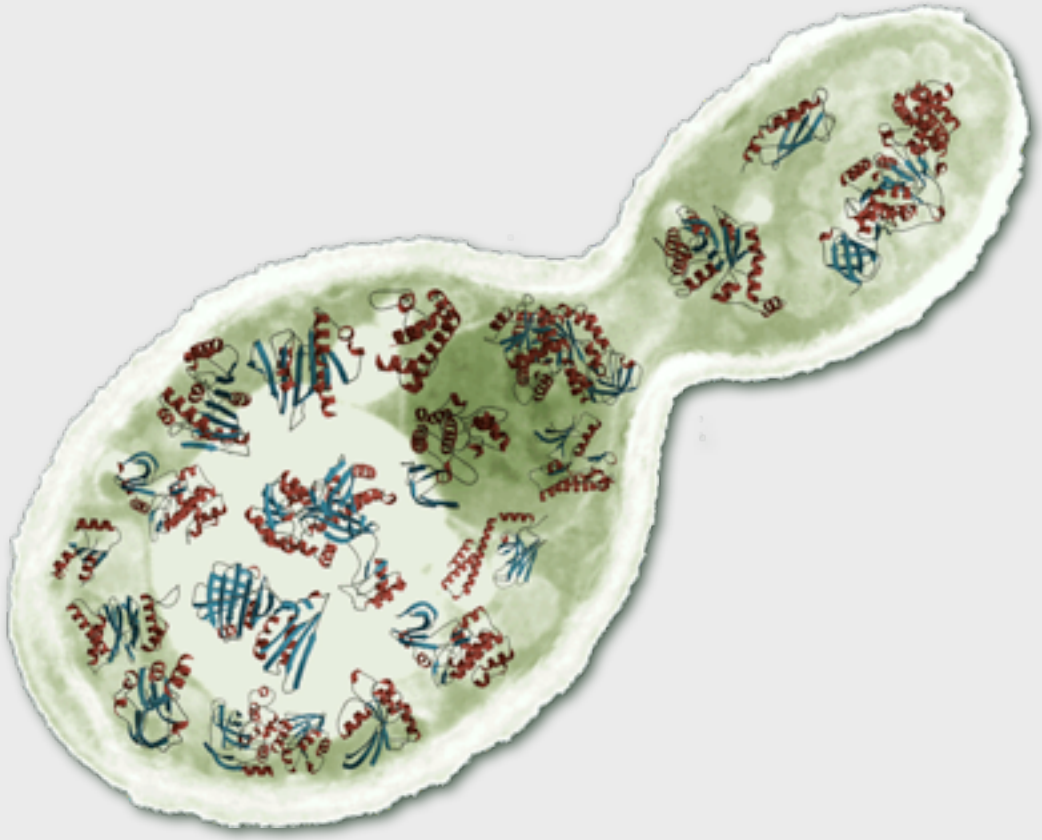
Comparative Protein Structure Prediction



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

Program

Intro to comparative protein structure prediction

- Template Search
- Target – Template Alignment
- Model Building
- Model Evaluation

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

Objective

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

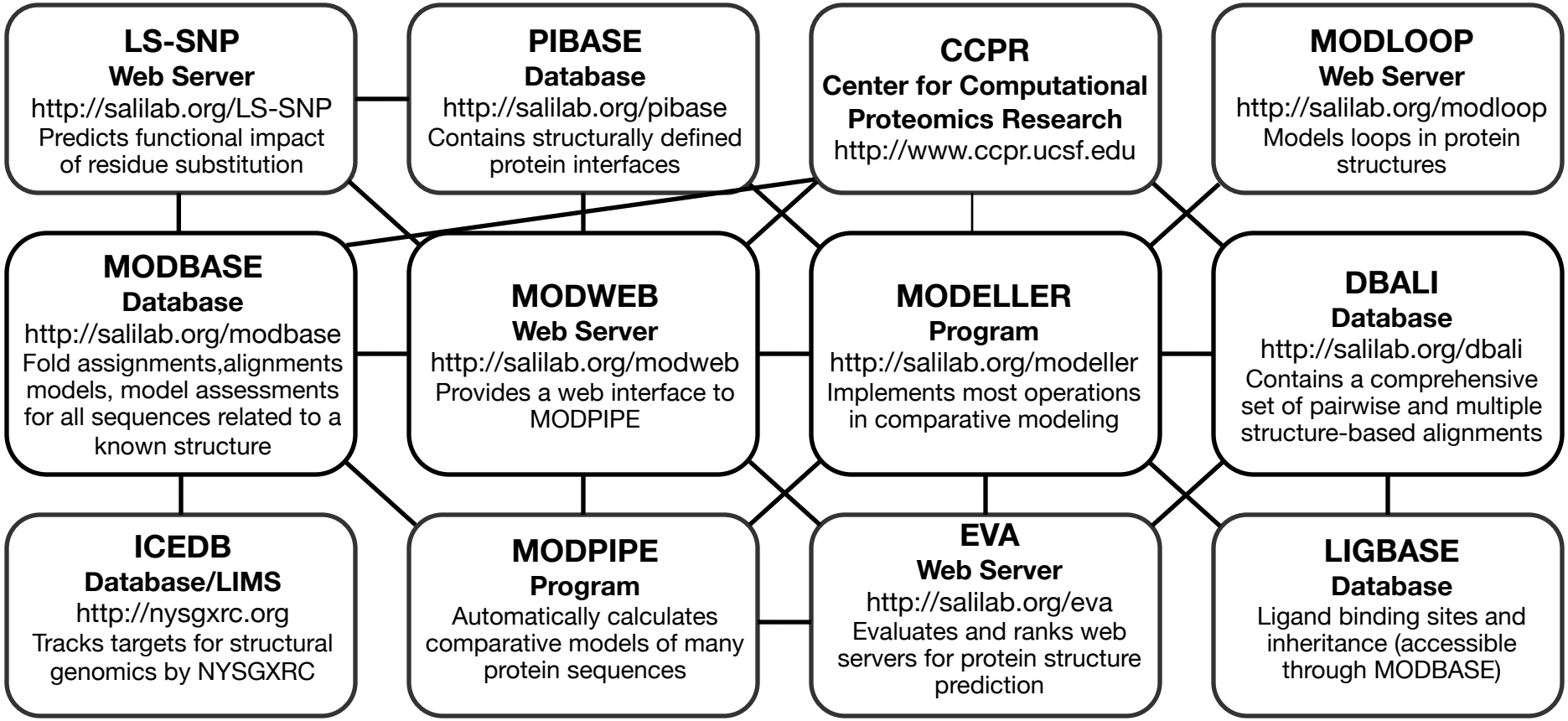
DISCLAIMER!

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

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

Programs, servers and databases

http://salilab.org



External Resources

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

Nomenclature

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

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

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

Target: Sequence corresponding to the protein to be modeled.

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

Model: Predicted 3D structure of the target sequence.

Nomenclature

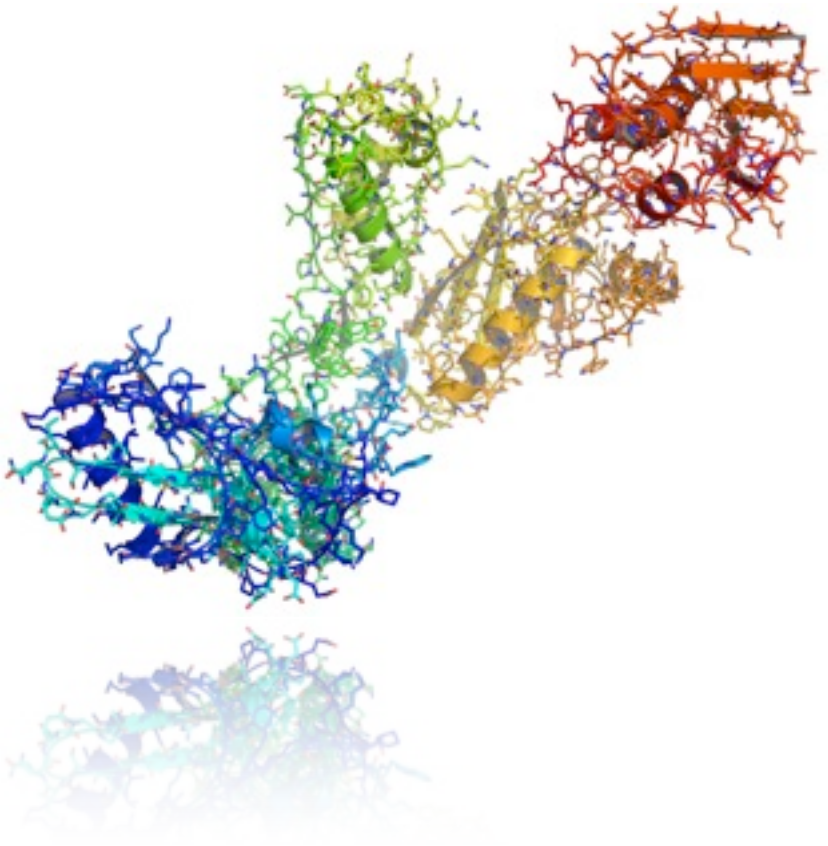
Fold: Three dimensional conformation of a protein sequence (usually at domain level).

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

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

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

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



General References

Protein Structure Prediction:

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

Comparative Modeling:

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

MODELLER:

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

Structural Genomics:

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

protein prediction .vs. protein determination

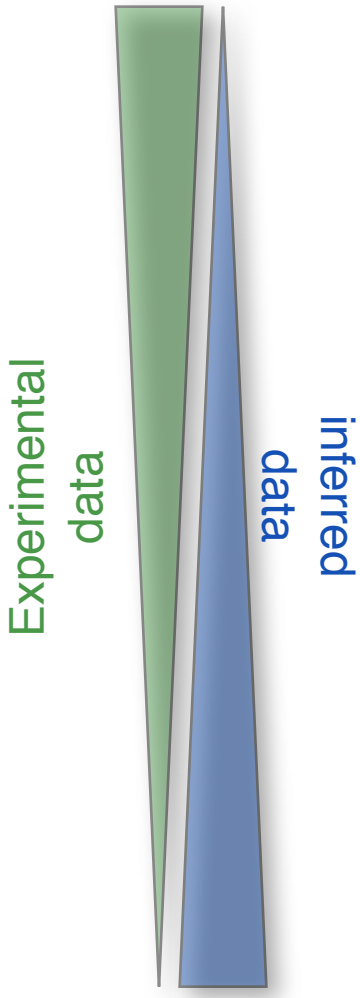
X-Ray

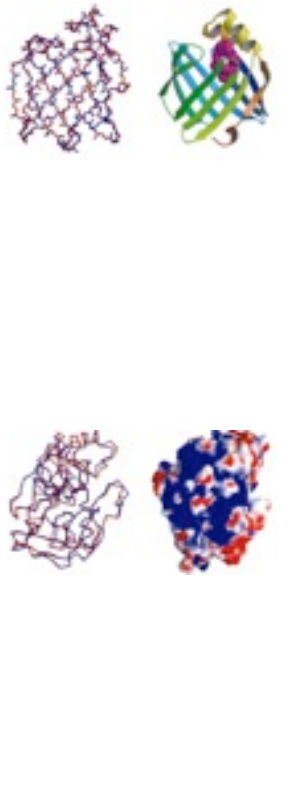
NMR

Comparative Modeling

Threading

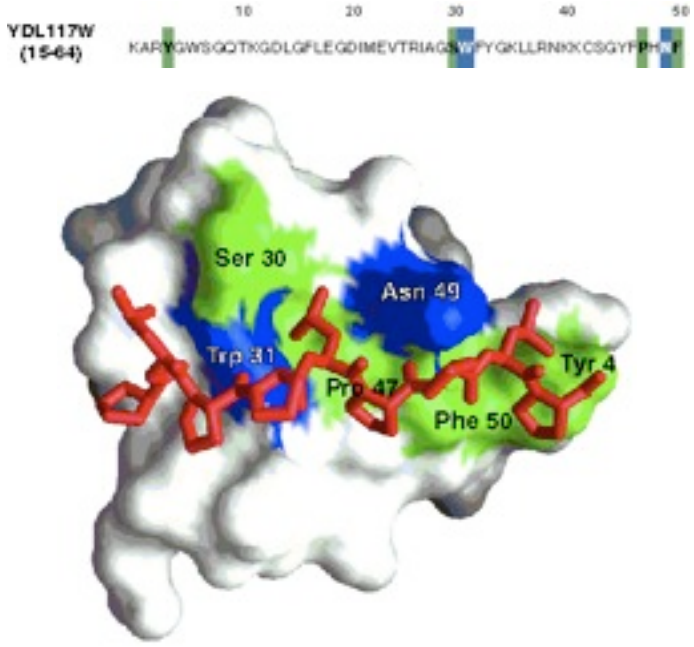
Ab-initio





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

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

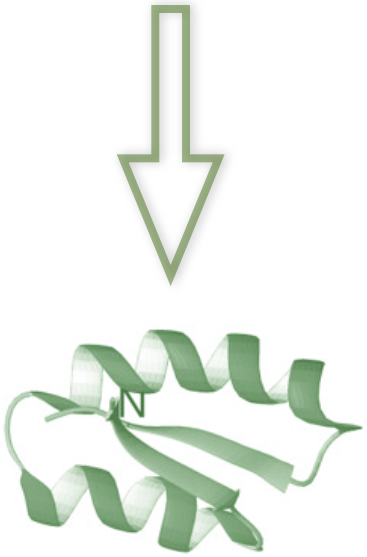


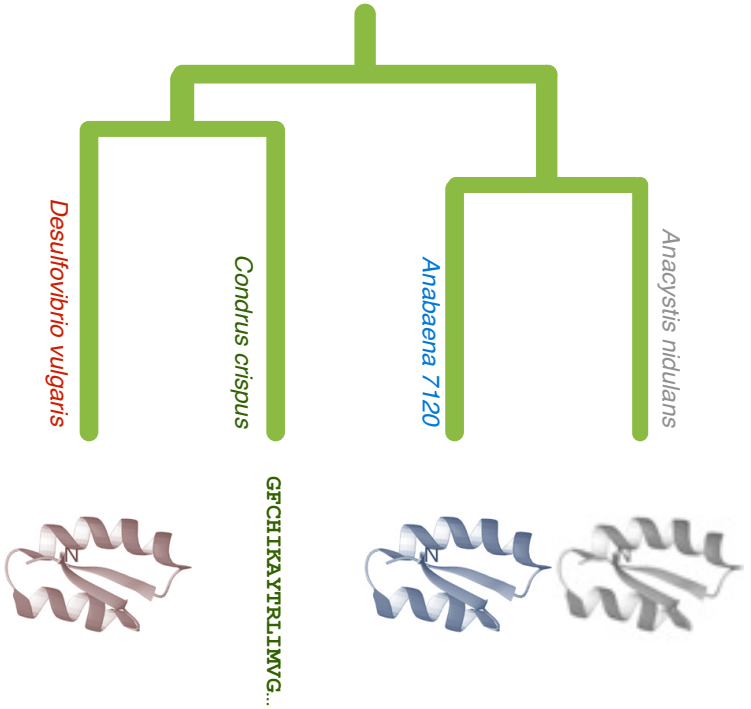
In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence**.

