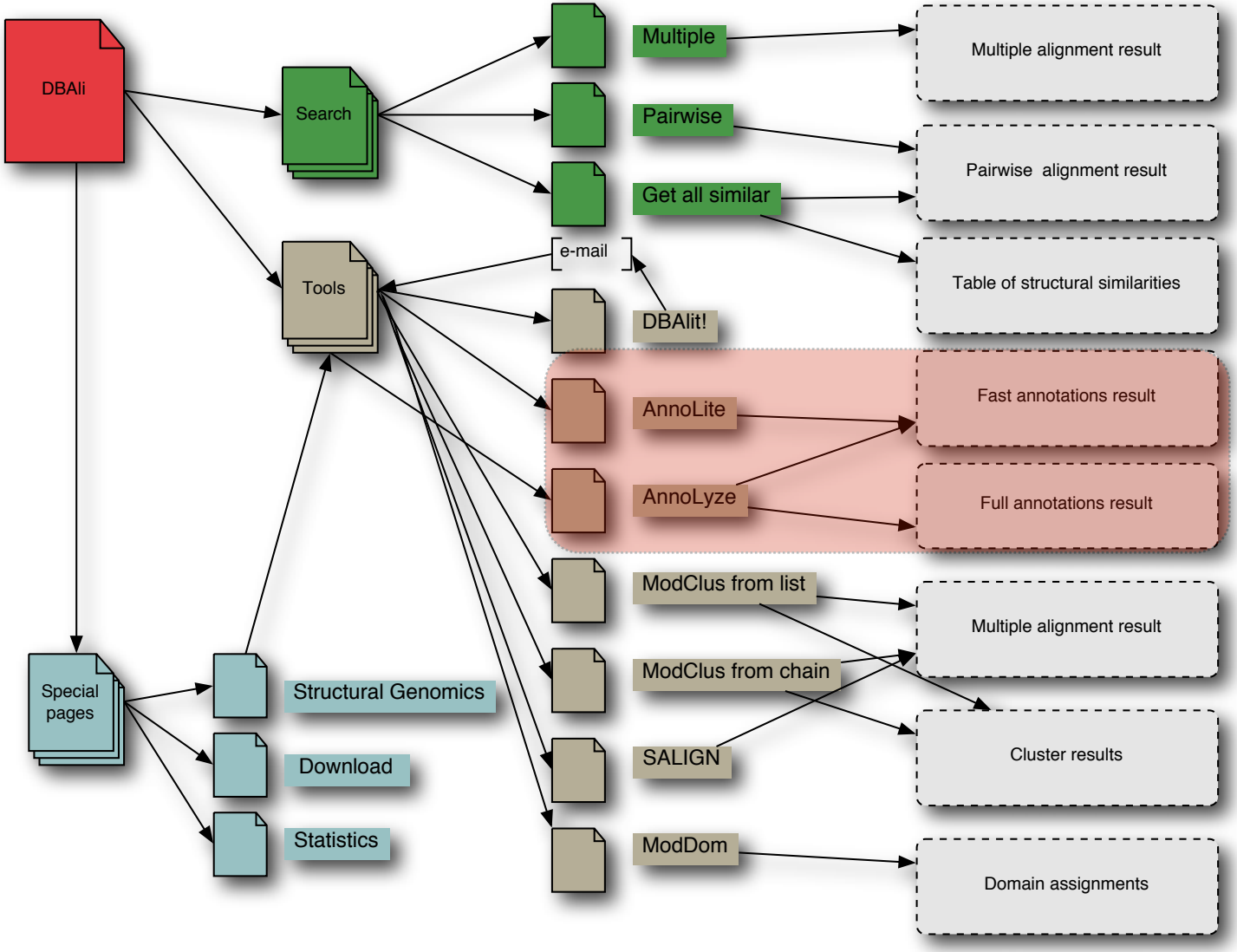
DBAli$_{v2.0}$ database

http://bioinfo.cipf.es/squ/services/DBAli/
http://www.salilab.org/DBAli/



DBAli$_{v2.0}$ database

http://bioinfo.cipf.es/squ/services/DBAli/
http://www.salilab.org/DBAli/



AnnoLite

	Conf.	P-value	Link	Description
CATH:		7.5e-99	2.70.100.10	1,4-Beta-D-Glucan Cellobiohydrolase I, subunit A
SCOP:		0.00	b.29.1.10	Glycosyl hydrolase family 7 catalytic core
PFAM:		0.00	PF00840	Glycosyl hydrolase family 7
InterPro:		1.3e-99	IPR001722	Glycoside hydrolase, family 7
		6.0e-51	IPR008985	Concanavalin A-like lectin/glucanase
		1.0e-42	IPR000254	Cellulose-binding region, fungal
EC Number:		1.2e-44	3.2.1.91	Cellulose 1,4-beta-cellobiosidase.
		6.0e-41	3.2.1.4	Cellulase.
GO Molecular Function:		6.0e-36	0030248	cellulose binding
		8.4e-36	0016162	cellulose 1,4-beta-cellobiosidase activity
		1.0e-35	0004553	hydrolase activity, hydrolyzing O-glycosyl compounds
		1.4e-30	0008810	cellulase activity
		3.1e-20	0016798	hydrolase activity, acting on glycosyl bonds
		1.0e+0	0016787	hydrolase activity
GO Biological Process:		1.1e-63	0030245	cellulose catabolism
		1.2e-54	0000272	polysaccharide catabolism
		3.6e-20	0005975	carbohydrate metabolism
GO Cellular Component:		1.2e-23	0005576	extracellular region

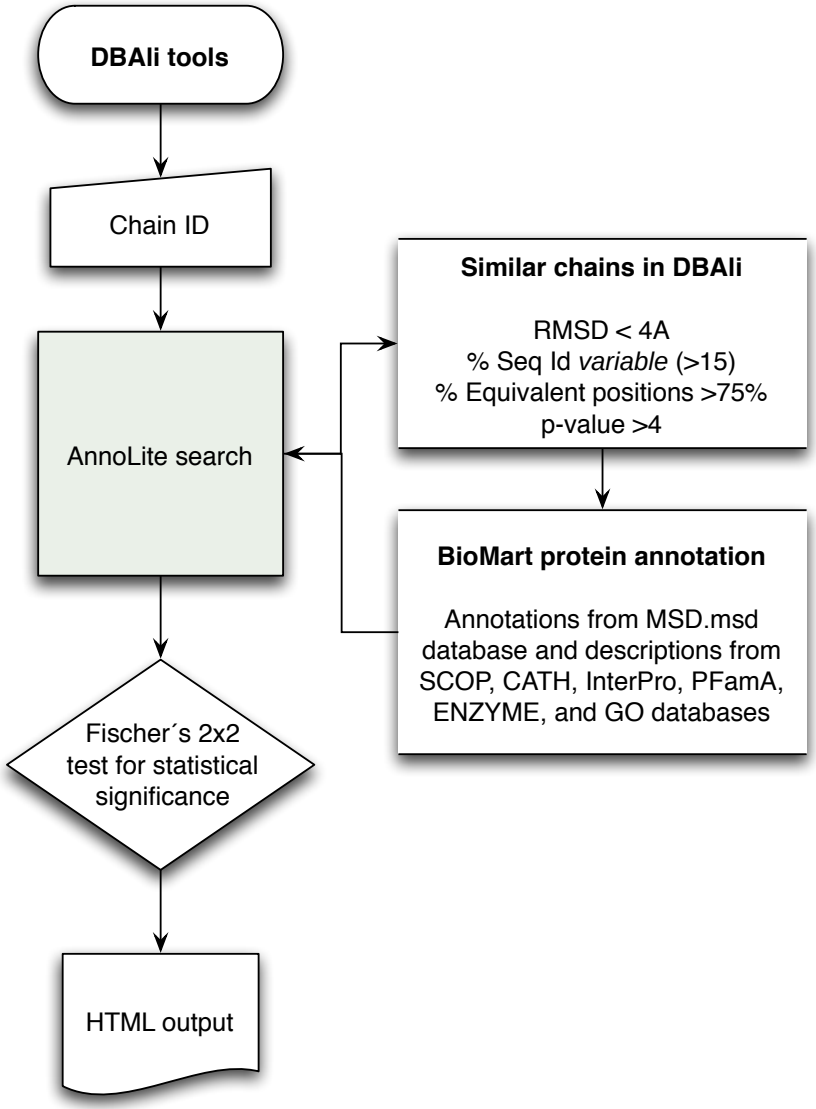
- Information annotated in the MSD database.
- High, medium and low confidence annotations not annotated in the MSD database.
- High, medium and low confidence annotations already annotated in the MSD database.

Benchmark set

	Number of chains
Initial set*	50,223
FULL annotation**	10,997
Non-redundant set***	1,879

**data from BioMart MSD.3 (release February 2005)*
***annotated with CATH, SCOP, Pfam, EC, InterPro, and GO terms in the MSD database*
***not two chains can be structurally aligned within 2A, superimposing more than 60% of their Cα atoms and have a length difference inferior to 30aa*

Method



AnnoLite results for chain 1gpi:A based on 44 structural similar chains.

	Conf.	P-value	Link	Description
CATH:		7.5e-99	2.70.100.10	1,4-Beta-D-Glucan Cellobiohydrolase I, subunit A
SCOP:		0.00	b.29.1.10	Glycosyl hydrolase family 7 catalytic core
PFAM:		0.00	PF00840	Glycosyl hydrolase family 7
InterPro:		1.3e-99	IPR001722	Glycoside hydrolase, family 7
		6.0e-51	IPR008985	Concanavalin A-like lectin/glucanase
		1.0e-42	IPR000254	Cellulose-binding region, fungal
EC Number:		1.2e-44	3.2.1.91	Cellulose 1,4-beta-cellobiosidase.
		6.0e-41	3.2.1.4	Cellulase.
GO Molecular Function:		6.0e-36	0030248	cellulose binding
		8.4e-36	0016162	cellulose 1,4-beta-cellobiosidase activity
		1.0e-35	0004553	hydrolase activity, hydrolyzing O-glycosyl compounds
		1.4e-30	0008810	cellulase activity
		3.1e-20	0016798	hydrolase activity, acting on glycosyl bonds
		1.0e+0	0016787	hydrolase activity
GO Biological Process:		1.1e-63	0030245	cellulose catabolism
		1.2e-54	0000272	polysaccharide catabolism
		3.6e-20	0005975	carbohydrate metabolism
GO Cellular Component:		1.2e-23	0005576	extracellular region

Information annotated in the MSD database.
High, medium and low confidence annotations not annotated in the MSD database.
High, medium and low confidence annotations already annotated in the MSD database.

Scoring function

Fisher's 2x2 contingency test

	Non-similar	Similar	Total
Annotated	a	b	a+b
Not Annotated	c	d	c+d
Total	a+c	b+d	n

1b78A SCOP c.51.4.1	Similar	Not similar	Total
Annotated	4	2	6
Not Annotated	0	71,096	71,096
Total	4	71,098	71,102

$$p = \binom{a+b}{a}\binom{c+d}{c} \Big/ \binom{n}{a+c}$$

$$= \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

$$p = 1.78e^{-19}$$

Sensitivity .vs. Precision

	Optimal cut-off	Sensitivity (%) Recall or TPR	Precision (%)
SCOP fold	1e-6	92.7	88.4
CATH fold	1e-3	95.7	90.1
InterPro	1e-3	88.4	78.2
PFam family	1e-4	90.5	82.8
EC number	1e-4	93.3	79.7
GO Molecular Function	1e-1	84.3	80.9
GO Biological Process	1e-3	85.5	74.8
GO Cellular Component	1e-2	77.6	58.6

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$

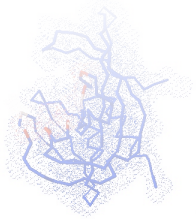
AnnoLyze

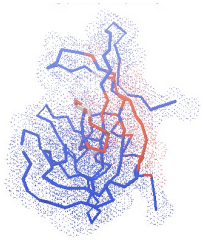
Inherited ligands: 4

Ligand	Av. binding site seq. id.	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
MO2	59.03	0.185	48 49 52 62 63 66 67 113 116
CRY	20.00	0.111	23 29 31 37 44 48 49 83 85 94 96 103 121
BOG	20.00	0.111	19 20 21 48 49 51 96 98 136
ACY	15.87	0.163	23 29 31 37 44 45 81 83 85 94 96 98 103 121 135

Inherited partners:1

Partner	Av. binding site seq. id.	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
d.113.1.1	23.68	0.948	19 20 50 51 52 53 54 55 56 57 58 77 78 79 80 81 82 83 84 85 93 95 97 99 134 135 138 142 145