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

Principles of protein structure

GFCHIKAYTRLIMVG...





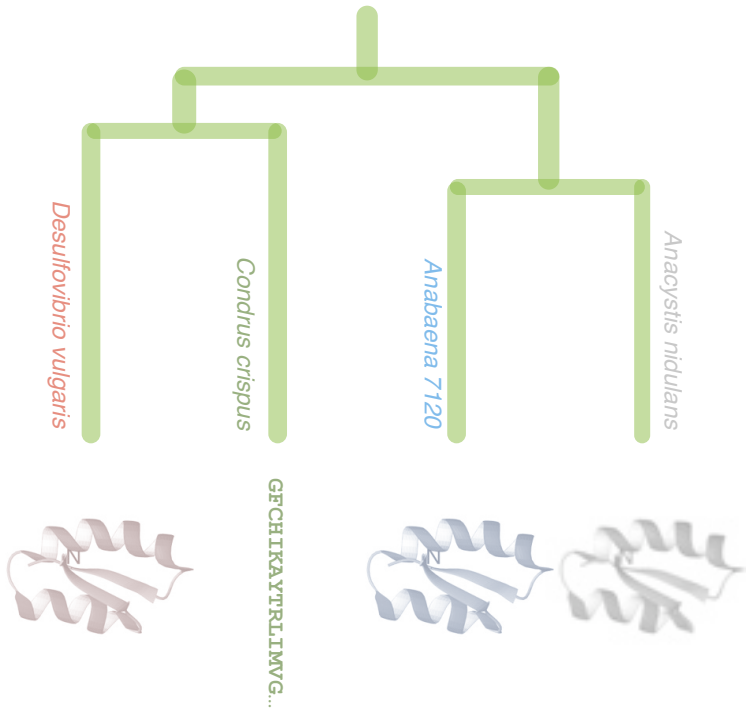
Folding (physics)

Ab initio prediction

Evolution (rules)

Threading
Comparative Modeling

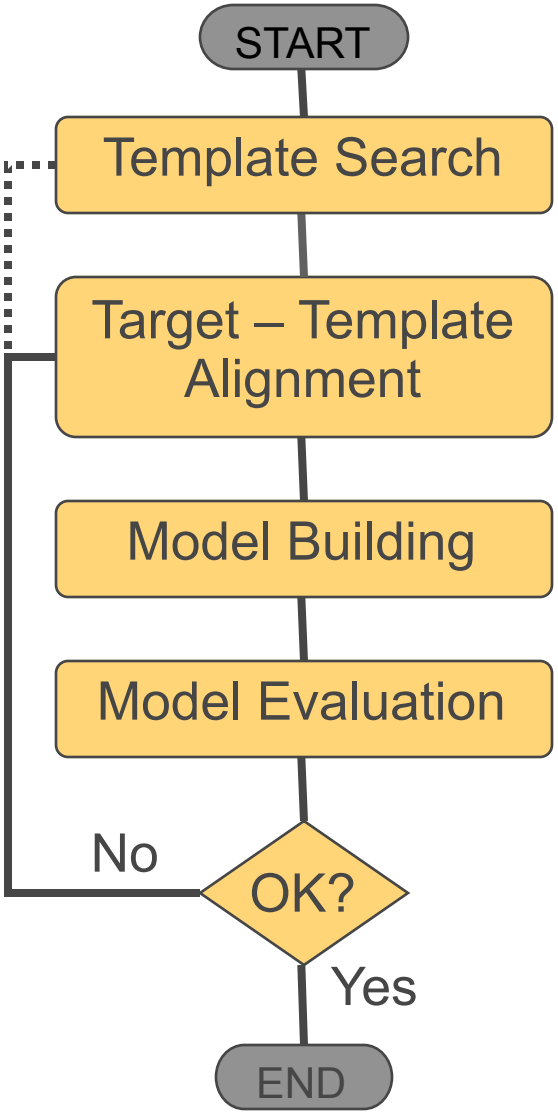
D. Baker & A. Sali. Science 294, 93, 2001.

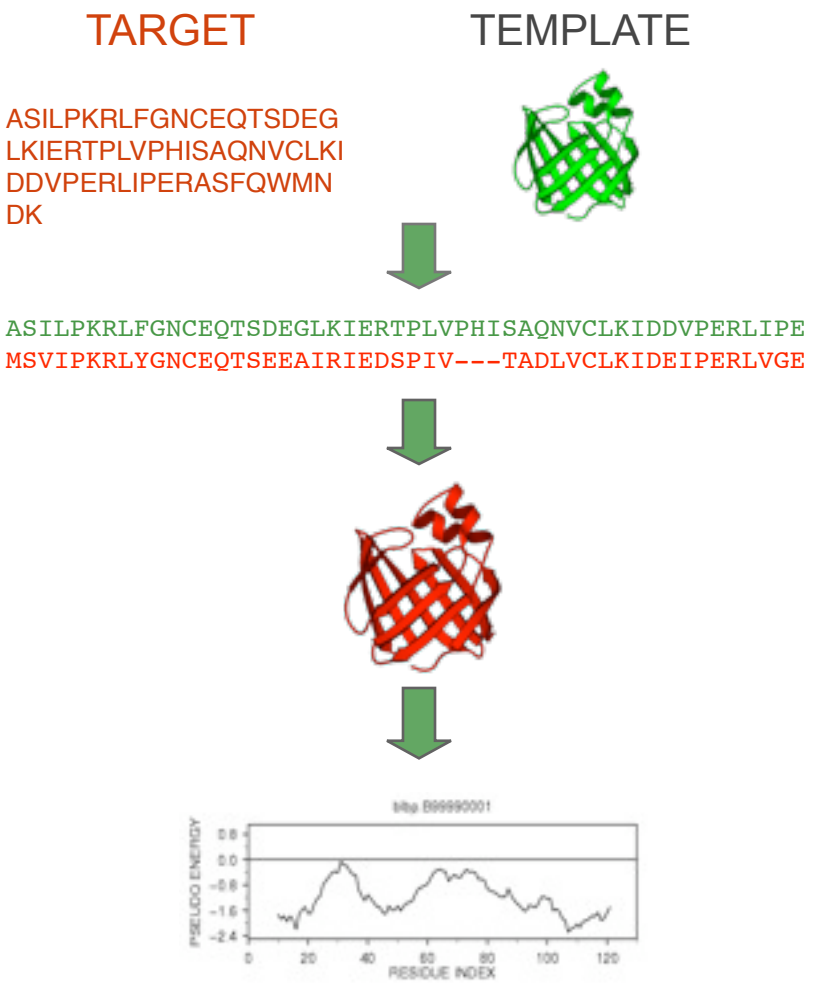


MODELLER

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

Steps in Comparative Protein Structure Modeling





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

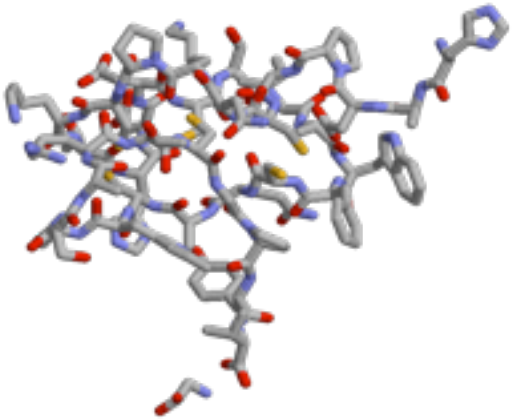
“Structural Space”

Structure-Structure alignments

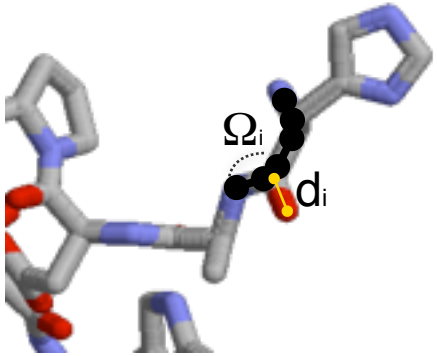
As any other bioinformatics problem…

- Representation
- Scoring
- Optimizer

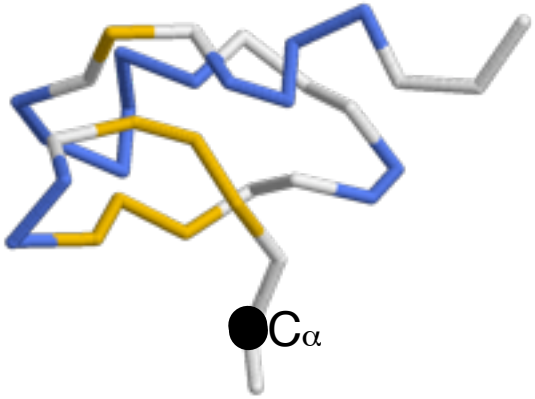
Structures



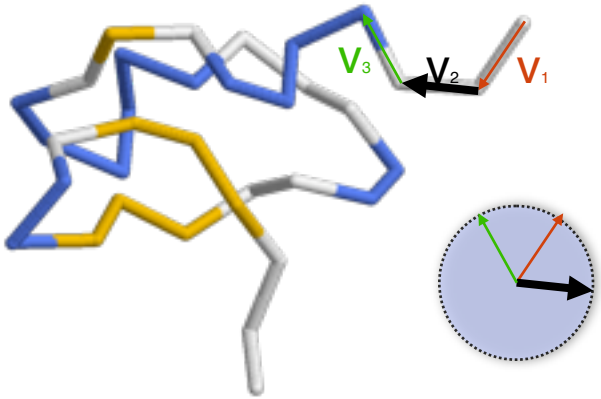
All atoms and coordinates



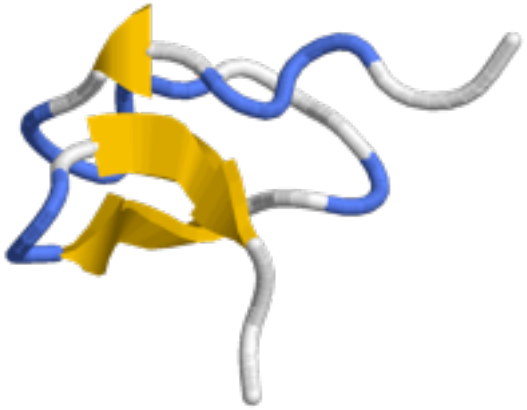
Dihedral space or distance space



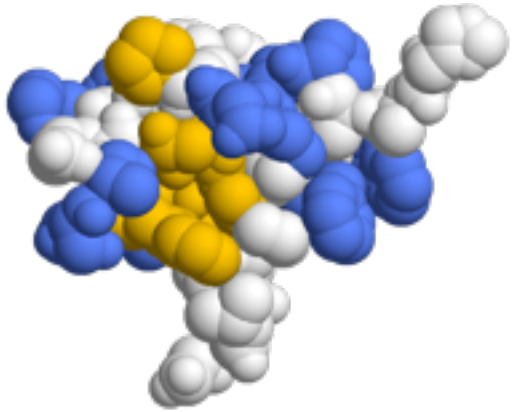
Reduced atom representation



Vector representation

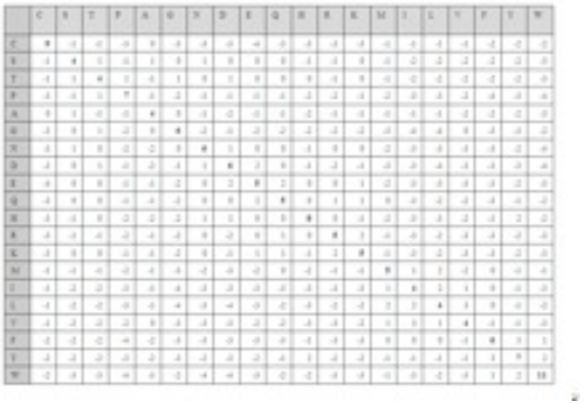


Secondary Structure



Accessible surface (and others)

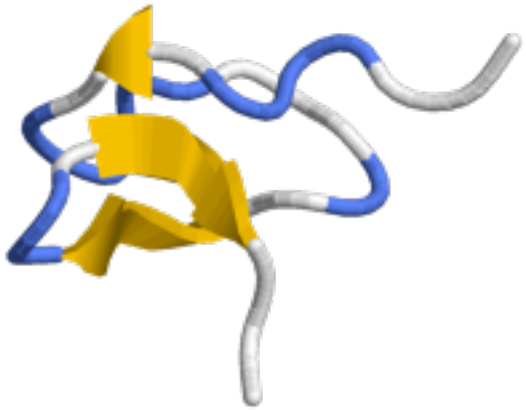
Raw scores



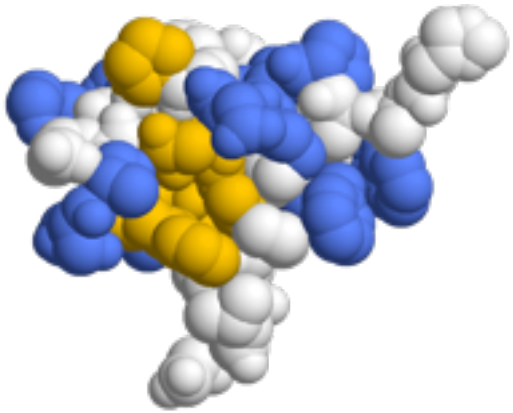
Aminoacid substitutions

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

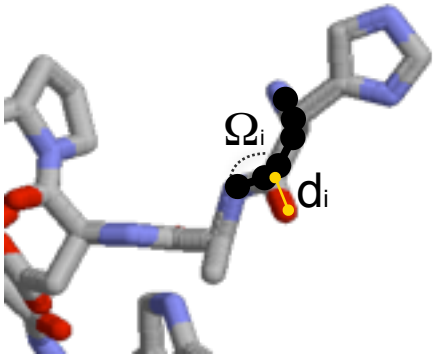
Root Mean Square Deviation



Secondary Structure (H,B,C)



Accessible surface (B,A [%])

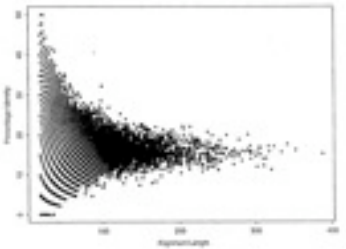


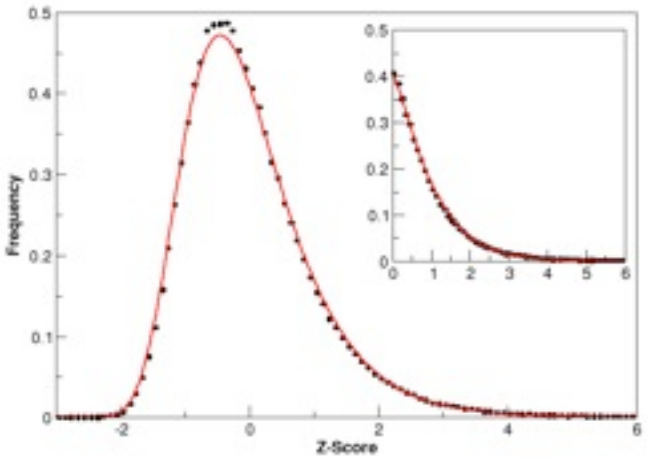
Angles or distances

Significance of an alignment (score)

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

Empirical





Sometimes approximated by Z-score (normal distribution).