Benchmark

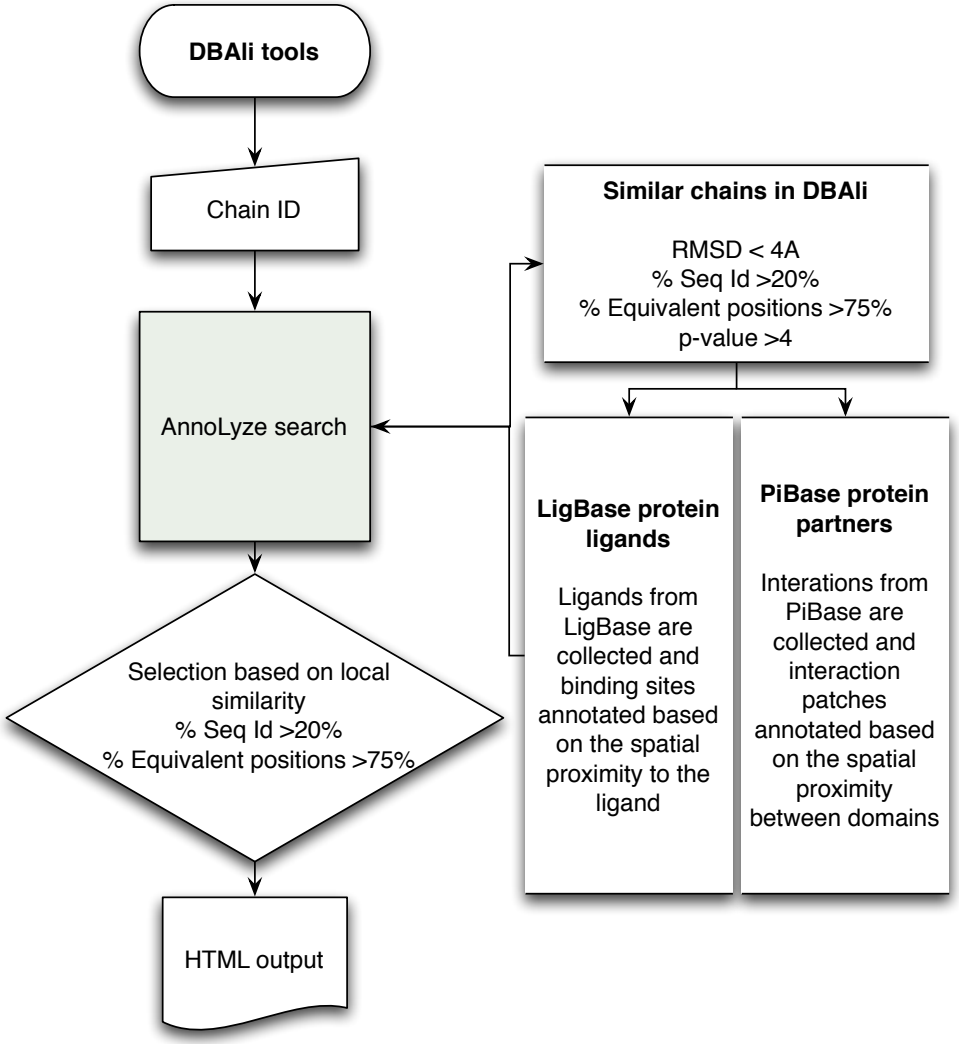
	Number of chains
Initial set*	78,167
LigBase**	30,126
Non-redundant set***	4,948 (8,846 ligands)

*all PDB chains larger than 30 aminoacids in length (8th of August, 2006)
**annotated with at least one ligand in the LigBase database
***not two chains can be structurally aligned within 3A, superimposing more than 75% of their Cα atoms, result in a sequence alignment with more than 30% identity, and have a length difference inferior to 50aa

	Number of chains
Initial set*	78,167
πBase**	30,425
Non-redundant set***	4,613 (11,641 partnerships)

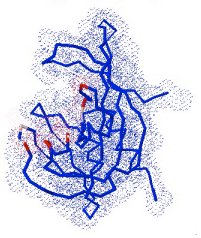
*all PDB chains larger than 30 aminoacids in length (8th of August, 2006)
**annotated with at least one partner in the πBase database
***not two chains can be structurally aligned within 3A, superimposing more than 75% of their Cα atoms, result in a sequence alignment with more than 30% identity, and have a length difference inferior to 50aa

Method



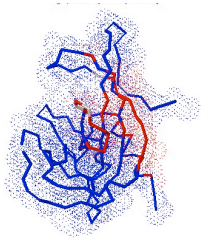
Inherited ligands: 4

Ligand	Av. binding site seq. id.	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
MO2	59.03	0.185	48 49 52 62 63 66 67 113 116
CRY	20.00	0.111	23 29 31 37 44 48 49 83 85 94 96 103 121
8OG	20.00	0.111	19 20 21 48 49 51 96 98 136
ACY	15.87	0.163	23 29 31 37 44 45 81 83 85 94 96 98 103 121 135



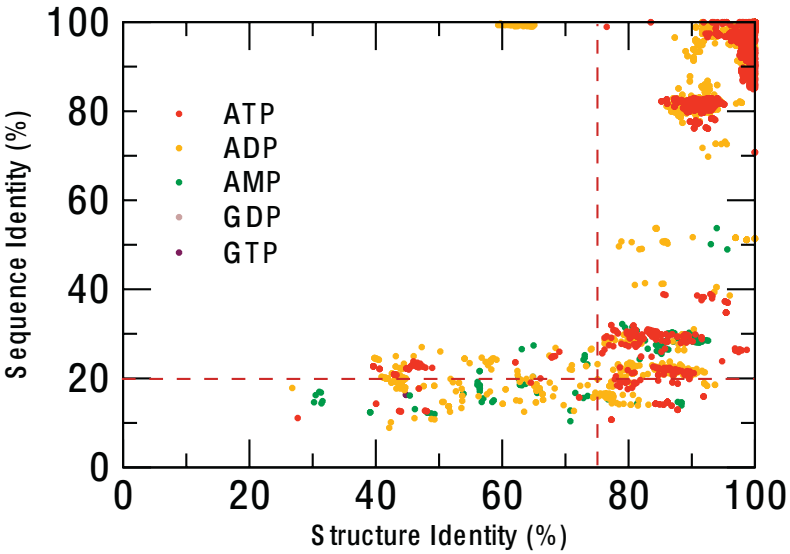
Inherited partners:1

Partner	Av. binding site seq. id.	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
d.113.1.1	23.68	0.948	19 20 50 51 52 53 54 55 56 57 58 77 78 79 80 81 82 83 84 85 93 95 97 99 134 135 138 142 145

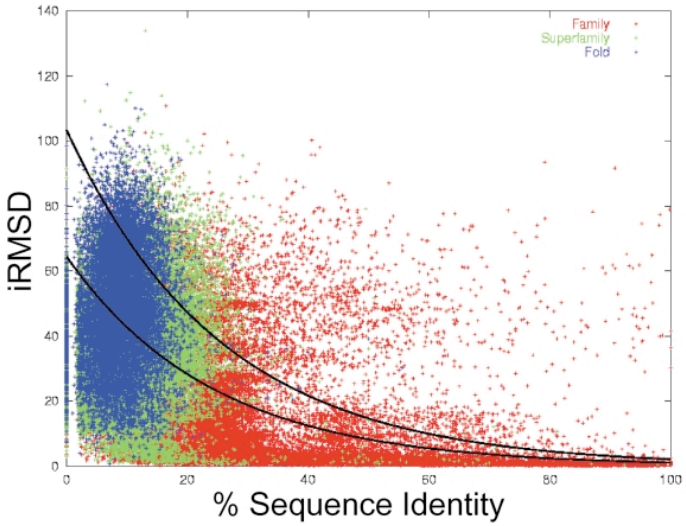


Scoring function

Ligands



Partners



Aloy *et al.* (2003) J.Mol.Biol. 332(5):989-98.

Sensitivity .vs. Precision

	Optimal cut-off	Sensitivity (%) Recall or TPR	Precision (%)
Ligands	30%	71.9	13.7
Partners	40%	72.9	55.7

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$

Example (2azwA)

Structural Genomics Unknown Function

Molecule: MutT/nudix family protein

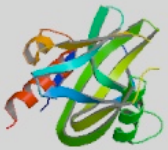
PDB ID: 2azw:A

Header:
STRUCTURAL GENOMICS, UNKNOWN FUNCTION

Compound:
MOL_ID: 1; MOLECULE: MUTT/NUDIX FAMILY PROTEIN; CHAIN: A; ENGINEERED: YES

Source:
MOL_ID: 1; ORGANISM_SCIENTIFIC: ENTEROCOCCUS FAECALIS V583; ORGANISM_COMMON: BACTERIA; EXPRESSION_SYSTEM: ESCHERICHIA COLI; EXPRESSION_SYSTEM_COMMON: BACTERIA; EXPRESSION_SYSTEM_STRAIN: BL21(DE3); EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID; EXPRESSION_SYSTEM_PLASMID: PET15B

Resolution: 1.90Å



Links: none

SCOP: none

CATH: none

Sequence:
Md5: 09b13d23ceae0fdfcaddec636e2ddfa6KTPTAAAS
Length: 146

```
KTPTFGKREE TLTYQTRYAA YIIVSKPENN TMVLVQAPNG AYFLPGGEIE
GTETKEEAIH REVLEELGIS VEIGCYLGEA DEYFYSNHRQ TAYYNPGYFY
VANTWRQLSE PLRTNTLHWV APEEAVRLLK RGSHRWAVEK WLAAAS
```

Ligands: none

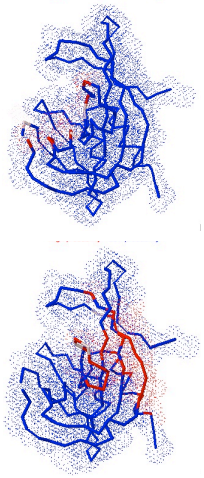
Interacting partners: none

Inherited ligands: 4

Ligand	Av. binding site seq. id.	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
MO2	59.03	0.185	48 49 52 62 63 66 67 113 116
CRY	20.00	0.111	23 29 31 37 44 48 49 83 85 94 96 103 121
8OG	20.00	0.111	19 20 21 48 49 51 96 98 136
ACY	15.87	0.163	23 29 31 37 44 45 81 83 85 94 96 98 103 121 135

Inherited partners: 1

Partner	Av. binding site seq. id.	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
d.113.1.1	23.68	0.948	19 20 50 51 52 53 54 55 56 57 58 77 78 79 80 81 82 83 84 85 93 95 97 99 134 135 138 142 145



Similar structures: 20
Similar sequences: 890

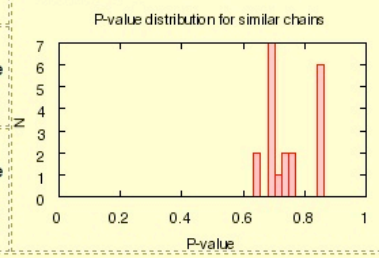
Most similar structure in DBAli:

Code	SeqId(%)	EqPos	RMSD	P-Value	See
1vc9:A	22.76	123	3.57	17.28	ali

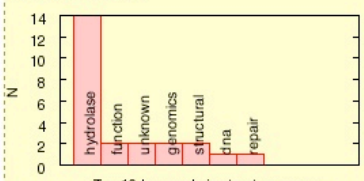
Most similar sequence in DBAli:

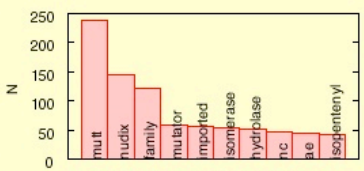
Code	SeqId(%)	EqPos	RMSD	P-Value	See
1vcd:B	24.59	122	3.47	17.00	ali

P-value distribution:

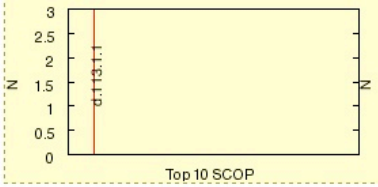


Keyword distribution:



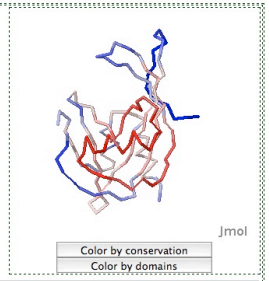


SCOP and CATH distribution for similar structures:



Chain: 2azwA
Equivalences selected: 6783
Coverage: 100.00 %
Number of domains: 1
Assignment score:0.5470
Alternative assignments: here

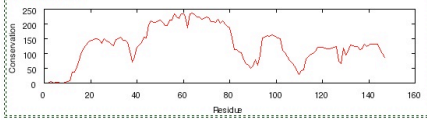
Highest score assignment:
Domain 1:
2-113,115-147
Length: 145
Conservation: 506.49



Protein sequence colored by conservation and underlined by domain color:

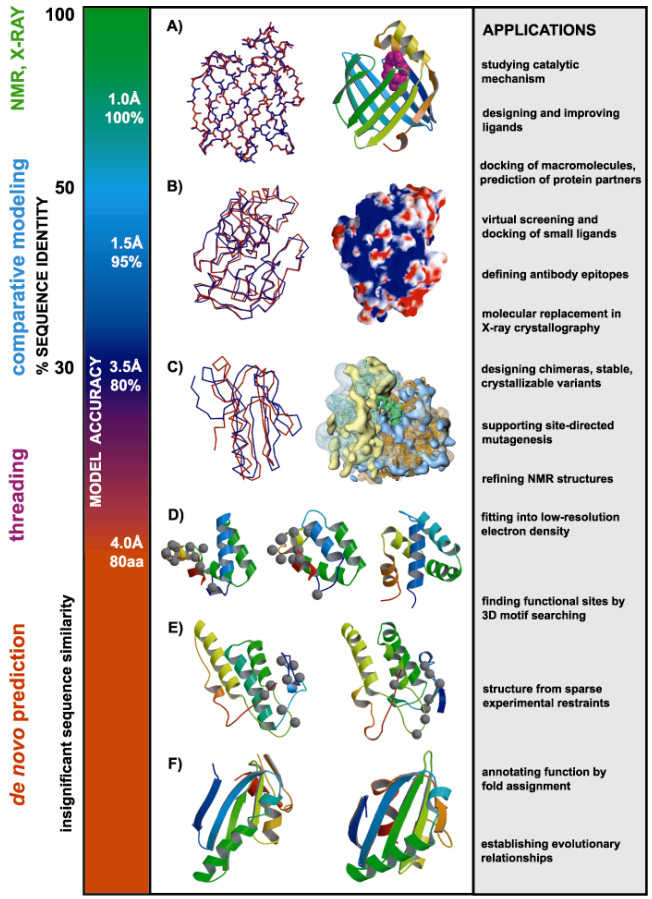
```
KTPTFGKREE TLTYQTRYAA YIIVSKPENN TMVLVQAPNG AYFLPGGEIE GTETKEEAIH REVLEELGIS
VEIGCYLGEA DEYFYSNHRQ TAYYNPGYFY VANTWRQLSE PLRTNTLHWV APEEAVRLLK RGSHRWAVEK
WLAAAS
```

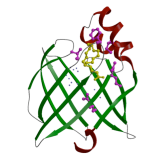
Conservation profile:

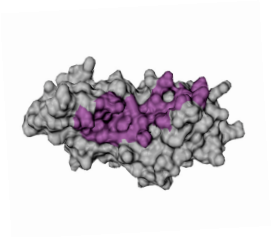


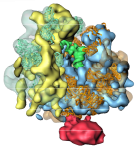
	Conf.	P-value	Link	Description
CATH:		1.1e-20	3.90.79.10	Nucleoside Triphosphate Pyrophosphohydrolase
SCOP:		4.2e-29	d.113.1.1	MutT-like
PFAM:		2.0e-74	PF00293	NUDIX domain
InterPro:		1.9e-65	IPR000086	NUDIX hydrolase
		2.7e-20	IPR003561	Mutator MutT
		2.9e-14	IPR002667	Isopentenyl-diphosphate delta-isomerase
EC Number:		1.7e-4	3.6.1.17	Bis(5'-nucleosyl)-tetraphosphatase (asymmetrical).
GO Molecular Function:		4.5e-19	0008413	8-oxo-7,8-dihydroguanine triphosphatase activity
		3.8e-13	0004452	isopentenyl-diphosphate delta-isomerase activity
		1.9e-6	0016787	hydrolase activity
		5.4e-3	0004081	bis(5'-nucleosyl)-tetraphosphatase (asymmetrical) activity
		1.9e-2	0000287	magnesium ion binding
GO Biological Process:		7.7e-11	0008299	isoprenoid biosynthesis
		1.5e-5	0006974	response to DNA damage stimulus
		1.7e-5	0006260	DNA replication
		2.4e-5	0006281	DNA repair

Can we use models to infer function?

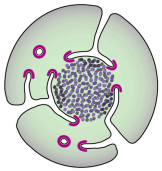








T. cruzi



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

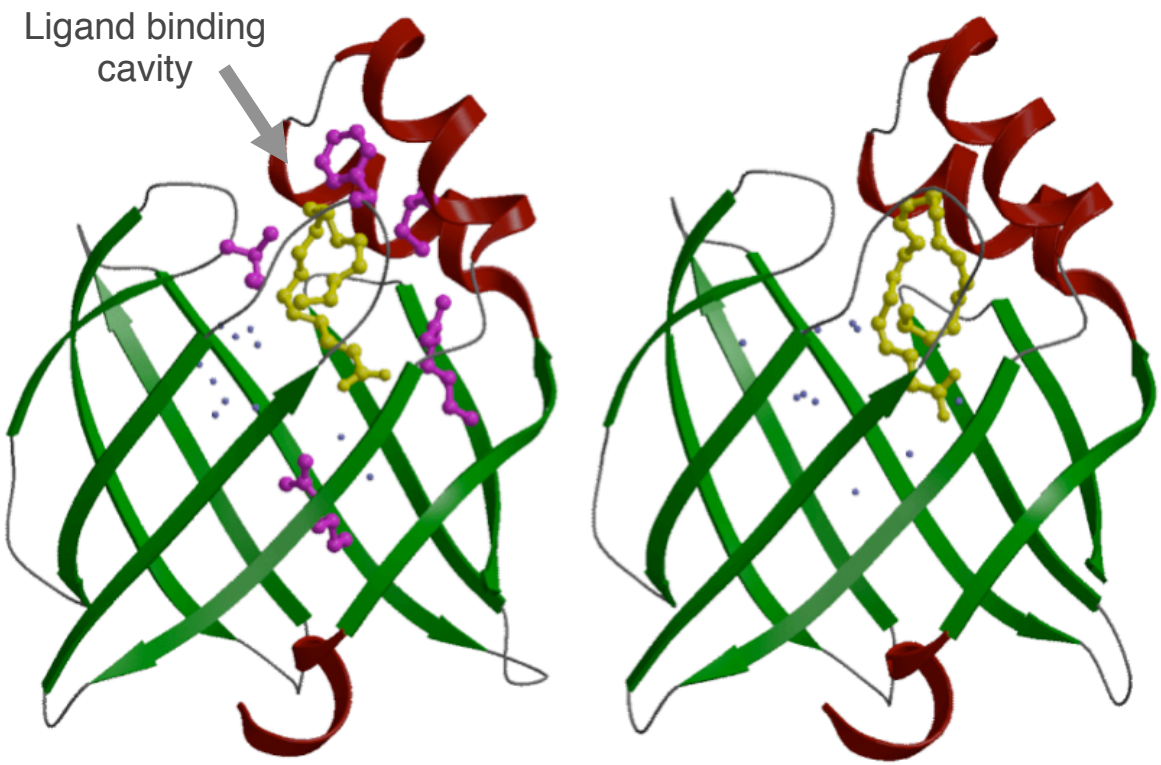
Predicting features of a model that are not present in the template

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 snuggly into the ligand binding cavity.

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

Structural analysis of missense mutations in human BRCA1 BRCT domains

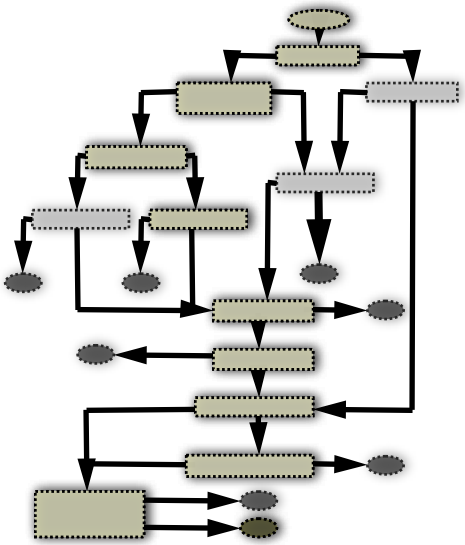
Nebojsa Mirkovic, Marc A. Marti-Renom, Barbara L. Weber, Andrej Sali and Alvaro N.A. Monteiro

Cancer Research (June 2004). 64:3790-97

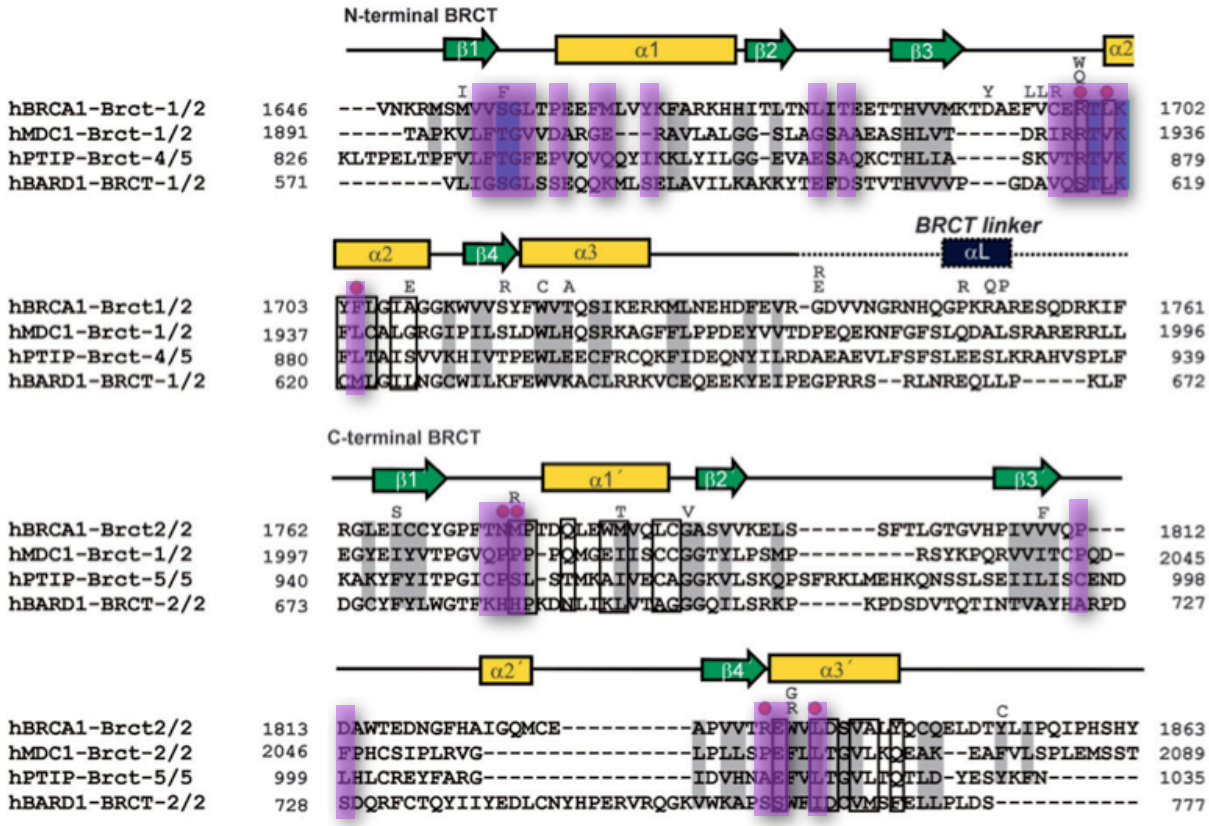
Cannot measure the functional impact of every possible SNP at all positions in each protein! Thus, prediction based on general principles of protein structure is needed.