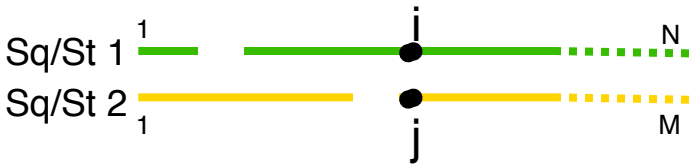
Analytic

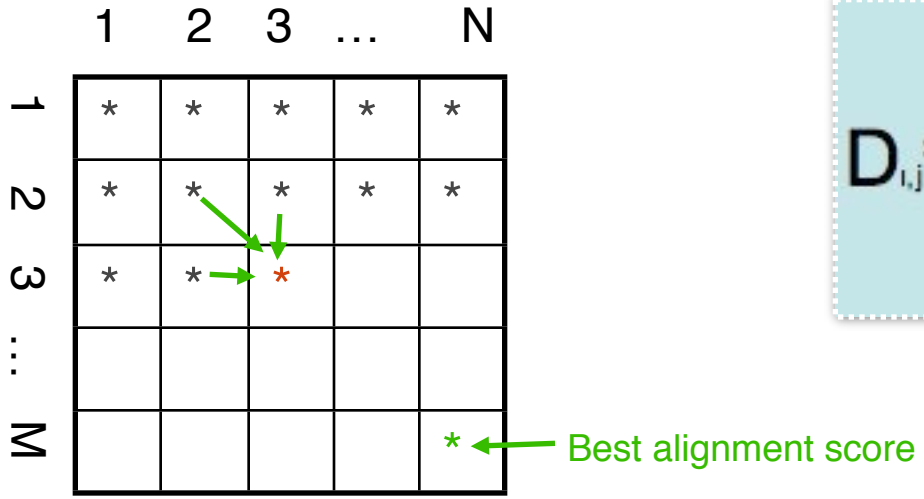
$$P(s) = e^{-\lambda (s-\mu)}$$

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

Karlin and Altschul, 1990 PNAS 87, pp2264

Global dynamic programming alignment



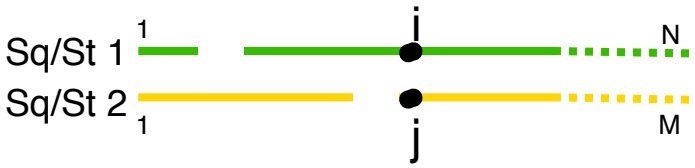


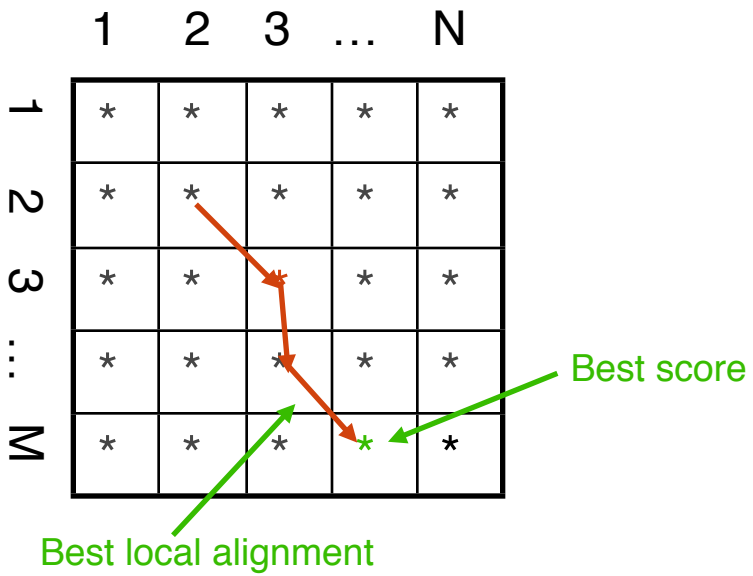
$$D_{i,j}=\min\begin{cases} D_{i,j-1}+Score_{(\Delta,rj)} \\ D_{i-1,j-1}+Score_{(ri,rj)} \\ D_{i-1,j}+Score_{(ri,\Delta)} \end{cases}$$

Backtracking to get the best alignment

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

Local dynamic programming alignment



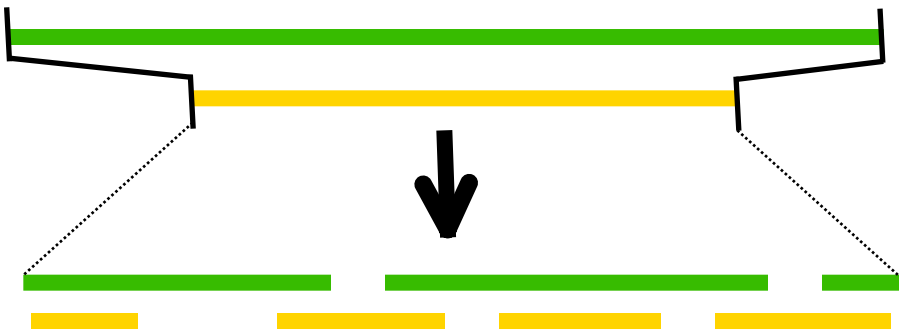


$$D_{i,j} = \min \begin{cases} D_{i,j-1} + Score_{(\Delta,rj)} \\ D_{i-1,j-1} + Score_{(ri,rj)} \\ D_{i-1,j} + Score_{(ri,\Delta)} \\ 0 \end{cases}$$

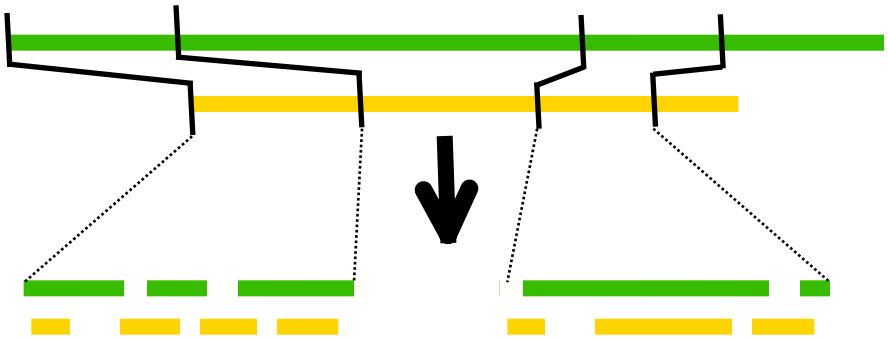
Backtracking to get the best alignment

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

Global .vs. local alignment



Global alignment



Local alignment

Multiple alignment

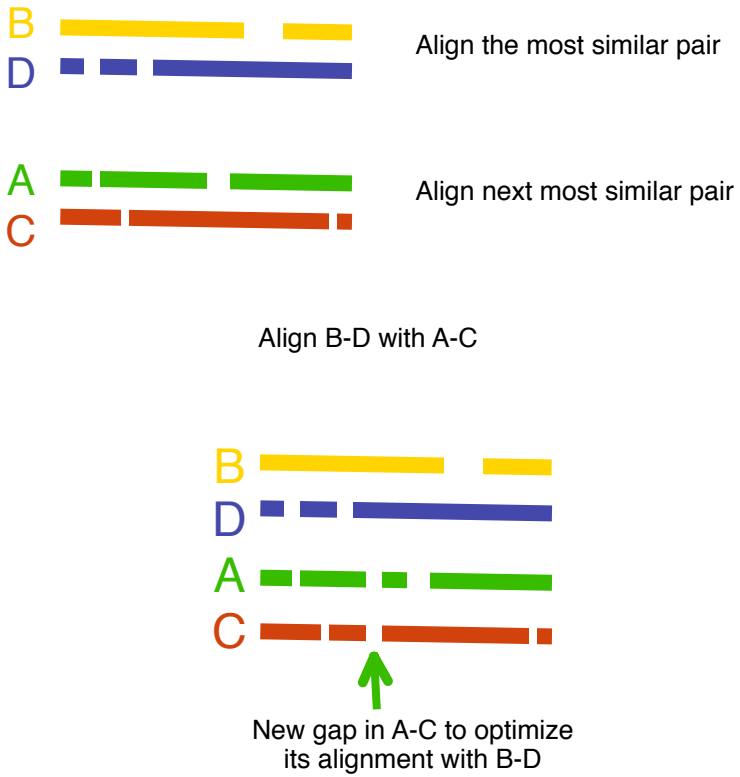
Pairwise alignments

Example – 4 sequences A, B, C, D.

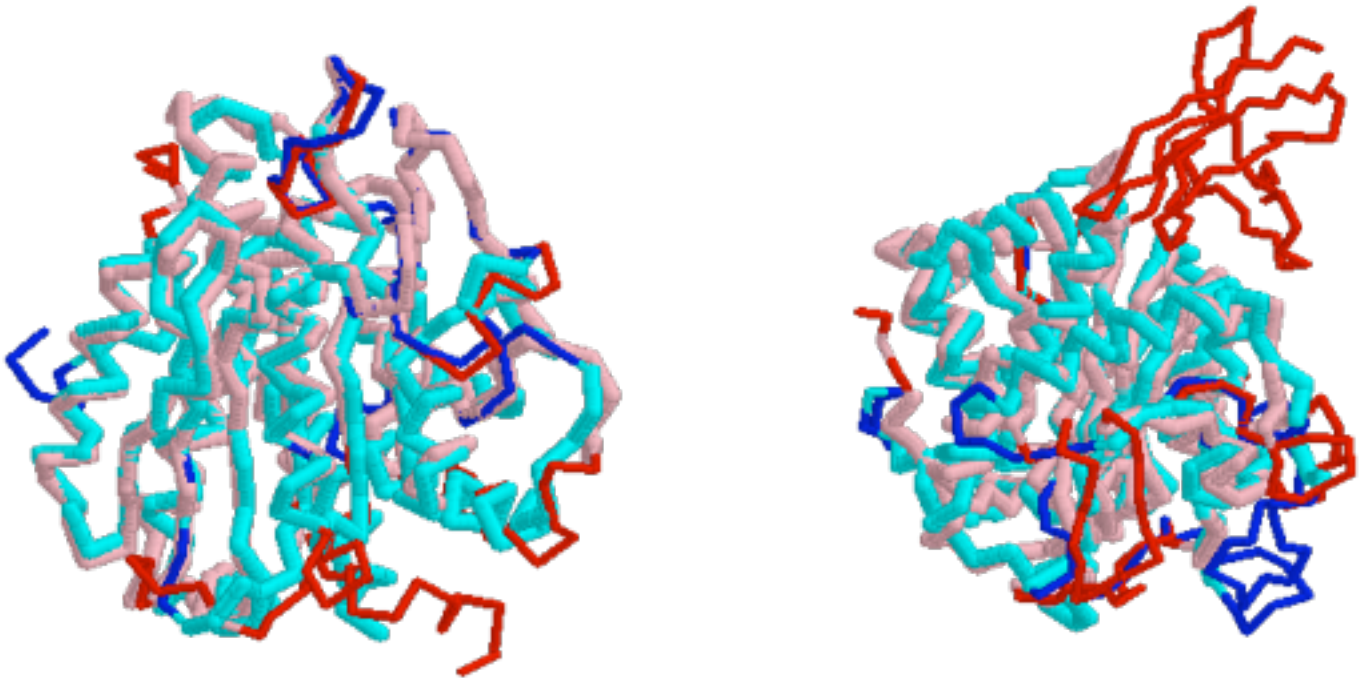
6 pairwise comparisons
then cluster analysis