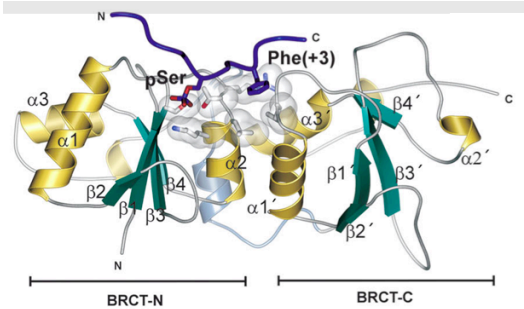
Missense mutations in BRCT domains by function

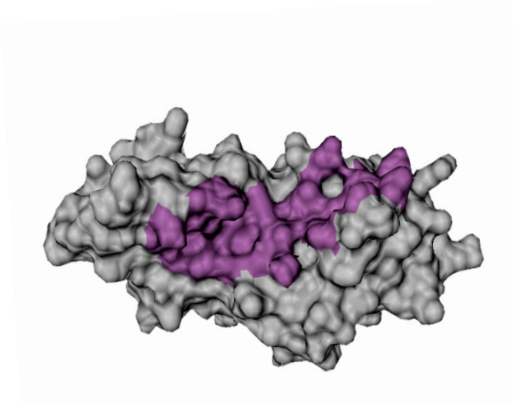
	cancer associate	not cancer associated	?
no transcription activation	C1697R R1699W A1708E S1715R P1749R M1775R		M1652K L1657P E1660G H1686Q R1699Q K1702E Y1703HF 1704S L1705PS 1715NS1 722FF17 34LG173 8EG1743 RA1752 PF1761I F1761S M1775E M1775K L1780P I1807S V1833E A1843T
transcription activation		M1652I A1669S	V1665M D1692N G1706A D1733G M1775V P1806A
?			M1652T V1653M L1664P T1685A T1685I M1689R D1692Y F1695L V1696L R1699L G1706E W1718C W1718S T1720A W1730S F1734S E1735K V1736A G1738R D1739E D1739G D1739Y V1741G H1746N R1751P R1751Q R1758G L1764P I1766S P1771L T1773S P1776S D1778N D1778G D1778H M1783T C1787S G1788D G1788V G1803A V1804D V1808A V1809A V1809F V1810G Q1811R P1812S N1819S A1823T V1833M W1837R W1837G S1841N A1843P T1852S P1856T P1859R



Putative binding site on BRCA1





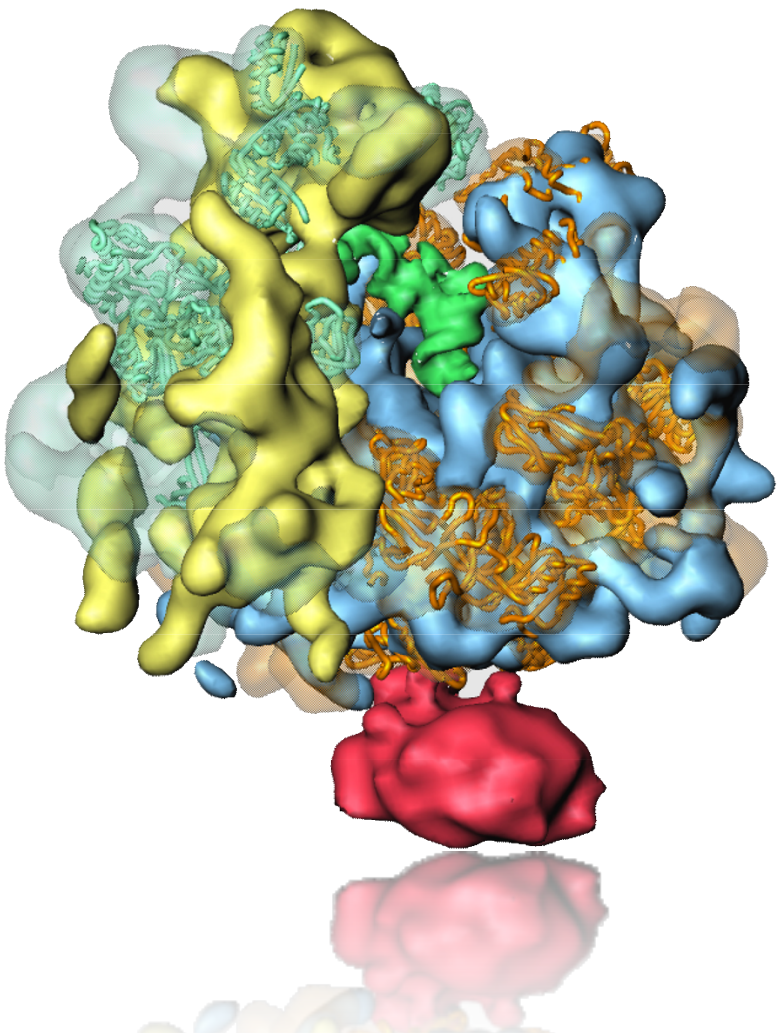


Putative binding site predicted in 2003 and accepted for publication on March 2004.

Williams *et al.* 2004 Nature Structure Biology. **June 2004 11**:519

Mirkovic *et al.* 2004 Cancer Research. **June 2004 64**:3790

S. cerevisiae ribosome



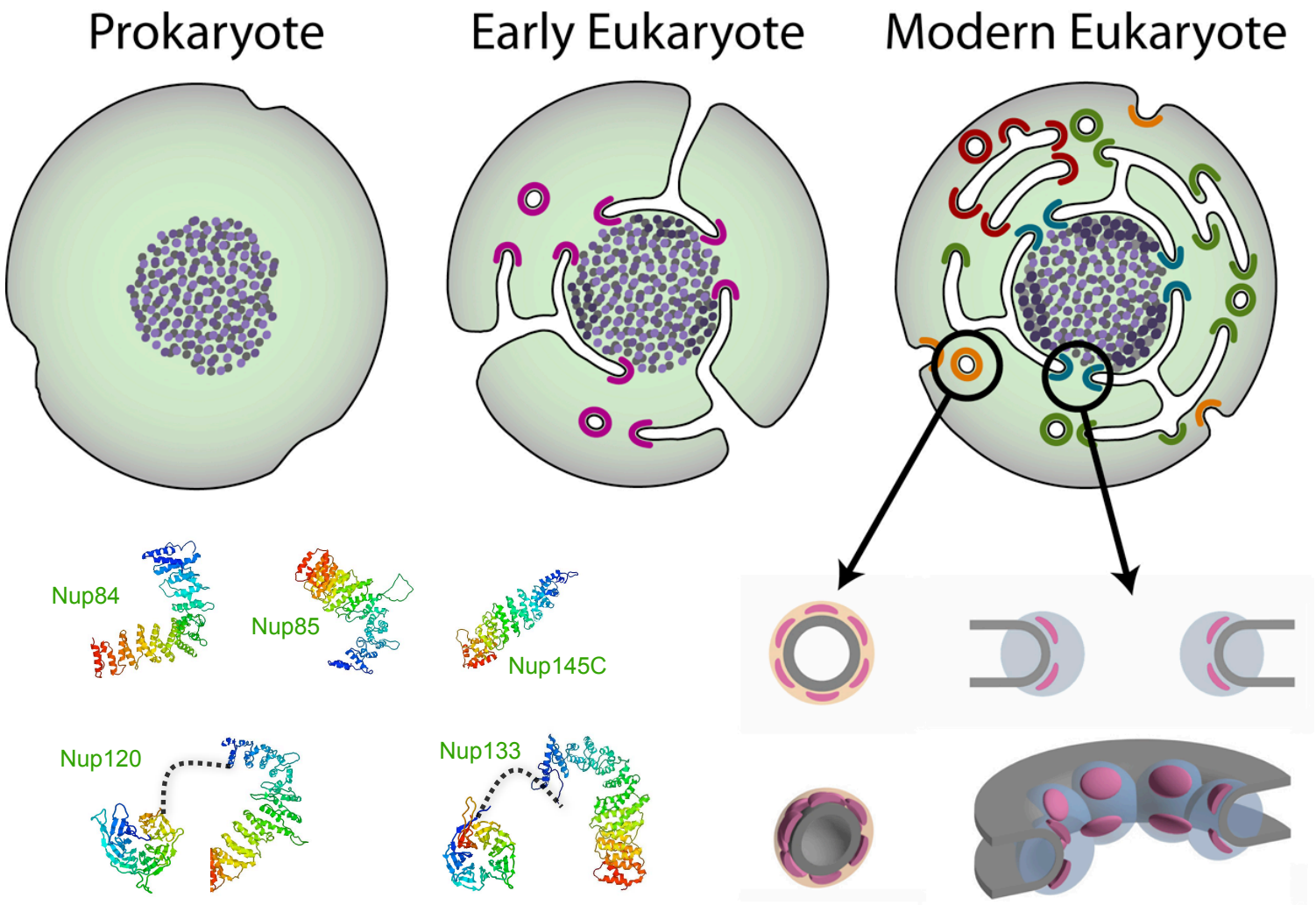
Fitting of comparative models into 15Å cryo-electron density map.

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.

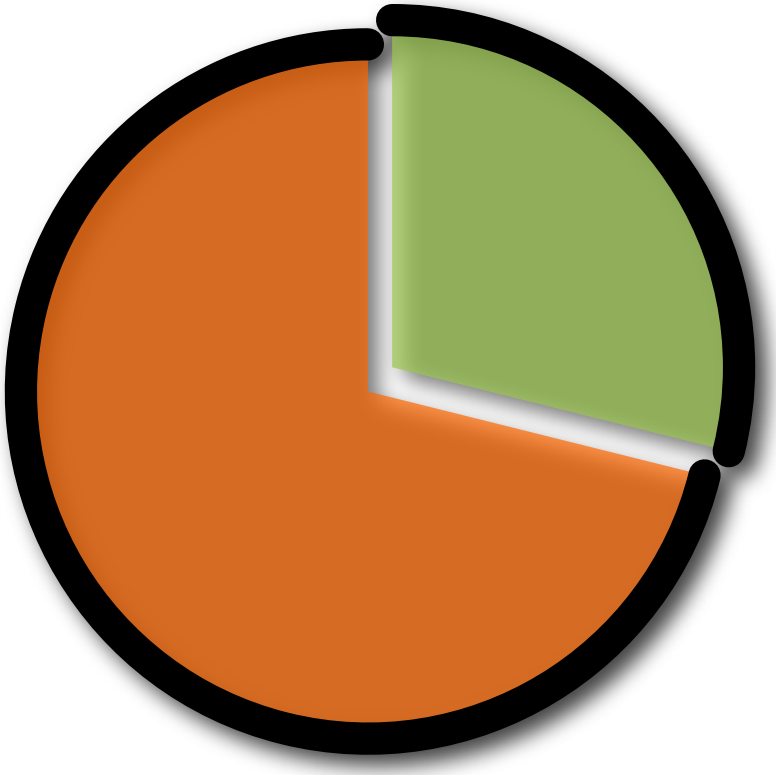
The Nucleopore complex
Cell evolution (?)



Devos et al. PLoS Biology 2, 1 (2004)

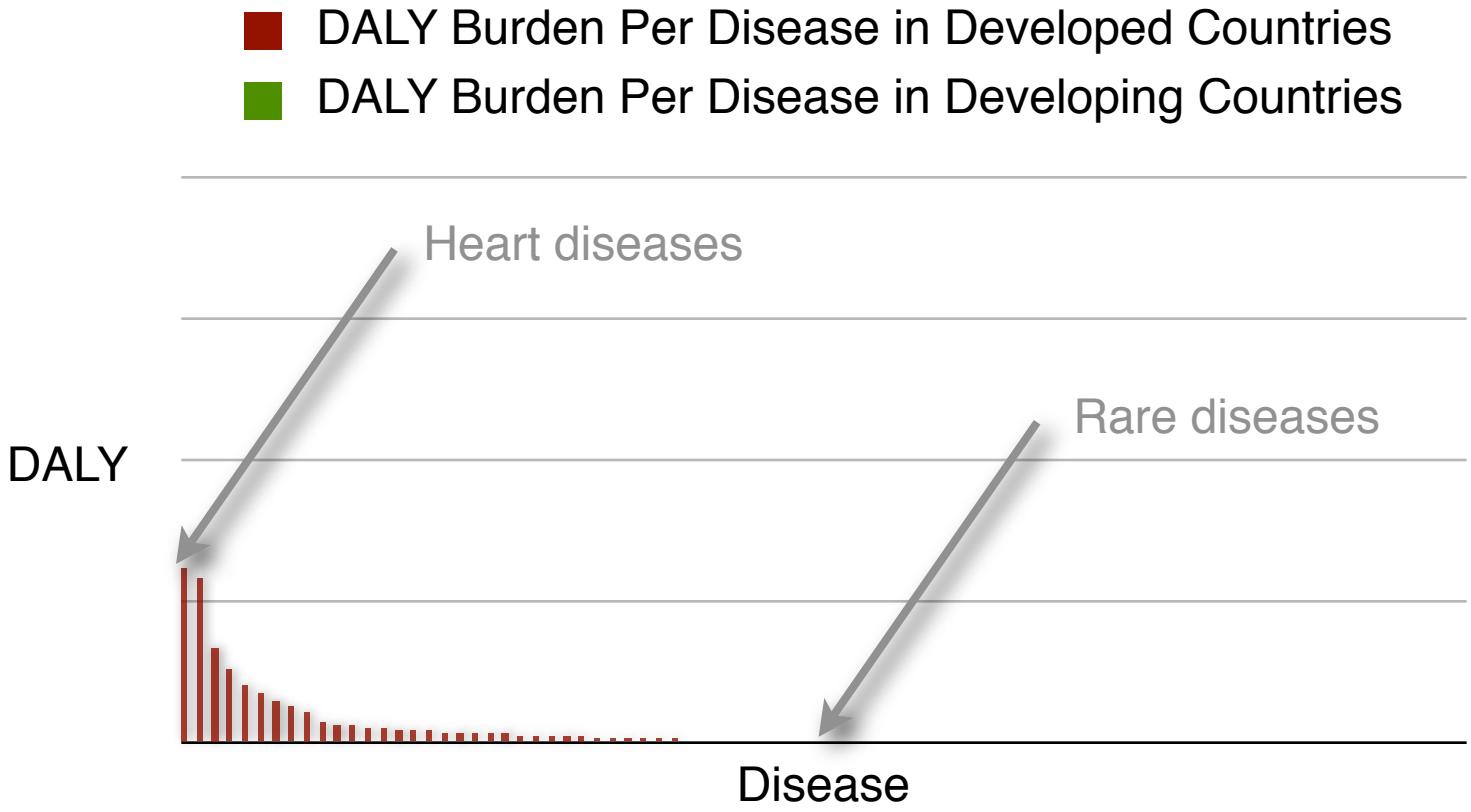
Tropical Disease Initiative (TDI)

Predicting binding sites in protein structure models.



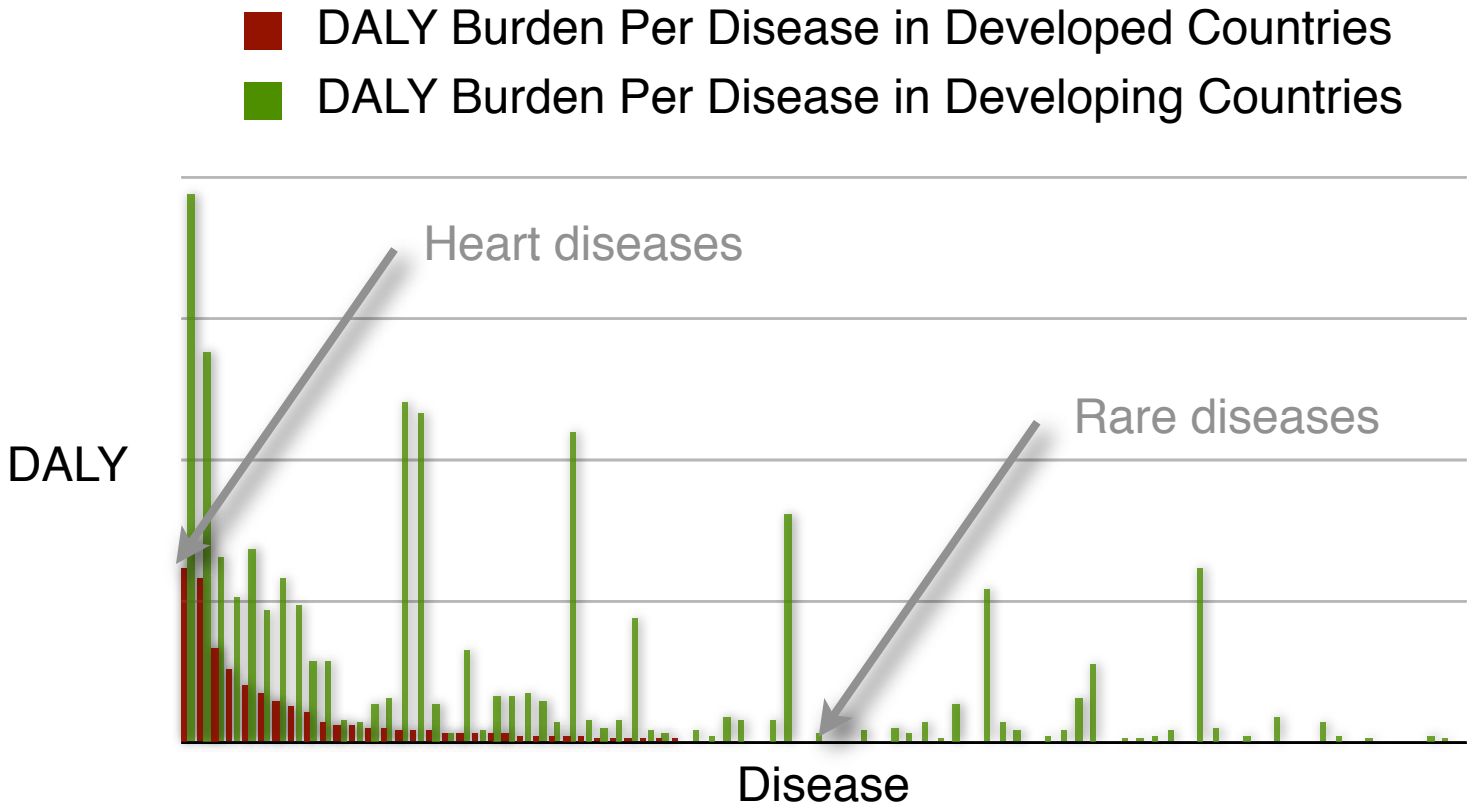
http://www.tropicaldisease.org

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.

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.

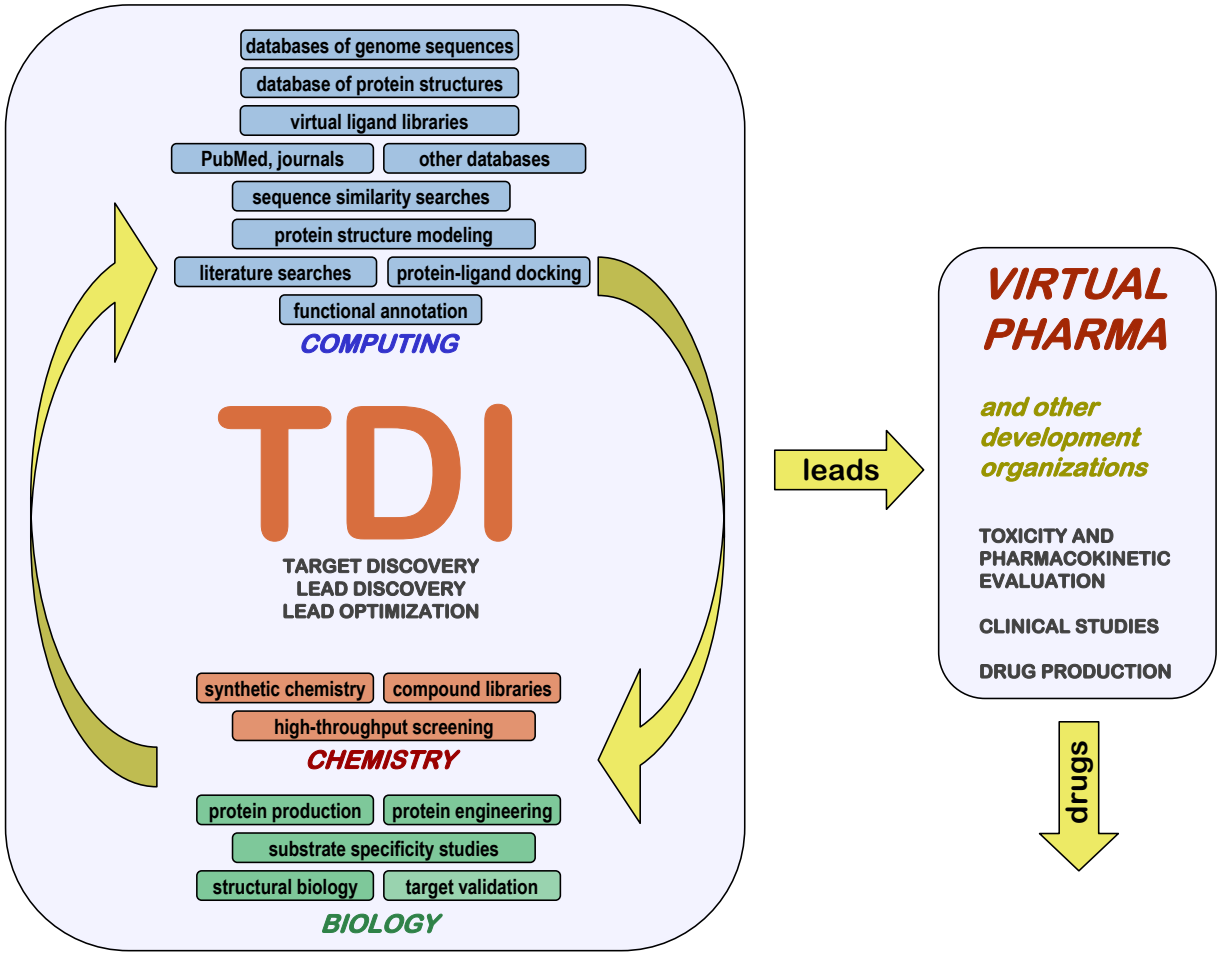
"Unprofitable" Diseases and Global DALY (in 1000's)

Malaria*	**46,486**	Trichuriasis	1,006
Tetanus	7,074	Japanese encephalitis	709
Lymphatic filariasis*	**5,777**	**Chagas Disease***	**667**
Syphilis	4,200	**Dengue***	**616**
Trachoma	2,329	**Onchocerciasis***	**484**
Leishmaniasis*	**2,090**	**Leprosy***	**199**
Ascariasis	1,817	Diphtheria	185
Schistosomiasis*	**1,702**	Poliomyelitise	151
Trypanosomiasis*	**1,525**	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.

TDI flowchart

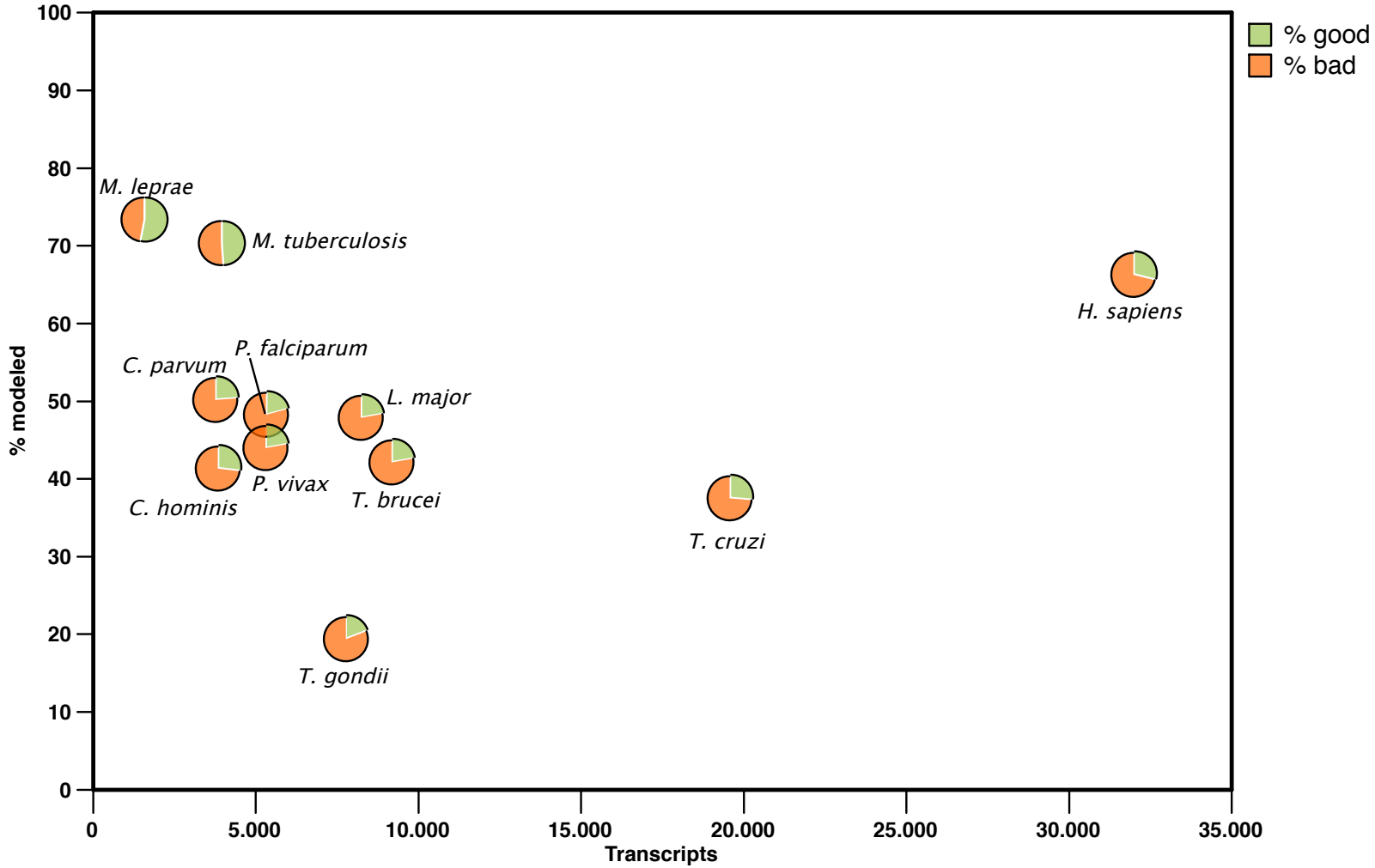
http://www.tropicaldisease.org



Sali, Rai, Maurer. PLoS Medicine (2004)
Kepler, et al. Australian Journal of Chemistry (2006)