Multiple alignments

Following the tree from step 1



Coverage .vs. Accuracy

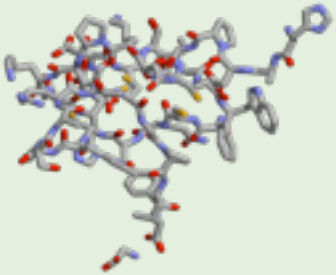


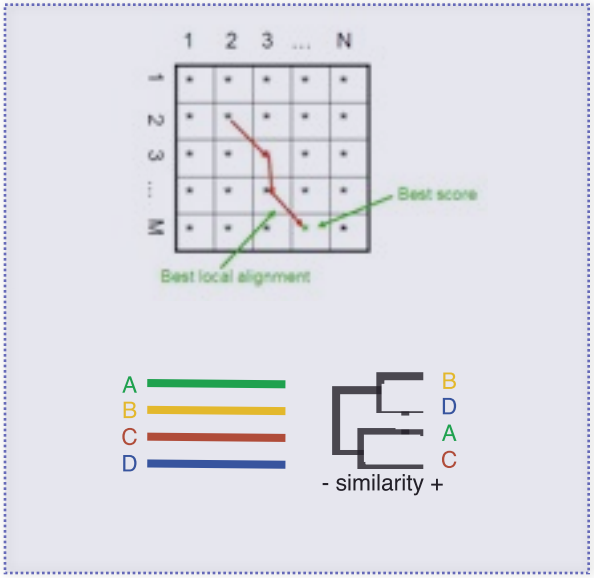
Same RMSD ~ 2.5Å

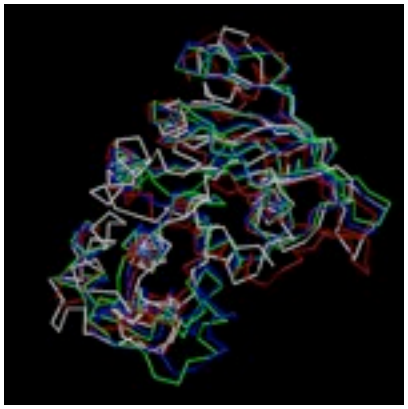
Coverage ~90% Cα

Coverage ~75% Cα

Structural alignment by properties conservation (SALIGN-MODELLER)

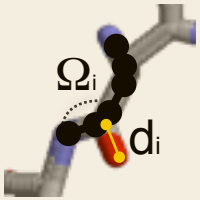


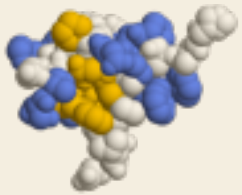




✓ Uses all available structural information
✓ Provides the optimal alignment

Computationally expensive





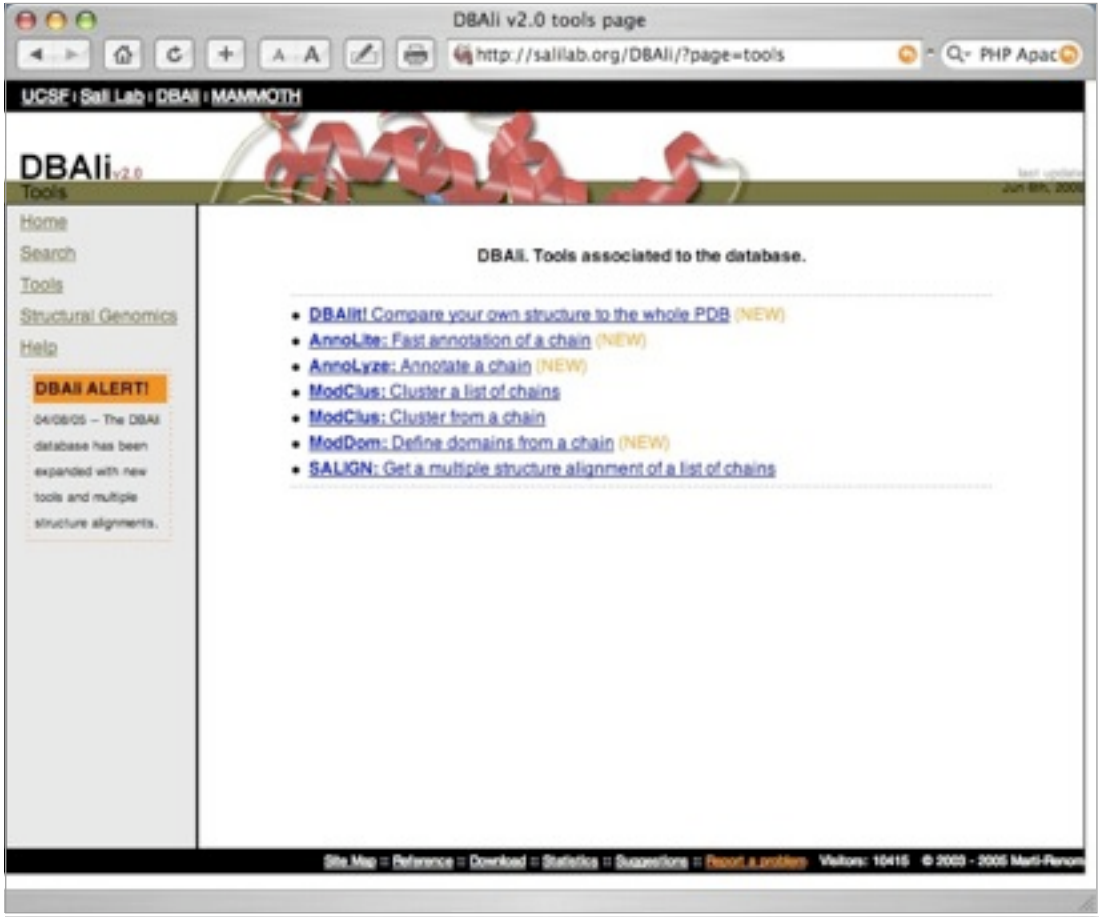
$$RMSD(x,y) = \sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i) - \mathbf{y}(i)\right\|^2\right)}$$

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

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

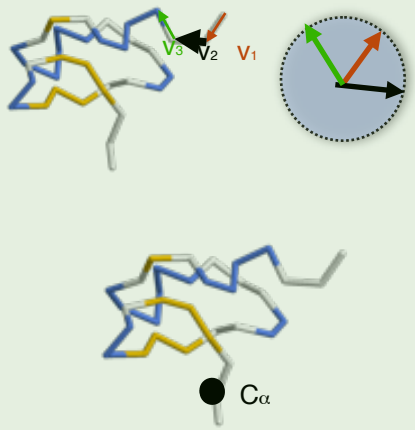
Structural alignment by properties conservation (SALIGN-MODELLER)

http://salilab.org/DBAli



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

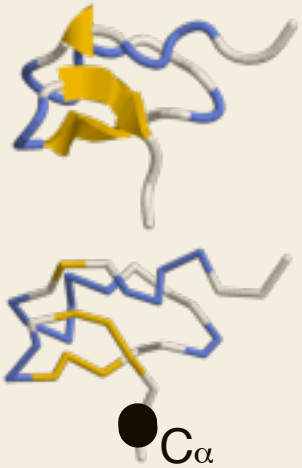
Vector Alignment Search Tool (VAST)

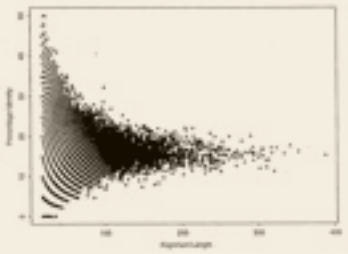


Graph theory search
of similar SSE
Refining by Monte Carlo
at all atom resolution

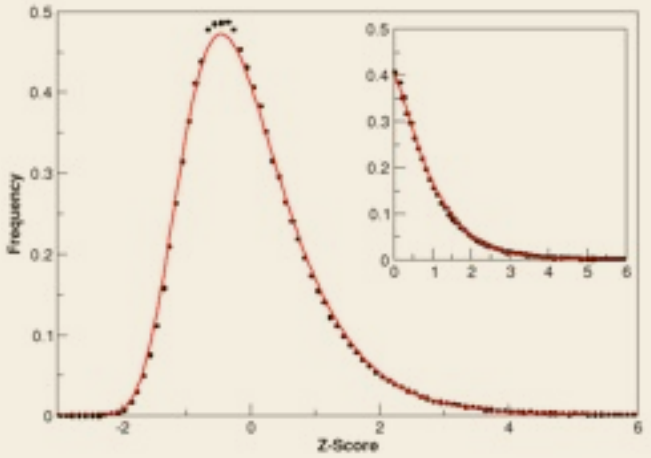
✓ Good scoring system with significance

Reduces the protein representation





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

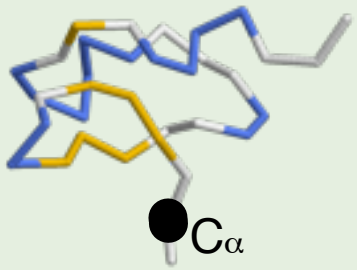


Gibrat JF et al. (1996) Curr Opin Struct Biol 3 pp377

Vector Alignment Search Tool (VAST)

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

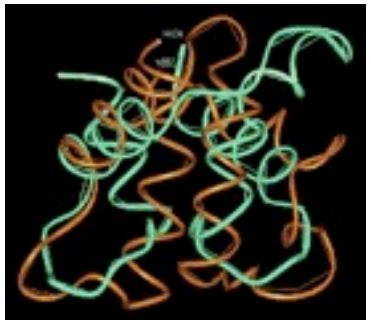
Incremental combinatorial extension (CE)



Exhaustive combination of fragments

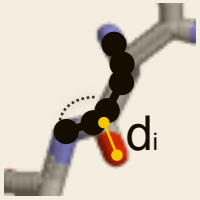
Longest combination of AFPs

Heuristic similar to PSI-BLAST



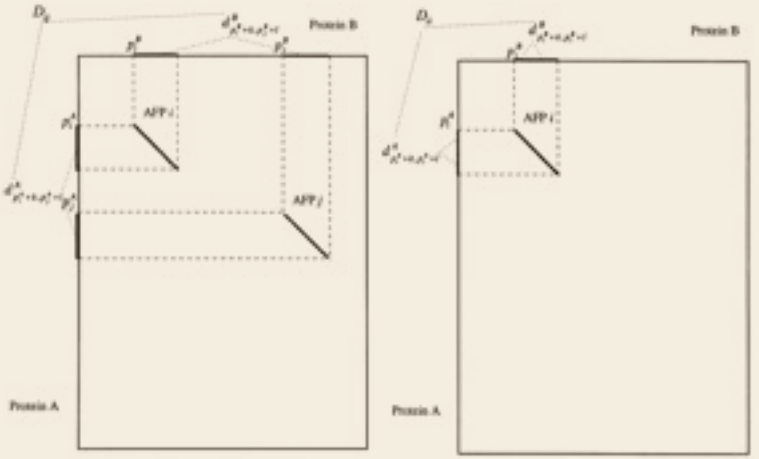
✓ FAST!
✓ Good quality of local alignments

Complicated scoring and heuristics



8 residues peptides

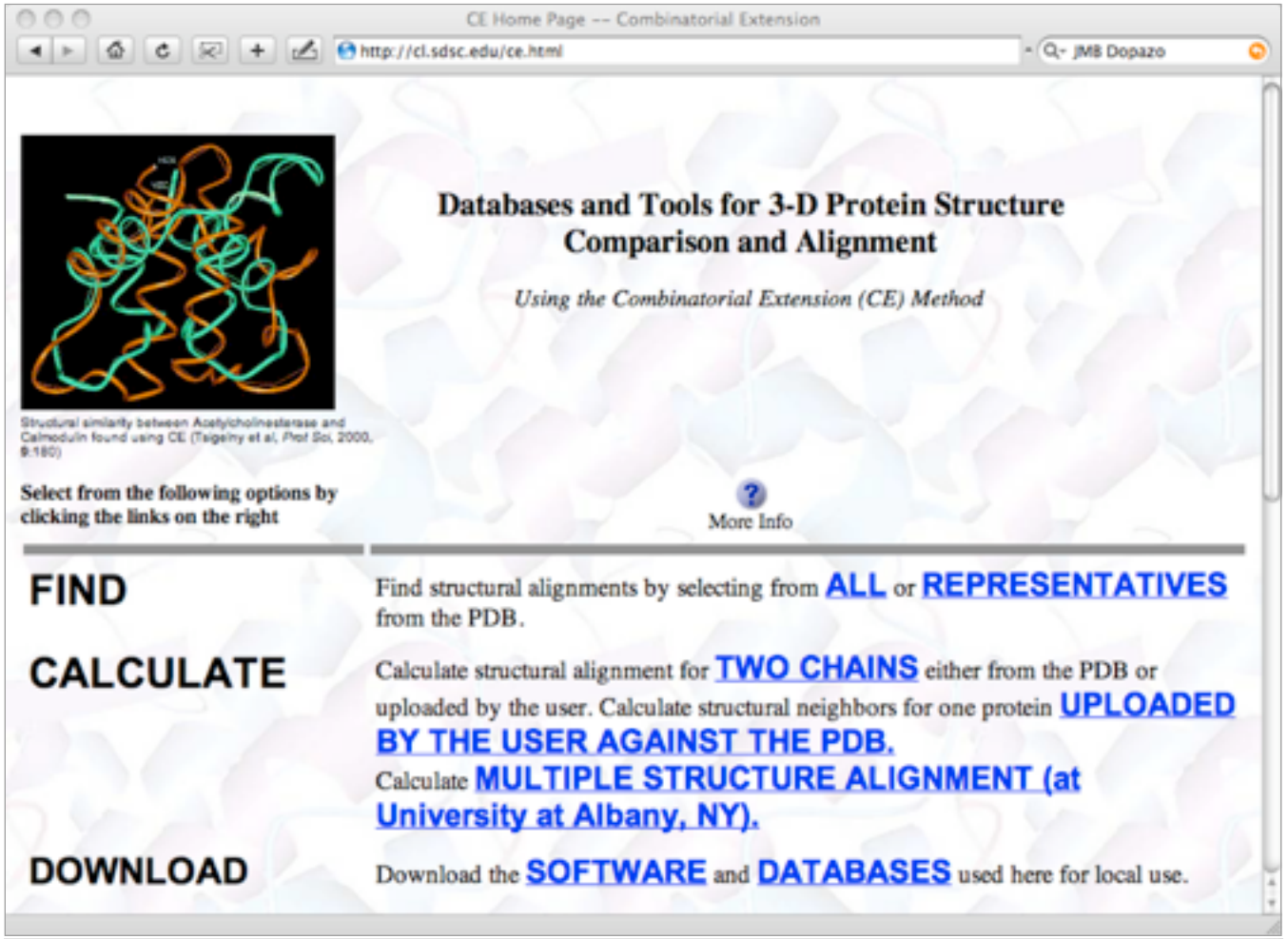
$$RMSD(x,y) = \sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i) - \mathbf{y}(i)\right\|^2\right)}$$