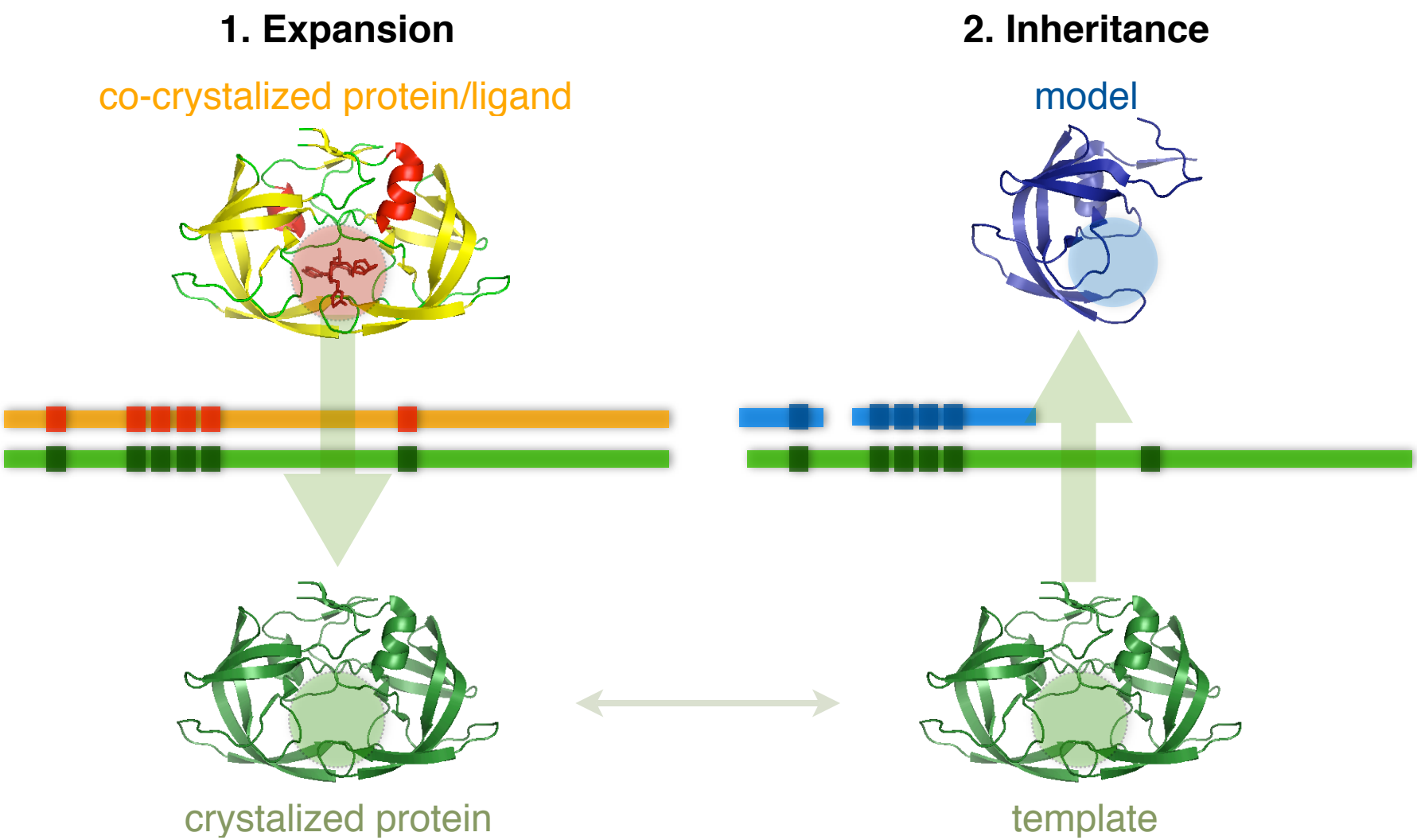
Modeling Genomes

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



A good model has MPQS of 1.1 or higher

Comparative docking



Ligand “expanded” space

from 6,859 templates used in “good” models

Expansion cut-off	Templates	Expanded	Unique
30%	4,639	64,800	3,178
50%	4,242	37,945	3,030
70%	3,323	20,603	2,786

Ligand “inherited” space

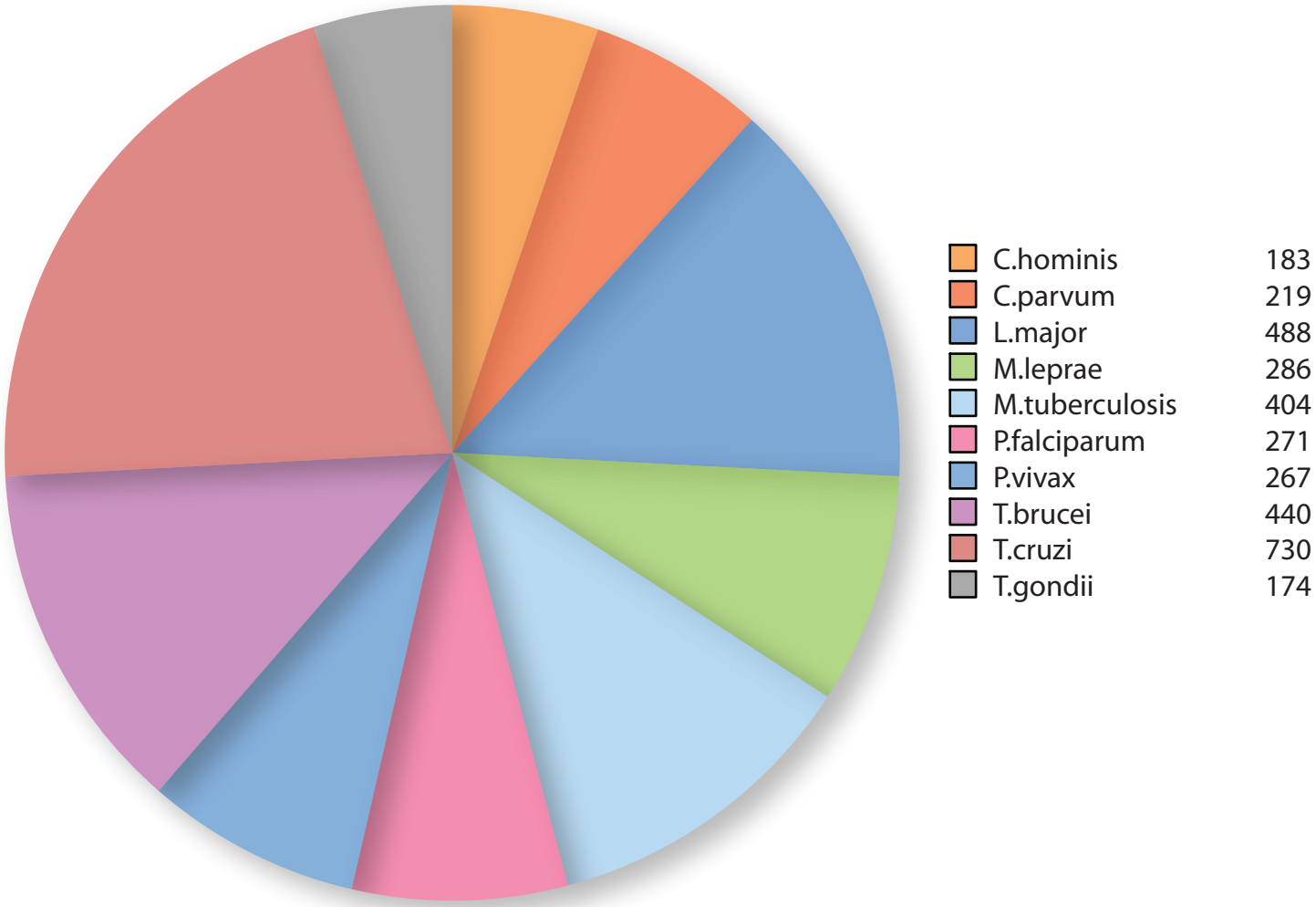
second cut-offs

Using a 70% “expansion” cut-off

Inheritance cut-offs	Models	Inherited	Unique
90% / 70%	5,181	23,286	1,137
90% / 80%	4,383	17,842	1,027
90% / 90%	3,462	11,803	827

Distribution of models with inherited ligands

from 3,882 “good” models
using a 90% / 90% “inherited” cut-offs



Summary table

models with inherited ligands

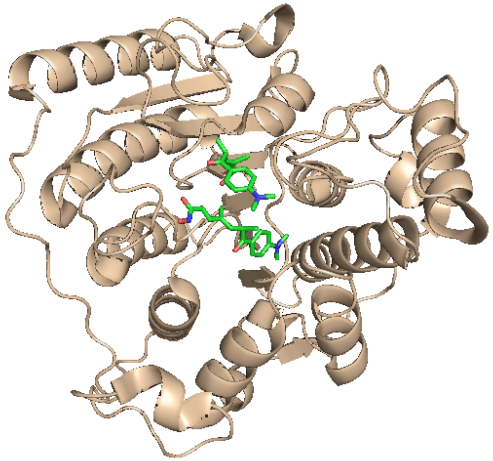
from 16,284 good models, 295 inherited a ligand/substance with at least one compound already approved by FDA and ready to be used from ZINC

	Transcripts	Good	Ligands	Lipinski	Lipinski+ZINC	FDA+ZINC
C. hominis	3,886	886	183	131	28	12 (10)
C. parvum	3,806	949	219	145	30	12 (10)
L. major	8,274	1,845	488	334	84	44 (34)
M. leprae	1,605	1,321	286	189	39	29 (25)
M. tuberculosis	3,991	2,887	404	285	71	44 (37)
P. falciparum	5,363	1,057	271	191	48	20 (16)
P. vivax	5,342	1,042	267	177	37	18 (15)
T. brucei	921	1,795	440	309	94	46 (36)
T. cruzi	19,607	3,915	730	493	127	62 (52)
T. gondii	7,793	587	174	124	28	8 (7)
TOTAL	**60,588**	**16,284**	**3,462**	**2,378**	**586**	**295 (242)**

Example of inheritance (expansion)

LmjF21.0680 from L. major *"Histone deacetylase 2" (model 1)*

Template 1t64A a human HDAC8 protein.



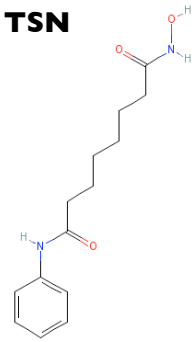
	Origen	Formula	Name	Cov.	Seq, Id. (%)
ZN	X-ray	Zn^{2+}	Zinc ion	--	--
NA	X-ray	Na^{+}	Sodium ion	--	--
CA	X-ray	Ca^{2+}	Calcium ion	--	--
TSN	X-ray	$C_{17}H_{22}N_2O_3$	Trichostatin A	--	--
SHH	Expanded	$C_{14}H_{20}N_2O_3$	Octadenioic acid hudroxyamide phenylamide	100.00	83.8

Example of inheritance (inheritance)

LmjF21.0680 from L. major *"Histone deacetylase 2" (model 1)*

	Formula	Name	Cov.	Seq, Id. (%)	Residues
TSN	$C_{17}H_{22}N_2O_3$	Trichostatin A	100.00	90.9	90 131 132 140 141 167 169 256 263 293 295
SHH	$C_{14}H_{20}N_2O_3$	Octadenioic acid hudroxyamide phenylamide	100.00	90.9	

TSN



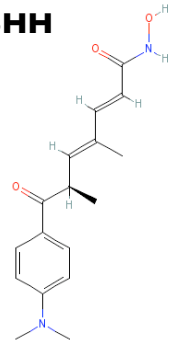
suberoylanilide hydroxamic acid

Pharmacological Action:

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

Inhibits histone deacetylase I and 3

SHH



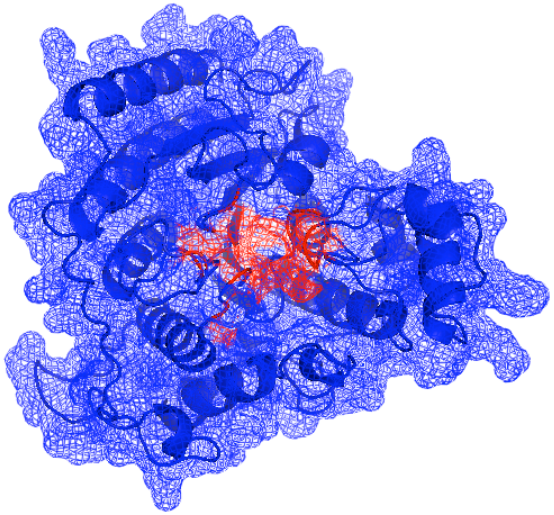
trichostatin A

Pharmacological Action:

- Antibiotics, Antifungal
- Enzyme Inhibitors
- Protein Synthesis Inhibitors

chelates zinc ion in the active site of histone deacetylases, resulting in preventing histone unpacking so DNA is less available for transcription