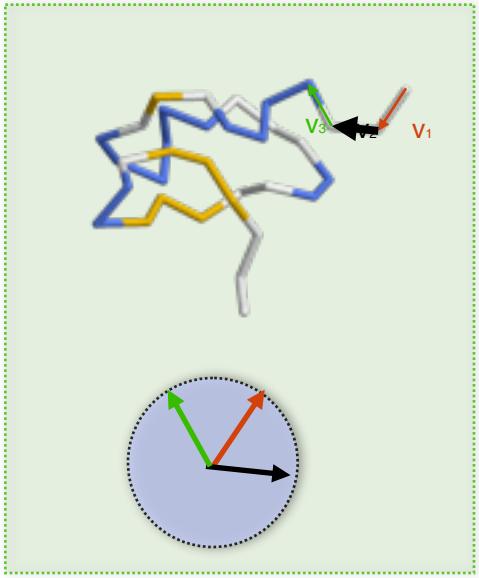
Shindyalov IN, amd Bourne PE. (1998) Protein Eng. 9 pp739

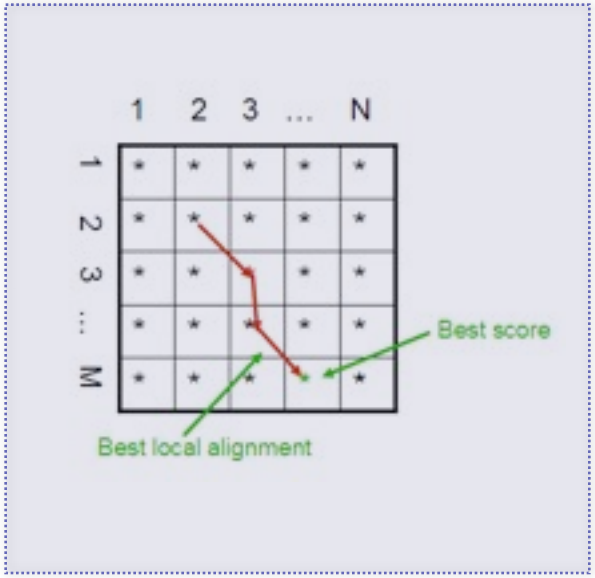
Incremental combinatorial extension (CE)

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



Matching molecular models obtained from theory (MAMMOTH)



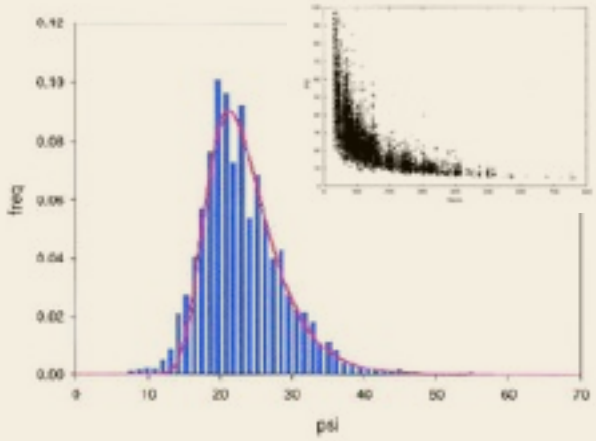


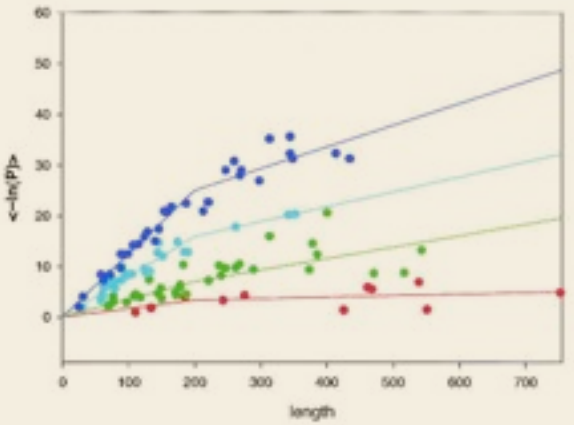
- ✓ VERY FAST!
- ✓ Good scoring system with significance

Reduces the protein representation

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

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



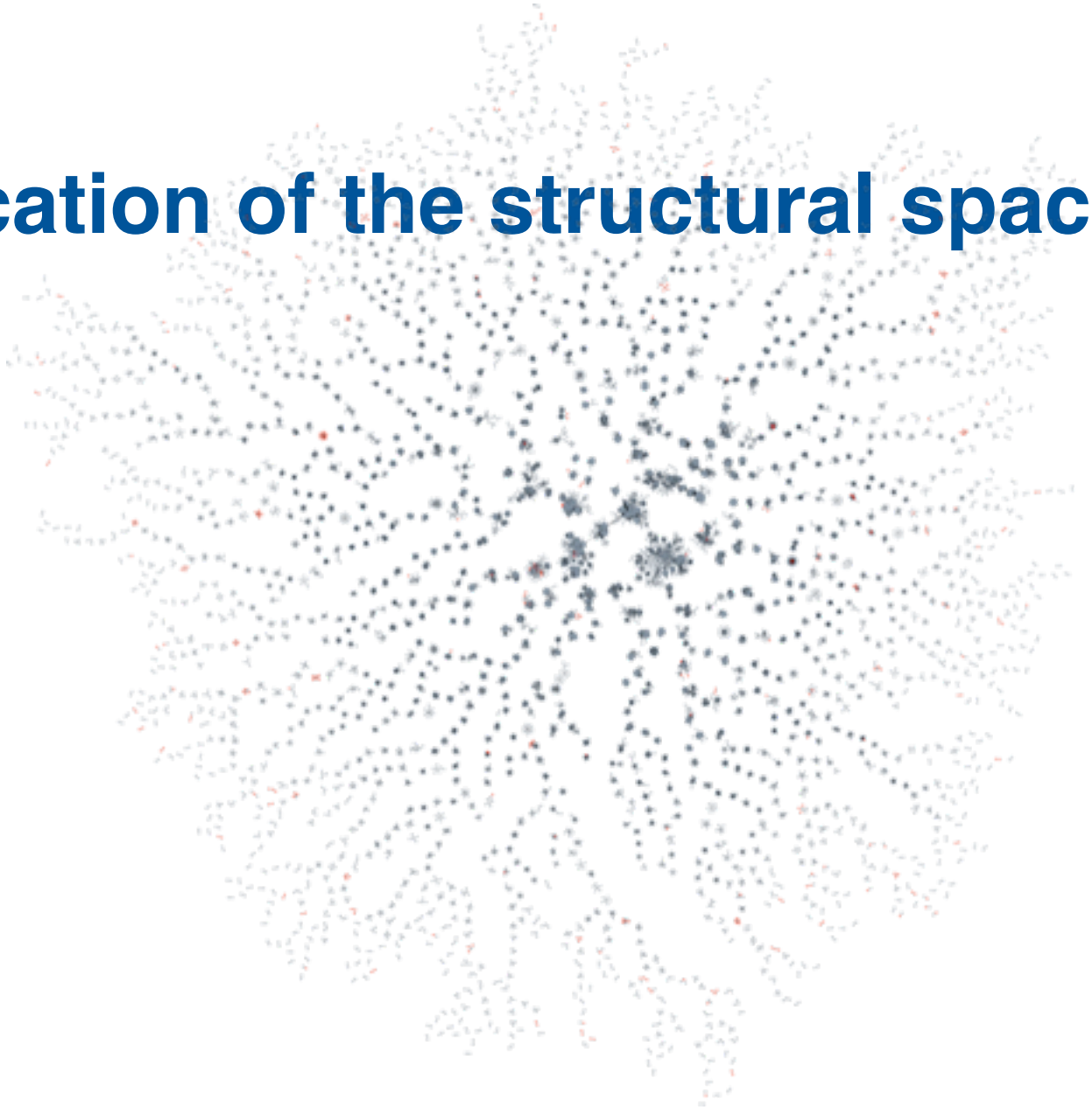


Ortiz AR, (2002) Protein Sci. 11 pp2606

Matching molecular models obtained from theory (MAMMOTH)

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

Classification of the structural space



SCOP$_{1.75}$ 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

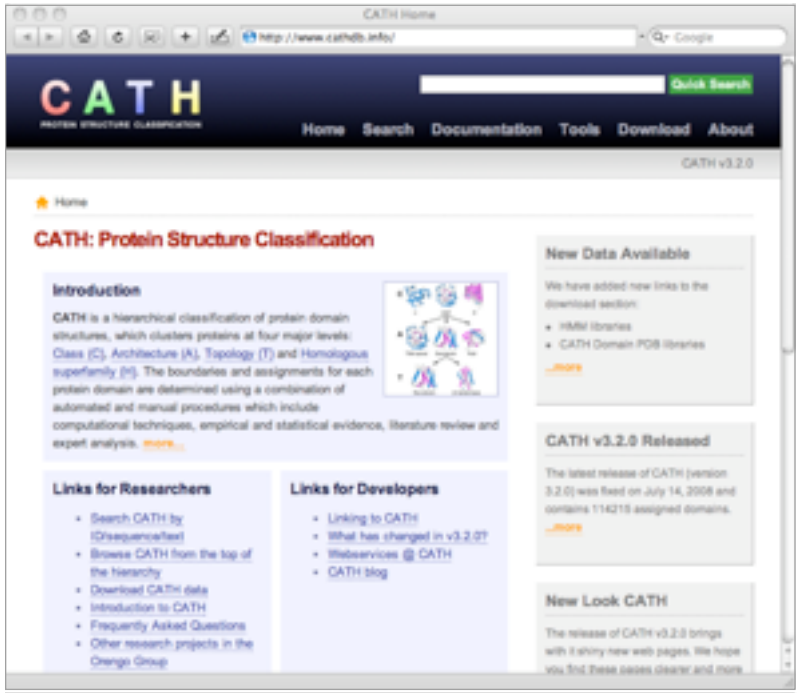
Manually classification
Not 100% up-to-date
Domain boundaries definition

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

CATH3.2 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

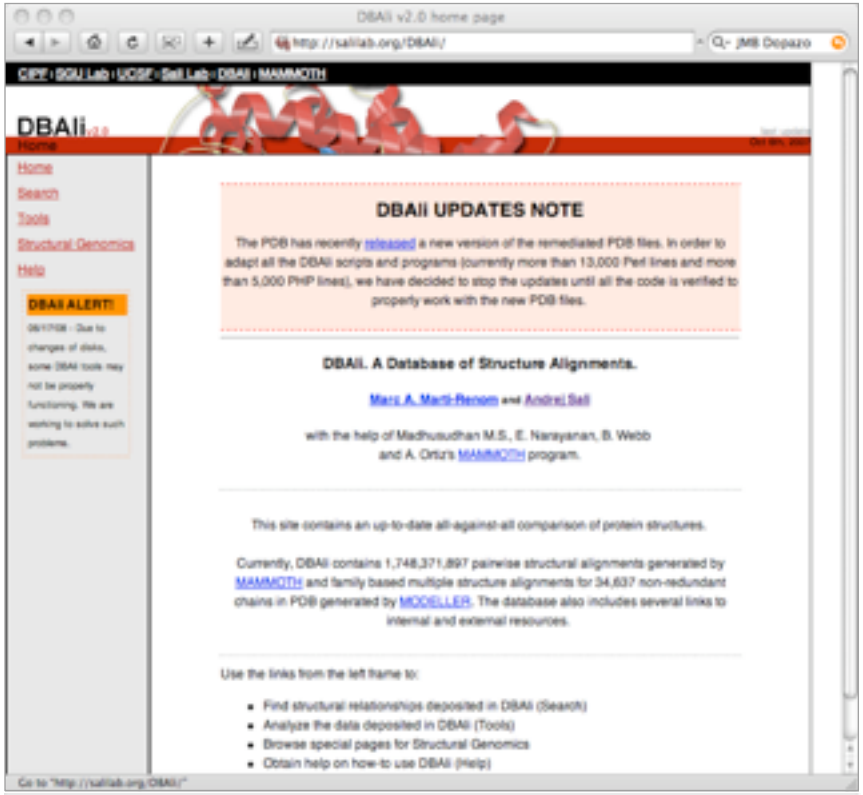
Semi-automatic classification
Domain boundaries definition

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$_{v2.0}$ 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

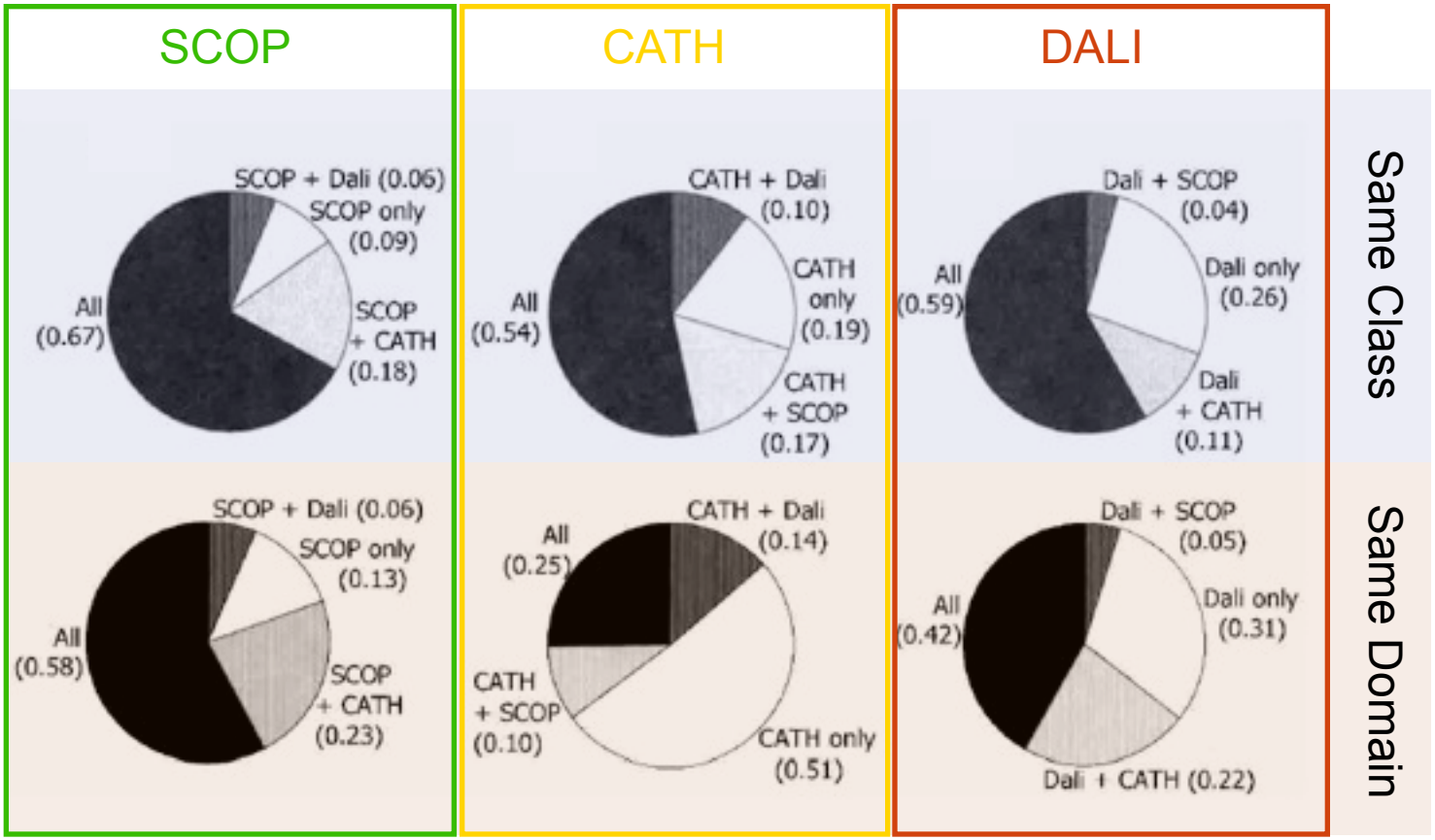
Pairwise structure alignments	
Last update:	October 6th, 2007
Number of chains:	96,804
Number of structure-structure comparisons:*	1,748,371,897
Multiple structure alignments	
Last update:	August 1st, 2007
Number of representative chains:	34,637
Number of families:	12,732

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

Classification of the structural space

Not an easy task!