	LmjF21.0680.1.pdb
Template	1t64A
Seq. Id (%)	**38.00**
MPQS	1.47



Example of inheritance (CDD-Roos-literature)

LmjF21.0680 from L. major *"Histone deacetylase 2" (model 1)*

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)

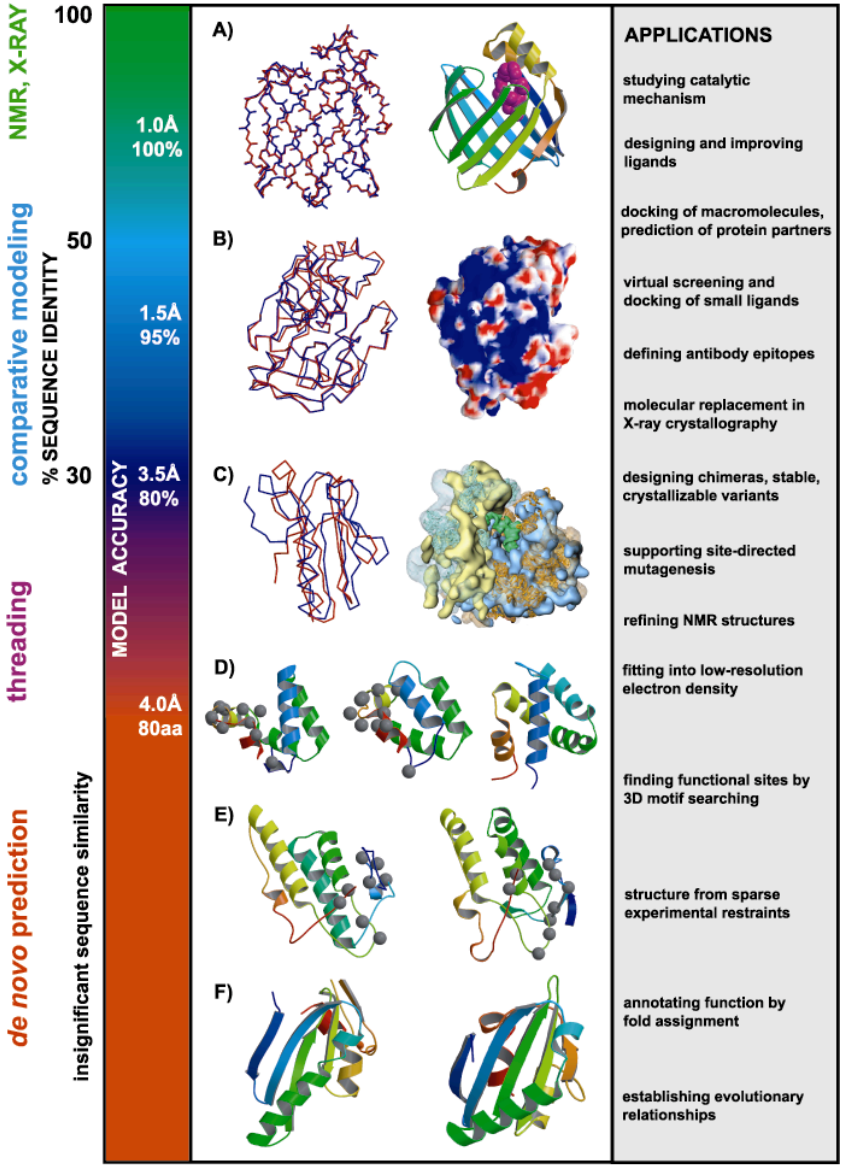
SANDRA J. DARKIN-RATTRAY*†, ANNE M. GURNETT*, ROBERT W. MYERS*, PAULA M. DULSKI*, TAMI M. CRUMLEY*, JOHN J. ALLOCCO*, CHRISTINE CANNOVA*, PETER T. MEINKE‡, STEVEN L. COLLETTI‡, MARIA A. BEDNAREK‡, SHEO B. SINGH§, MICHAEL A. GOETZ§, ANNE W. DOMBROWSKI§, JON D. POLISHOOK§, AND DENNIS M. SCHMATZ*

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

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

“take home” message



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

Comparative Protein Structure Prediction

MODELLER tutorial

```
$>mod9v1 model.py
```

Marc A. Marti-Renom

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

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

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Obtaining MODELLER and related information

- MODELLER (9v1) web page
- http://www.salilab.org/modeller/
- http://www.salilab.org/modeller/tutorial
  - Download Software (Linux/Windows/Mac/Solaris)
  - HTML Manual
  - **Join Mailing List**

Using MODELLER

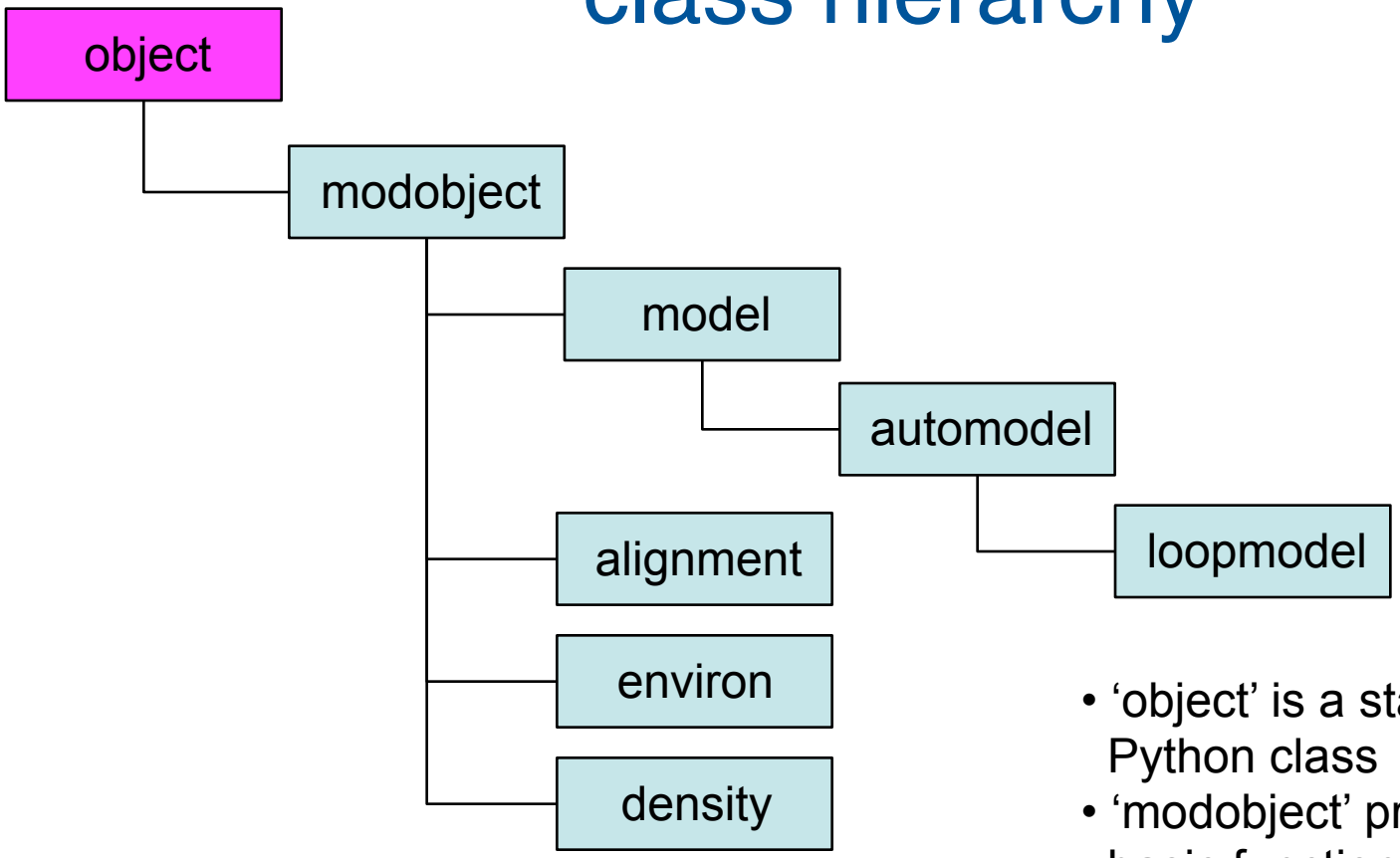
- No GUI! ☹
- Controlled by command file ☹☹
- Script is written in PYTHON language ☺
- You may know Python language is simple ☺☺

MODELLER 9v1
Python interface

- Modeller Python interface uses classes, e.g.:
  - ‘alignment’ *holds and manipulates aligned sequences*
  - ‘model’ *holds and manipulates protein models*
  - ‘environ’ *keeps the configuration of the environment*
  - ‘profile’ *holds and manipulates sequence profiles*
  - ‘sequence_db’ *is for sequence databases*
- These behave just like ordinary Python classes, but Modeller Fortran code is linked to them
- The Modeller data is automatically freed when the Python object is deleted (explicitly or implicitly)

MODELLER 8

class hierarchy



- ‘object’ is a standard Python class
- ‘modobject’ provides basic functions for most Modeller classes
- Not all classes are shown in this diagram

Using MODELLER

- INPUT:
  - Target Sequence (FASTA/PIR format)
  - Template Structure (PDB format)
  - Python file
- OUTPUT:
  - Target-Template Alignment
  - Model in PDB format
  - Other data

Modeling of BLBP
Input

- Target: Brain lipid-binding protein (BLBP)
- BLBP sequence in PIR (MODELLER) format:

```
>P1;blbp
sequence:blbp:::::::
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSID
DRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v1 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v1 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v1 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v1 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Output

```
>P1;1hms
structureX:1hms:   1 : : 131 : :undefined:undefined:-1.00:-1.00
VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA
DDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE*
>P1;blbp
sequence:blbp:     : :     : : : : 0.00: 0.00
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Output

```
>P1;1hms
structureX:1hms:   1 : : 131 : :undefined:undefined:-1.00:-1.00
VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA
DDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE*
>P1;blbp
sequence:blbp:     : :     : : : : 0.00: 0.00
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Output

```
 _aln.pos          10        20        30        40        50        60
1hms      VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGV
blbp      VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGE
 _consrvd ****  **** ** *** *** **********   **** **   *      *  ******* * **

 _aln.p   70        80        90       100       110       120       130
1hms      EFDETTADDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE
blbp      EFEETSIDDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA
 _consrvd ** **  ***  ** * *** ** * ***** **   **  ***   *** *  *  * ***
```

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

```
# Homology modelling by the automodel class
from modeller.automodel import *     # Load the automodel class
log.verbose()                        # request verbose output
env = environ()                      # create a new MODELLER environment

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # code of the target
a.starting_model= 1                  # index of the first model
a.ending_model  = 1                  # index of the last model
                                     # (determines how many models to calculate)
a.make()                             # do the actual homology modelling
```

Run by typing `mod9v1 model.py` in the directory where you have the python file.
MODELLER will produce a `model.log` file

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

```
# Homology modelling by the automodel class
from modeller.automodel import *     # Load the automodel class
log.verbose()                        # request verbose output
env = environ()                      # create a new MODELLER environment

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # code of the target
a.starting_model= 1                  # index of the first model
a.ending_model  = 1                  # index of the last model
                                     # (determines how many models to calculate)
a.make()                             # do the actual homology modelling
```

Run by typing `mod9v1 model.py` in the directory where you have the python file.
MODELLER will produce a `model.log` file

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

```
# Homology modelling by the automodel class
from modeller.automodel import *     # Load the automodel class
log.verbose()                        # request verbose output
env = environ()                      # create a new MODELLER environment