Domain definition AND domain classification



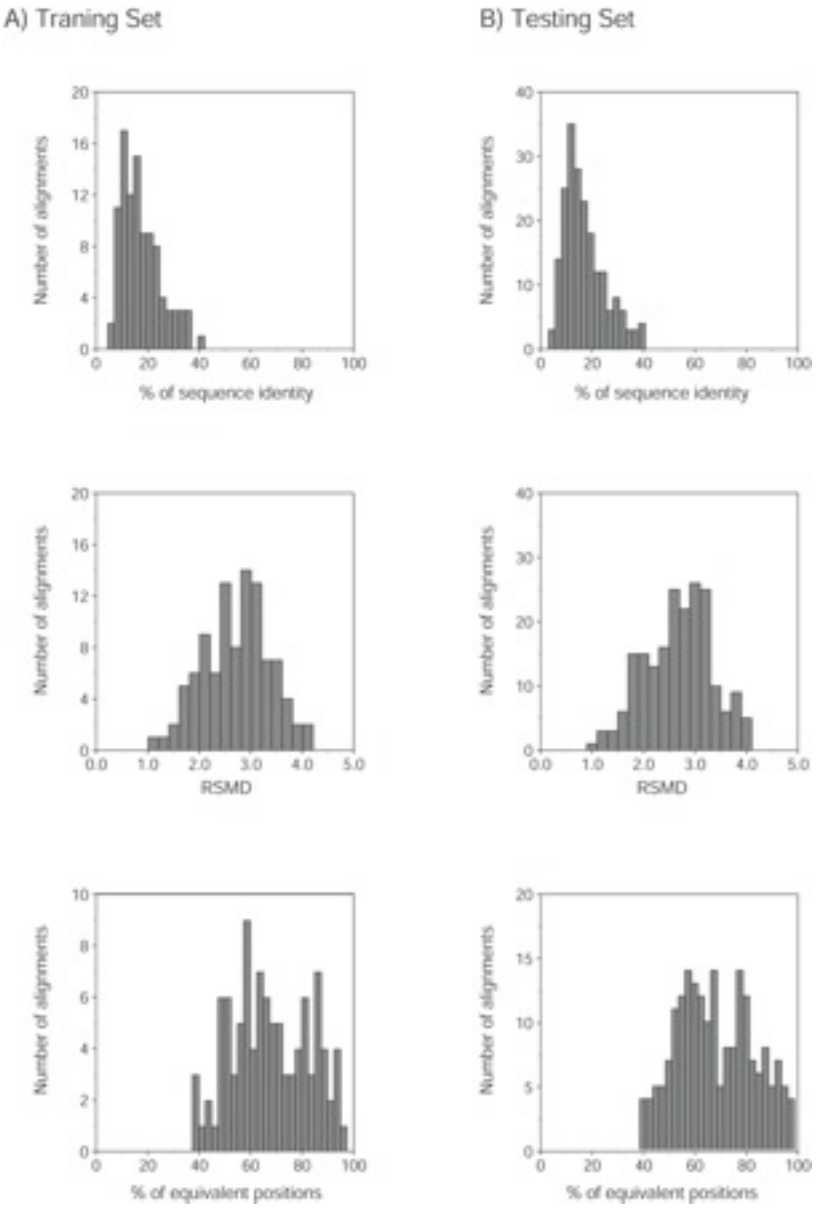
*Day, et al. (2003) Protein Sciences, **12** pp2150*

template search and template-target alignment

(pp_scan)

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

PP_SCAN or profile-profile alignments



Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local heuristic method

Seq.-Str.

SEA: Local structure prediction method

Prof.-Seq.

SAM: HMM method

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

LOBSTER: HHM + Phylogeny Method

Prof.-Prof.

CLUSTALW: DP multiple sequence method.

COMPASS: DP profile-profile method

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

PP_SCAN protocols

Profile generation

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

Profile comparison

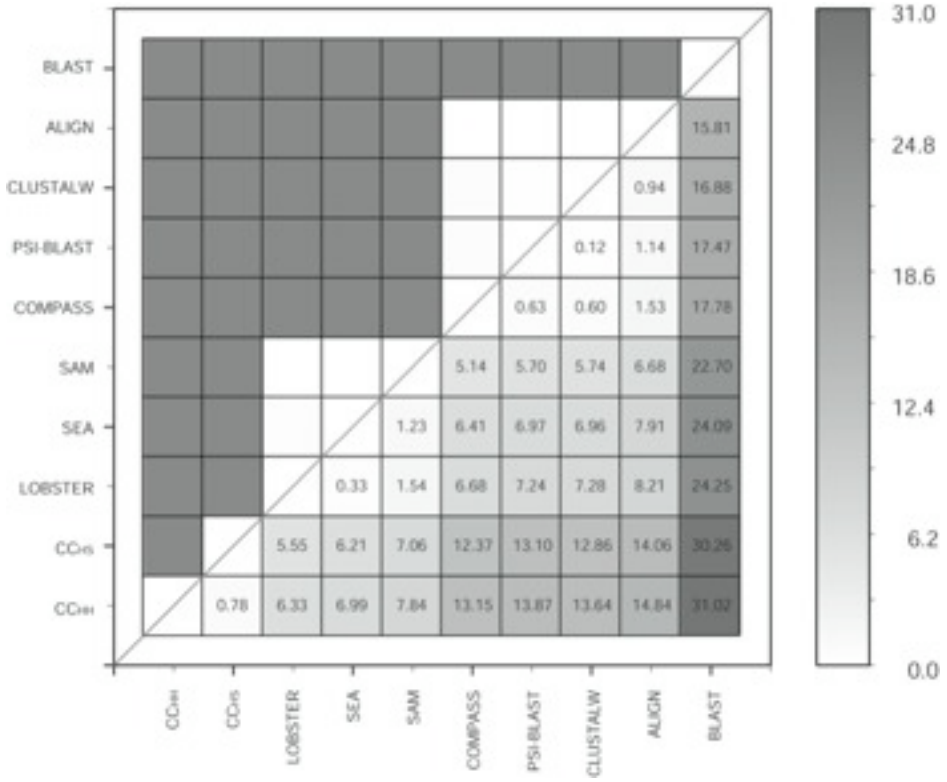
- Correlation coefficient (CC)
- Euclidean distance (ED)
- Dot product (DP)
- Jensen-Shannon distance (JS)
- Average value (Ave)

PP_SCAN protocols accuracy

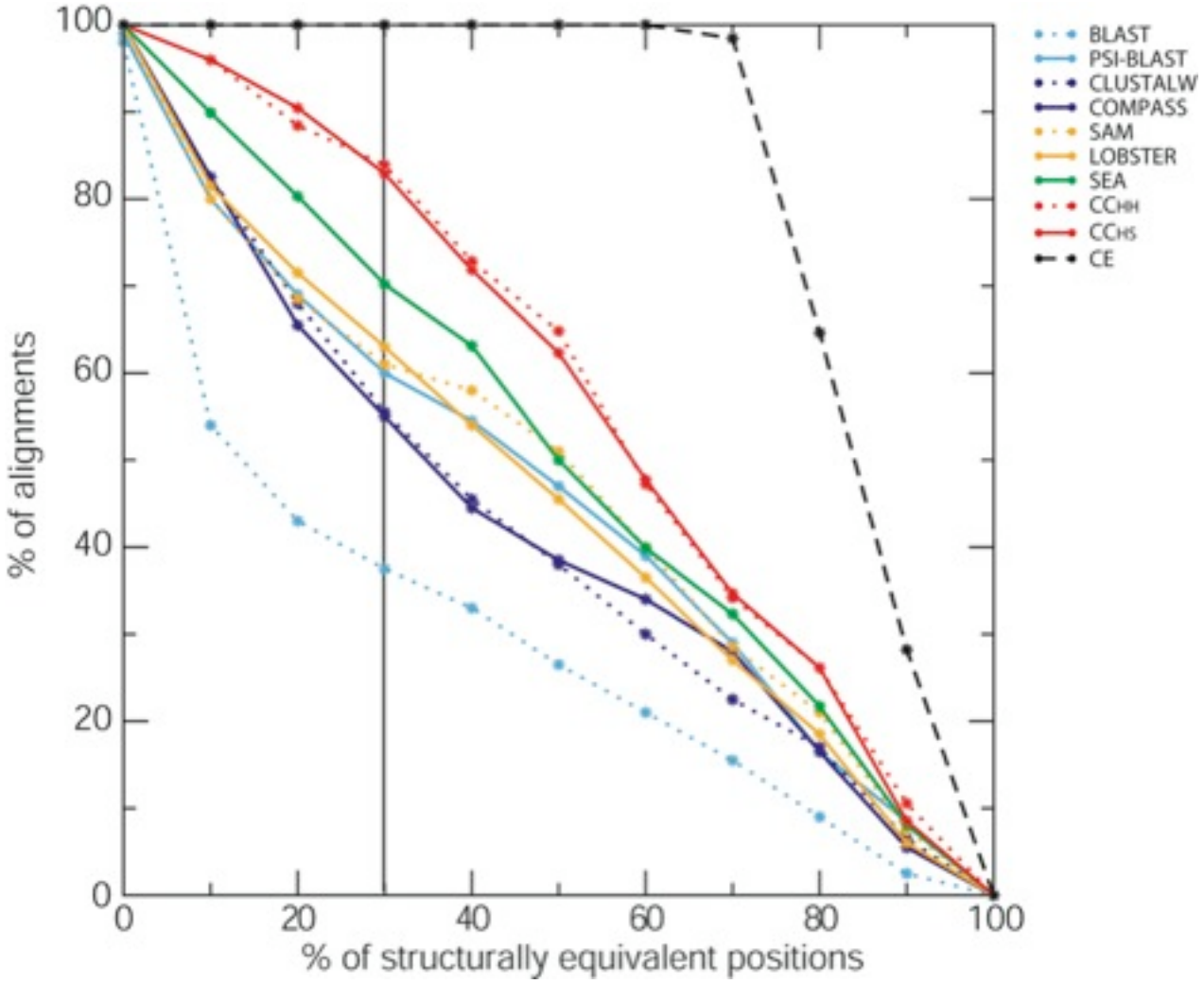
SALIGN protocol	CE overlap [%]	Shift score
CC$_{PBP}$	55 ± 23	0.61 ± 0.24
CC$_{HH}$	**56 ± 23**	**0.61 ± 0.24**
CC$_{HS}$	**56 ± 24**	**0.62 ± 0.23**
CC$_{MAT}$	51 ± 25	0.55 ± 0.27
ED$_{PBP}$	54 ± 24	0.60 ± 0.25
ED$_{HH}$	54 ± 24	0.59 ± 0.26
ED$_{HS}$	55 ± 24	0.59 ± 0.26
DP$_{PBP}$	55 ± 23	0.61 ± 0.24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

PP_SCAN accuracy

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



PP_SCAN success



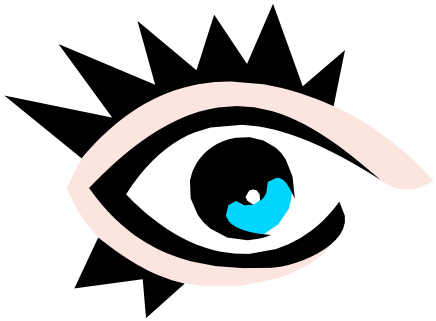
Alignment accuracy (CE overlap)

200 pairwise DBAli alignments

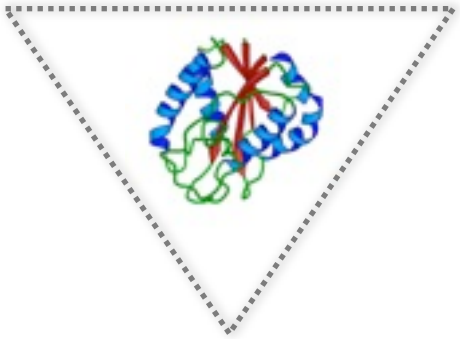
PSI-BLAST (sequence-profile alignment)	43%
SEA (local structure alignment)	49%
PP_SCAN (profile-profile alignment)	56%