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # code of the target
a.starting_model= 1                  # index of the first model
a.ending_model  = 1                  # index of the last model
                                     # (determines how many models to calculate)
a.make()                             # do the actual homology modelling
```

Run by typing `mod9v1 model.py` in the directory where you have the python file. MODELLER will produce a `model.log` file

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

PDB file

Can be viewed with Chimera

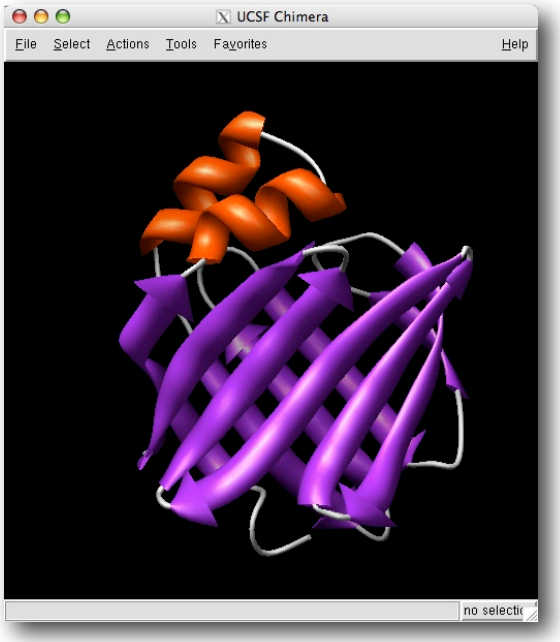
http://www.cgl.ucsf.edu/chimera/

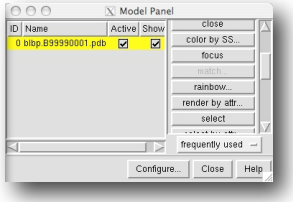
Rasmol

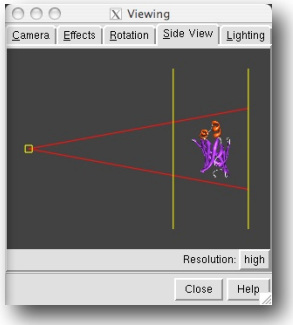
http://www.openrasmol.org

PyMol

http://pymol.sourceforge.net/







Model file →

`blbp.B99990001.pdb`

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

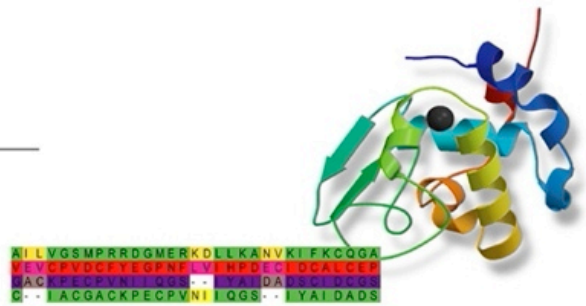
Tutorial

http://salilab.org/modeller/tutorial/

To main Sali lab pages

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



- About MODELLER
- MODELLER News
- Download & Installation
  - Release Notes
- Registration
- Discussion Forum
  - Subscribe
  - Browse archives
  - Search archives
- Documentation
  - FAQ
  - Tutorial
  - Online manual
  - Wiki
- Developers' Pages
- Contact Us

Tutorial

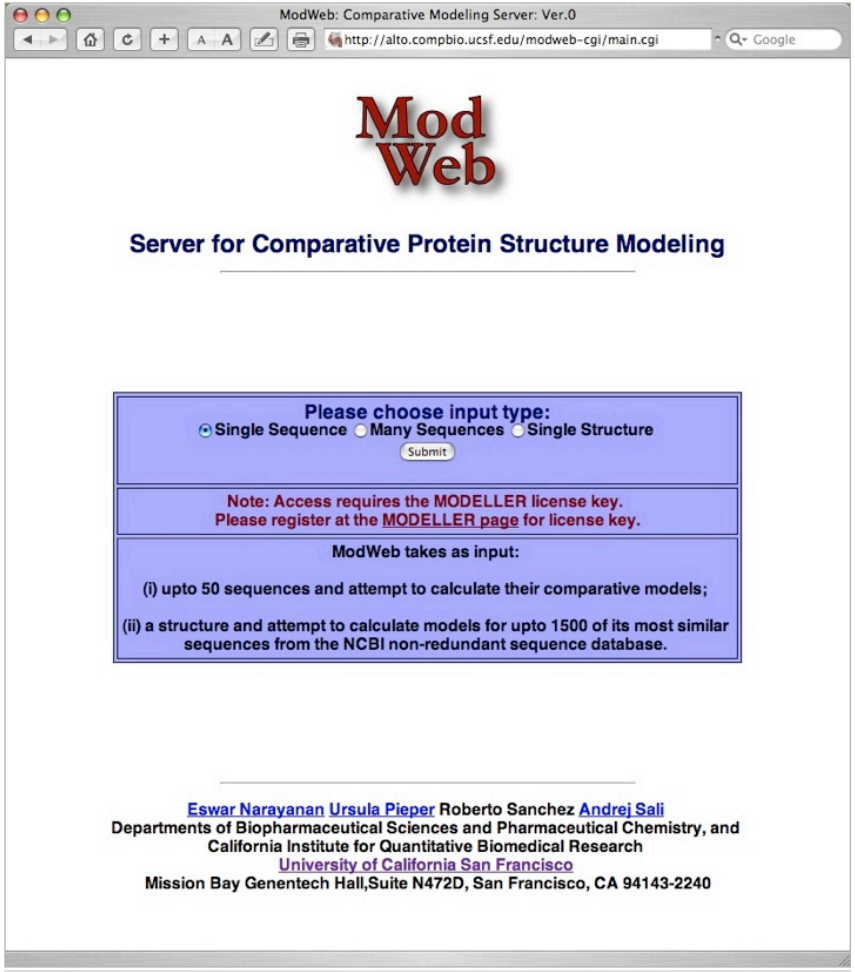
MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2,3). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms.

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 dificult case of protein structure prediction.

MODWEB

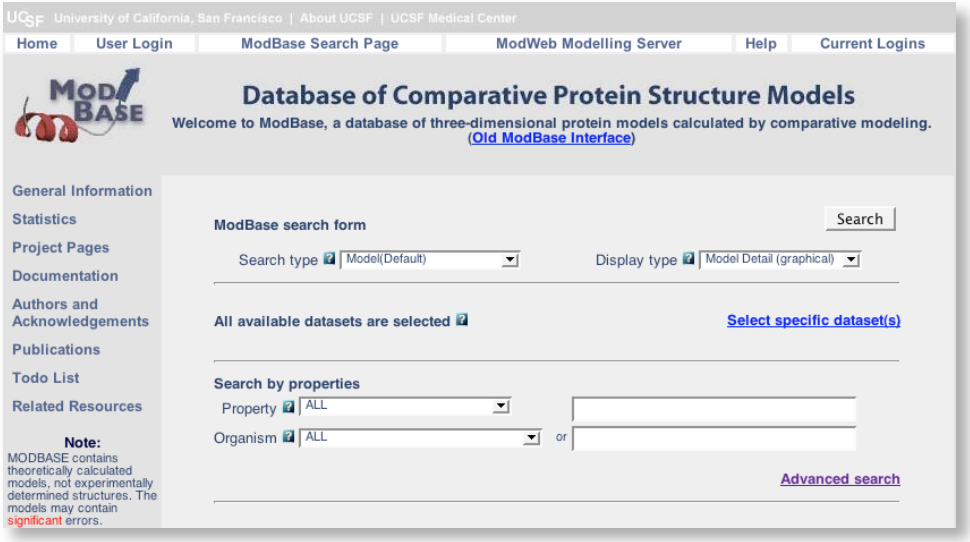
http://salilab.org/modweb



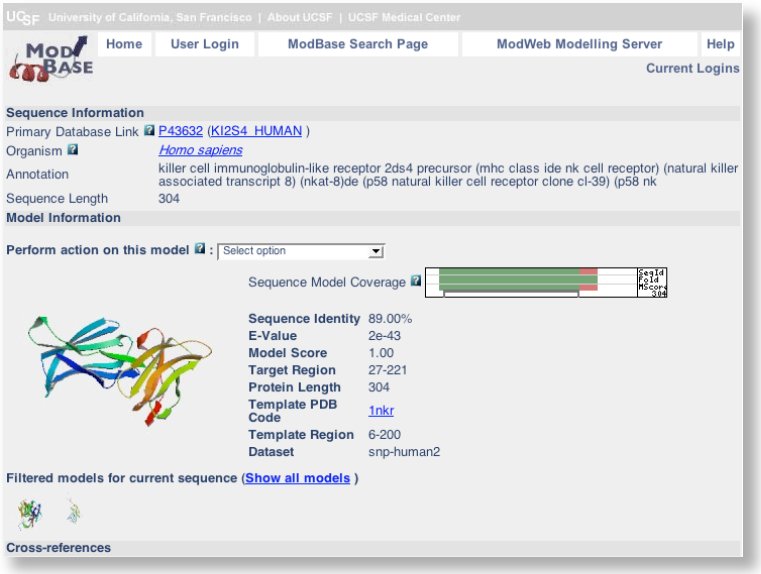
MODBASE

http://salilab.org/modbase

Search Page



Model Details



Sequence Overview

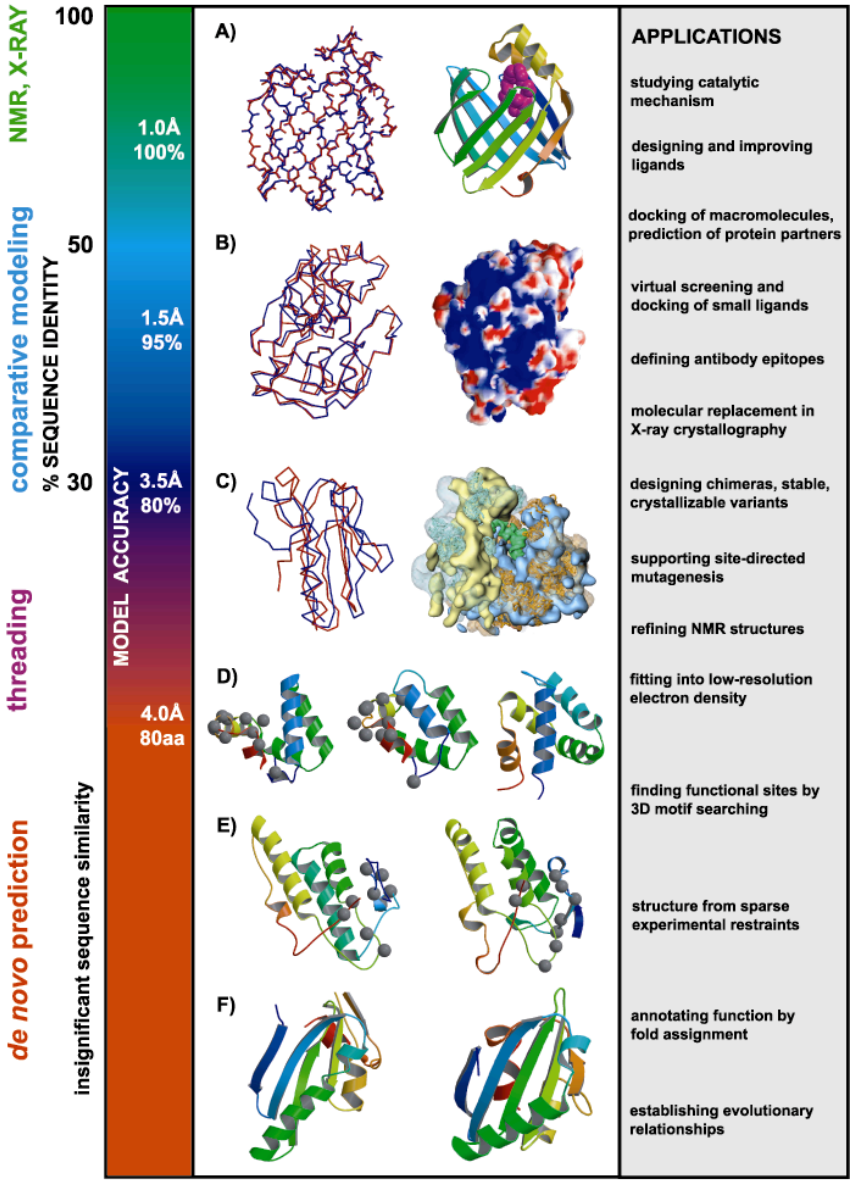
SeqId Fold MScore	Q8G8A6	hypothetical protein	*Pseudomonas aeruginosa*	3738
SeqId Fold MScore	Q8G9W1	hypothetical protein	*Escherichia coli*	1140
SeqId Fold MScore	Q8CY62	hypothetical protein spr1965	*Streptococcus pneumoniae, Streptococcus pneumoniae R6*	1038

Model Overview

Q8G8C7	hypothetical protein	*Pseudomonas aeruginosa*	4996	2089-2158	70	37.00	7e-14	1.00	1dnyA	8-78
Q8G8C7	hypothetical protein	*Pseudomonas aeruginosa*	4996	492-1017	526	36.00	1e-82	1.00	1amuA	19-529
Q8G9W1	hypothetical protein	*Escherichia coli*	1140	349-1135	787	35.00	0	1.00	1r9dA	6-783

Pieper et al. (2004) Nucleic Acids Research 32, D217-D222

>>> TAKE HOME MESSAGE <<<



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

Acknowledgments

COMPARATIVE MODELING
Andrej Sali
M. S. Madhusudhan
Narayanan Eswar
Min-Yi Shen
Ursula Pieper
Ben Webb
Maya Topf

MODEL ASSESSMENT
David Eramian
Min-Yi Shen
Damien Devos

FUNCTIONAL ANNOTATION
Andrea Rossi
Fred Davis

FUNDING
Prince Felipe Research Center
Marie Curie Reintegration Grant
STREP EU Grant

MODEL ASSESSMENT
Francisco Melo (CU)
Alejandro Panjkovich (CU)

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

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)
Barri Bunin (CDD)

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)