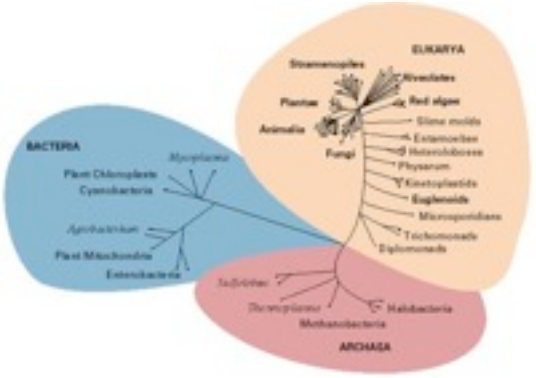
model building and model assessment

Information about a protein can come from three distinct sources



Experimental observations





Statistical rules

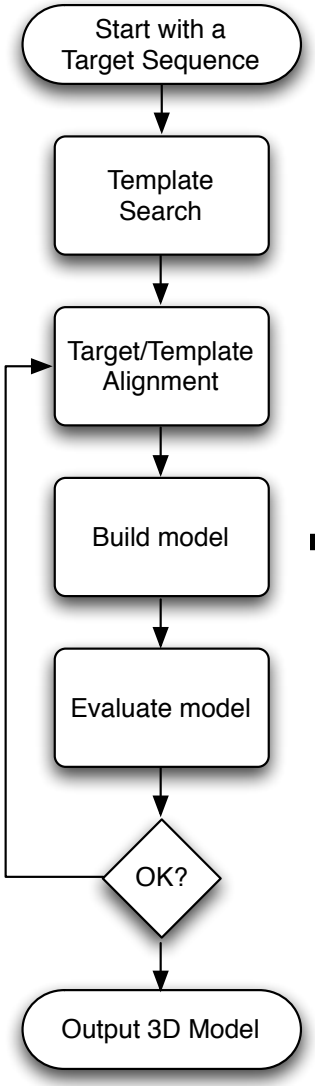
Laws of physics

Classes of methods for comparative protein structure modeling

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

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

Comparative modeling by satisfaction of spatial restraints
MODELLER

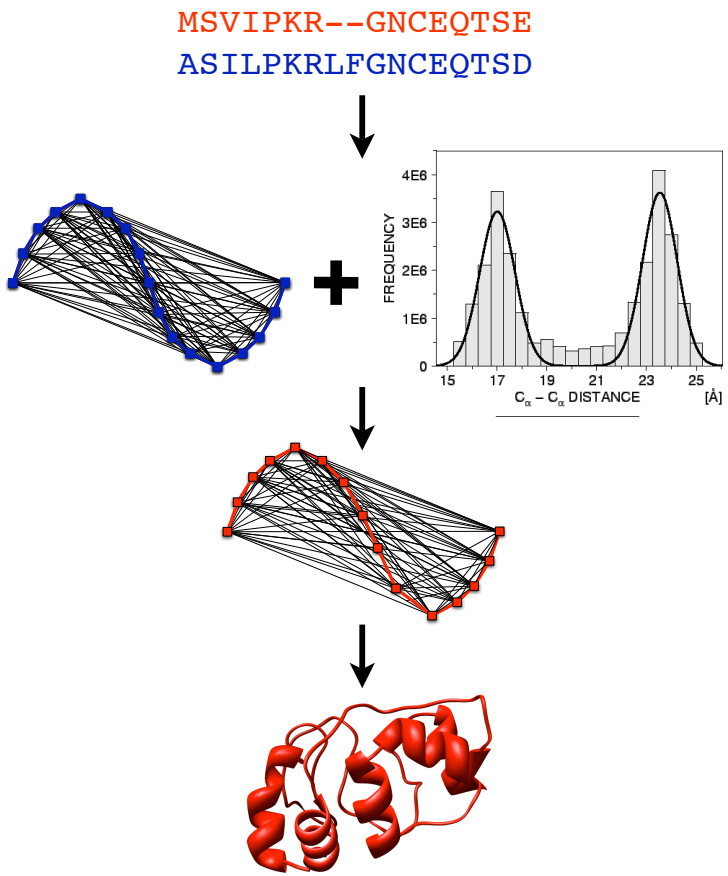


Given an alignment...

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



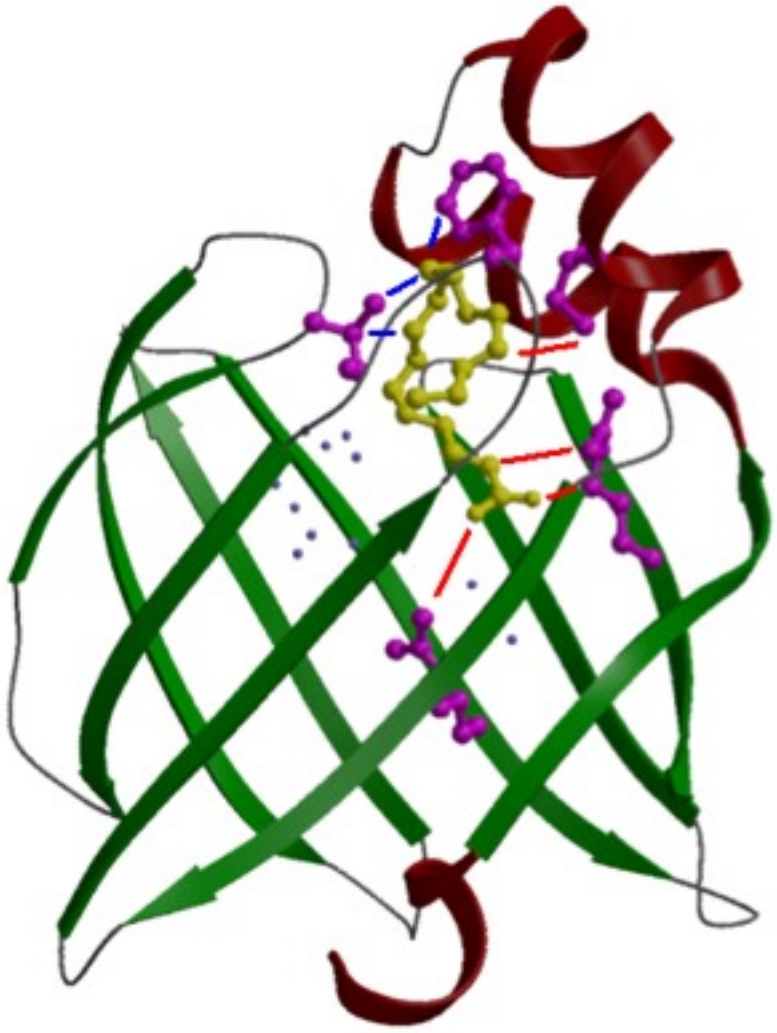
A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.

Multiple Templates

Local similarity extracted from closest template

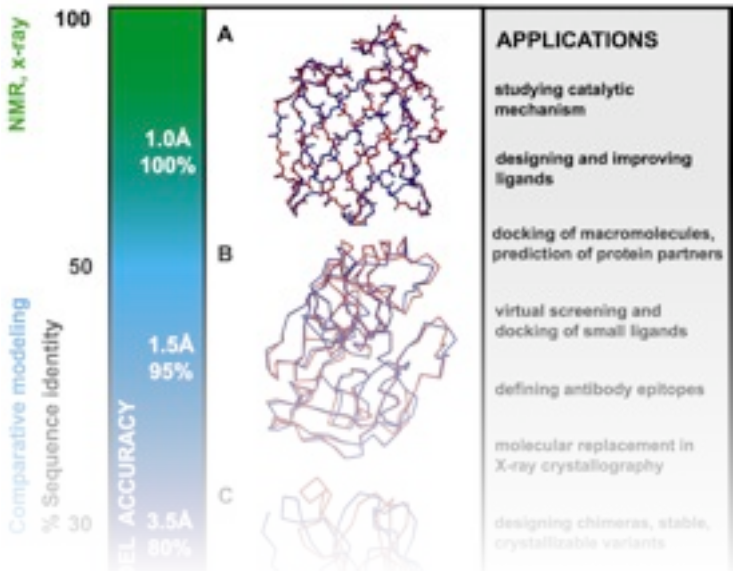
Templates	KSINPIHGDNCEQTSDEGLKIERTPL--------QWLKSSICDMRGLIPE
	ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
Target	MSVIPKRLYGNCEQTSEEAIRIEDSPIVRWISAQLVCLKIDEIPERLVGE

Modeling ligands and using external restraints



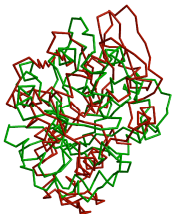
Homology derived restraint

External Restraint

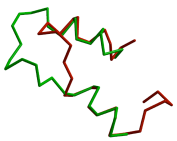


Accuracy and applicability of comparative models

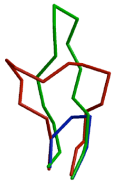
Comparative modeling by satisfaction of spatial restraints Types of errors and their impact



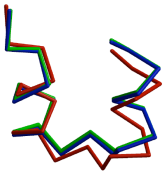
Wrong fold



Miss alignments

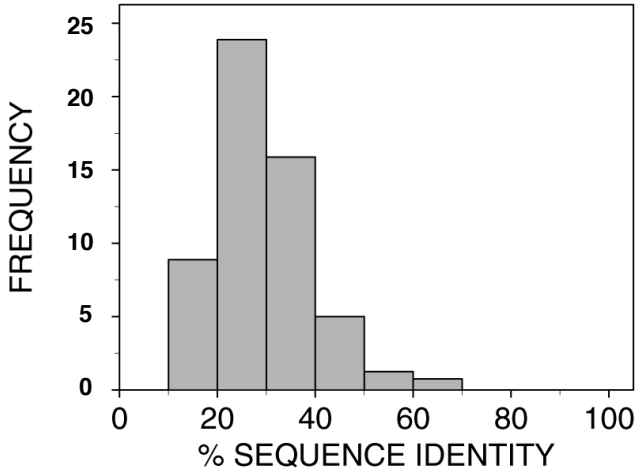


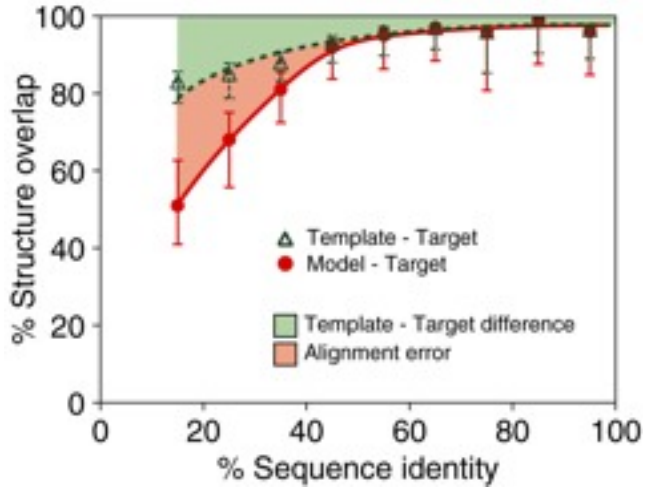
Loop regions



Rigid body distortions

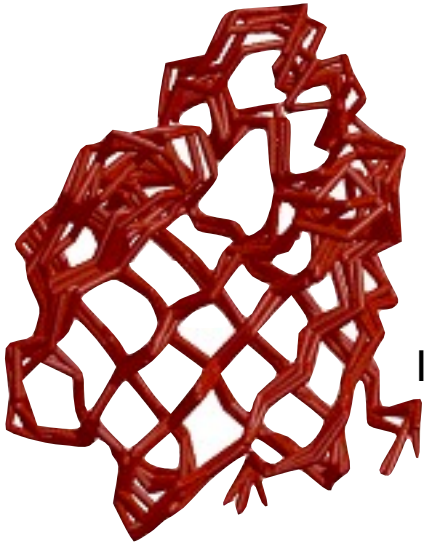
Side-chain packing





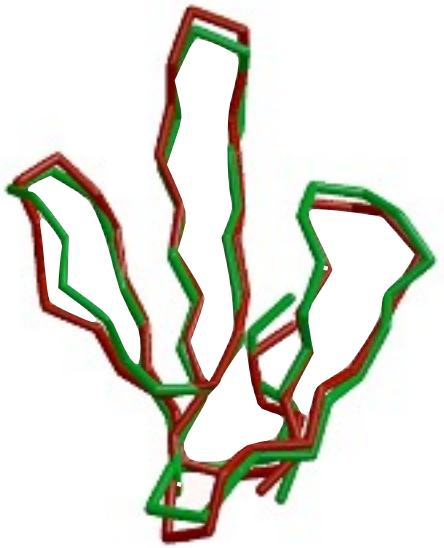
Marti-Renom etal. Ann Rev Biophys Biomol Struct (2000) 29, 291

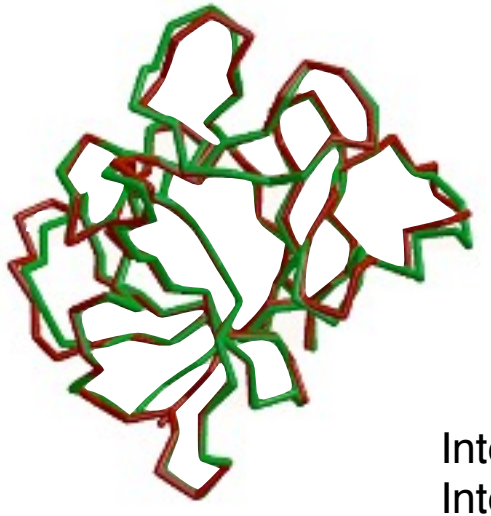
“Biological” significance of modeling errors



NMR
Ileal lipid-binding protein
1eal

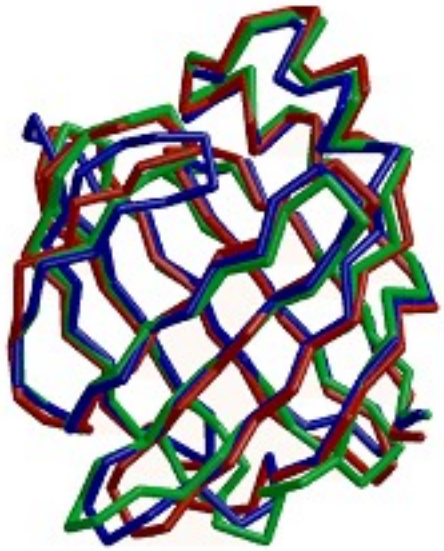
NMR – X-RAY
Erabutoxin 3ebx
Erabutoxin 1era





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

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



Model Accuracy

HIGH ACCURACY

NM23 Seq id 77%

Cα equiv 147/148
RMSD 0.41Å

Sidechains
Core backbone
Loops

MEDIUM ACCURACY

CRABP Seq id 41%

Cα equiv 122/137
RMSD 1.34Å

Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

EDN Seq id 33%

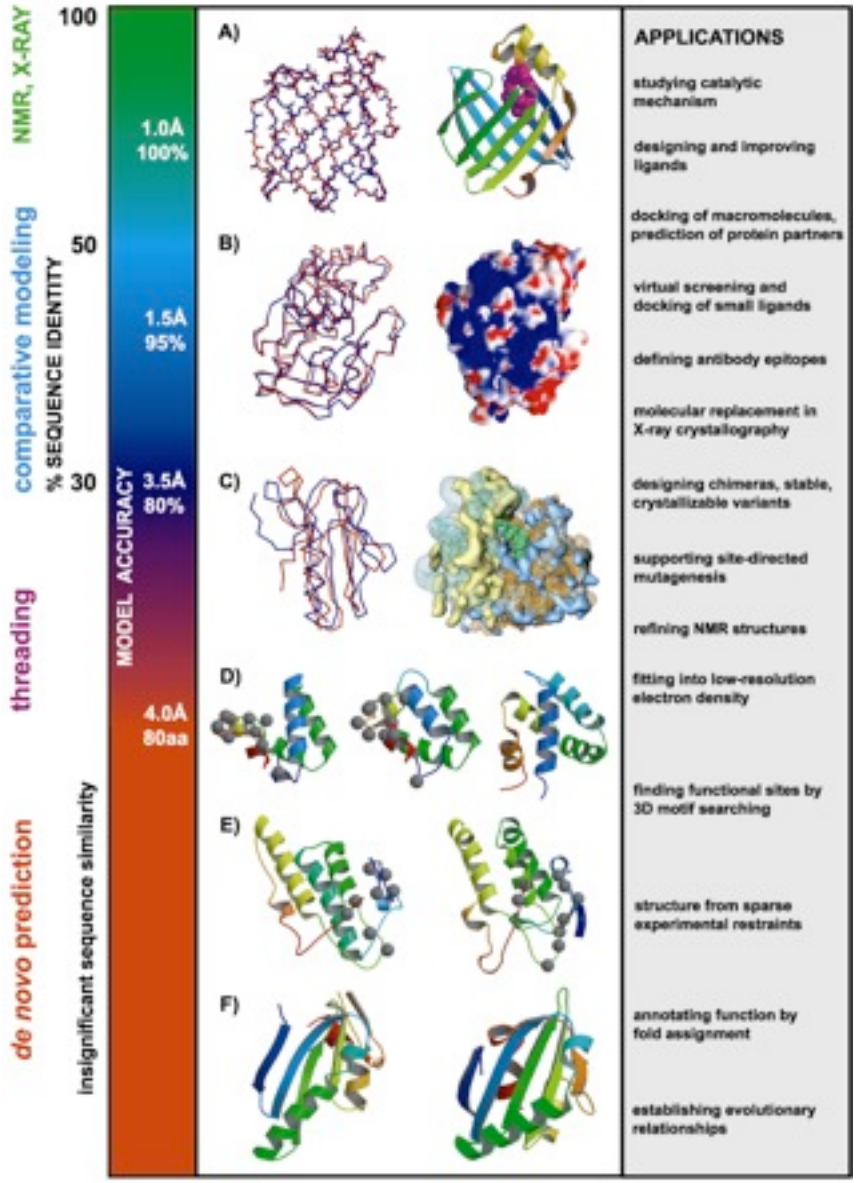
Cα equiv 90/134
RMSD 1.17Å

Sidechains
Core backbone
Loops
Alignment
Fold assignment

X-RAY / MODEL

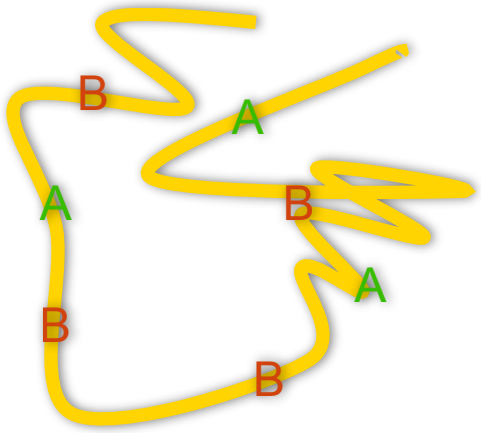
Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

Utility of protein structure models, despite errors



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

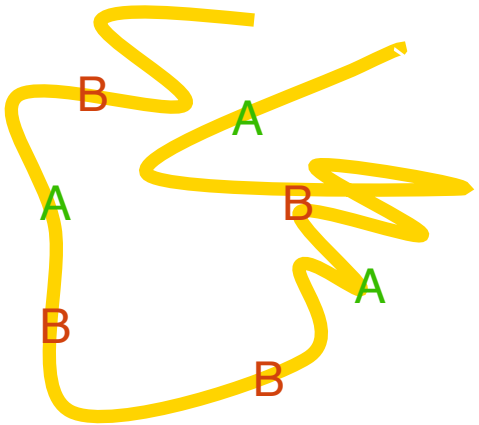
Model Assessment (PMF)



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 (ΔE) and the ratio of their occupancies (N_1:N_2) are related [9]:

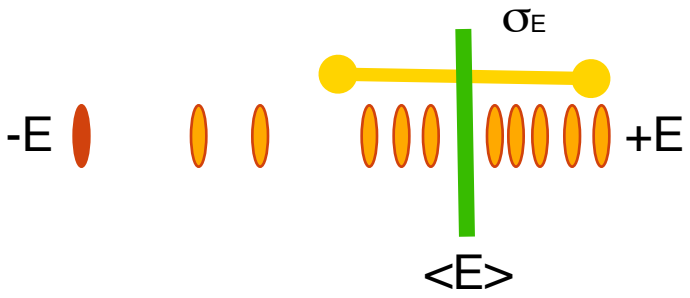
$$\Delta E = -kT \ln\left(\frac{N_1}{N_2}\right) \qquad (1)$$

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

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

Significance of an alignment (score)

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



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

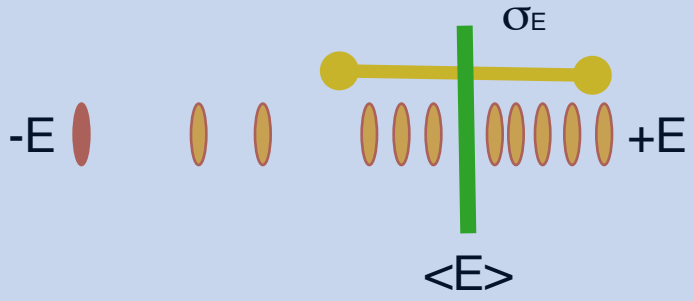
Prosall

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

Deriving

Structural space

Scoring



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

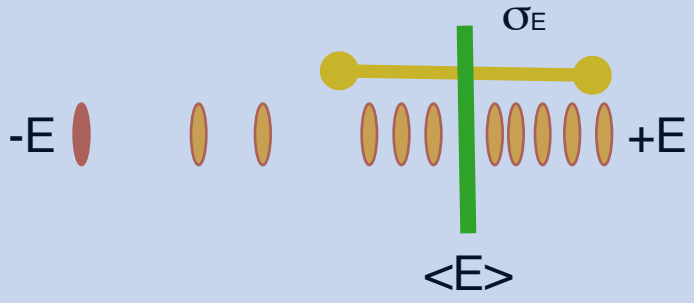
ANOLEA

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

Deriving

Structural space

Scoring



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

all atom potential

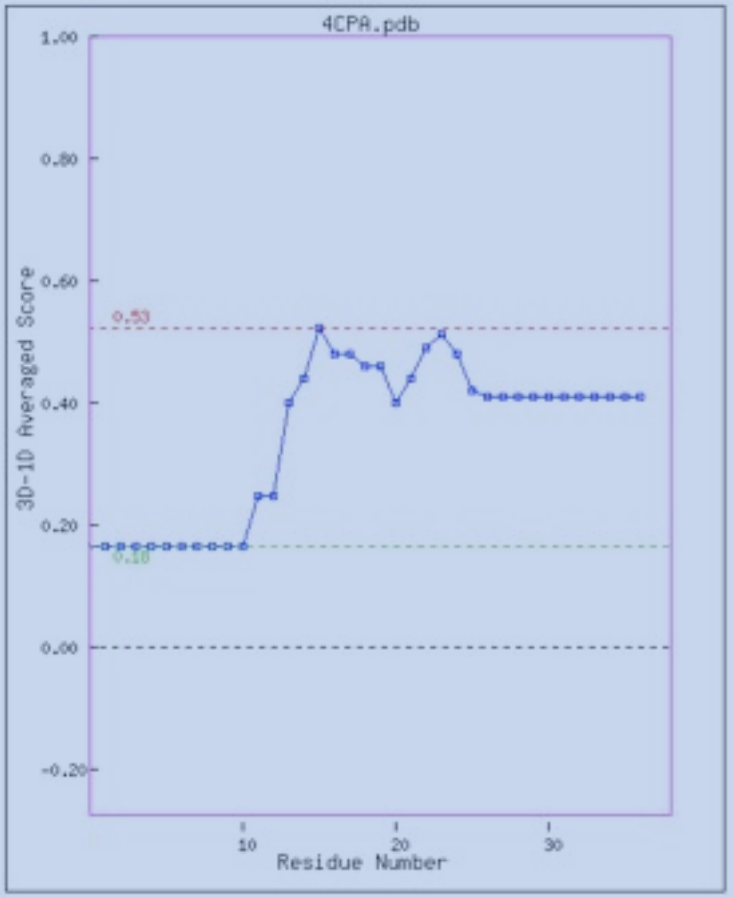
Verify3D

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

Deriving

Structural space

Scoring



DFIRE

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

Deriving

Structural space

Scoring

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

DOPE (MODELLER)

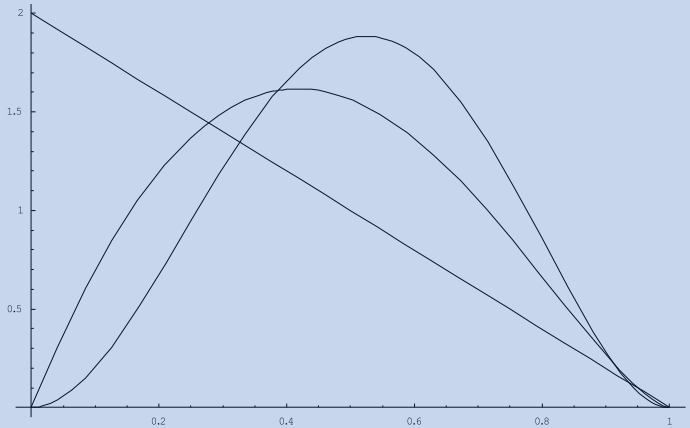
http://www.salilab.org/modeller/

Deriving

Structural space

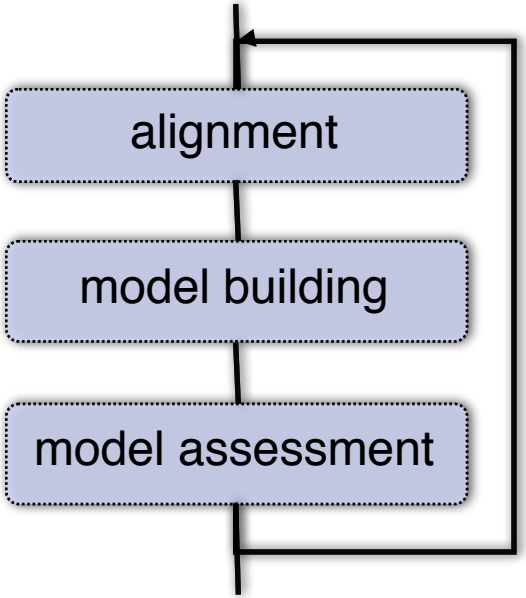
Scoring

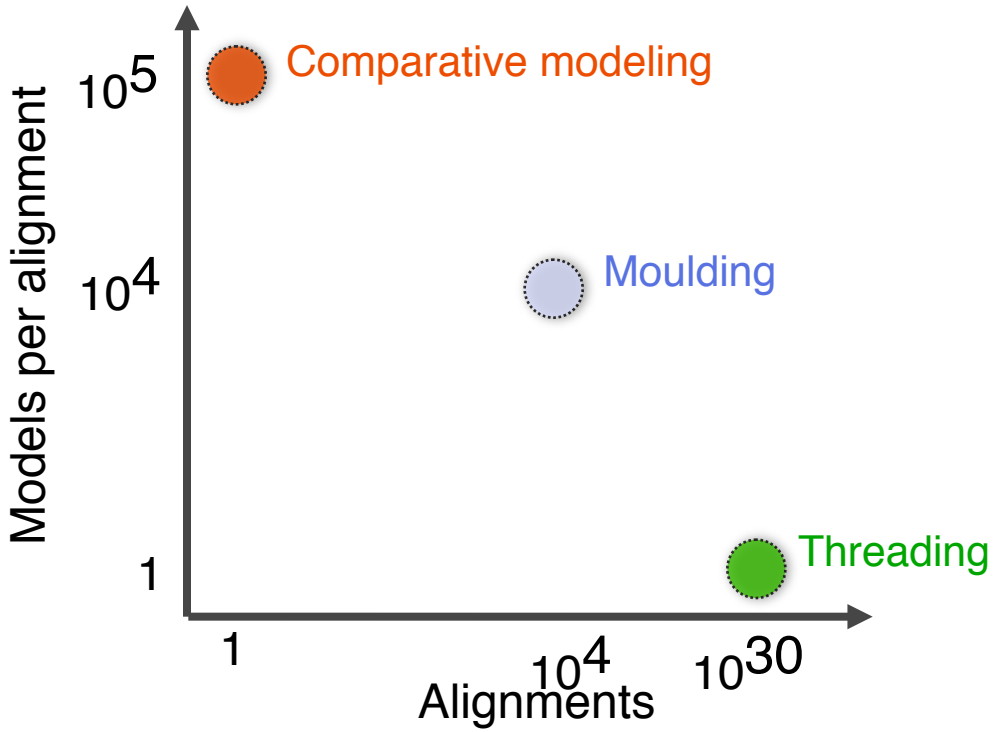
Pseudo-Energy with respect a ideal spherical protein as a reference state



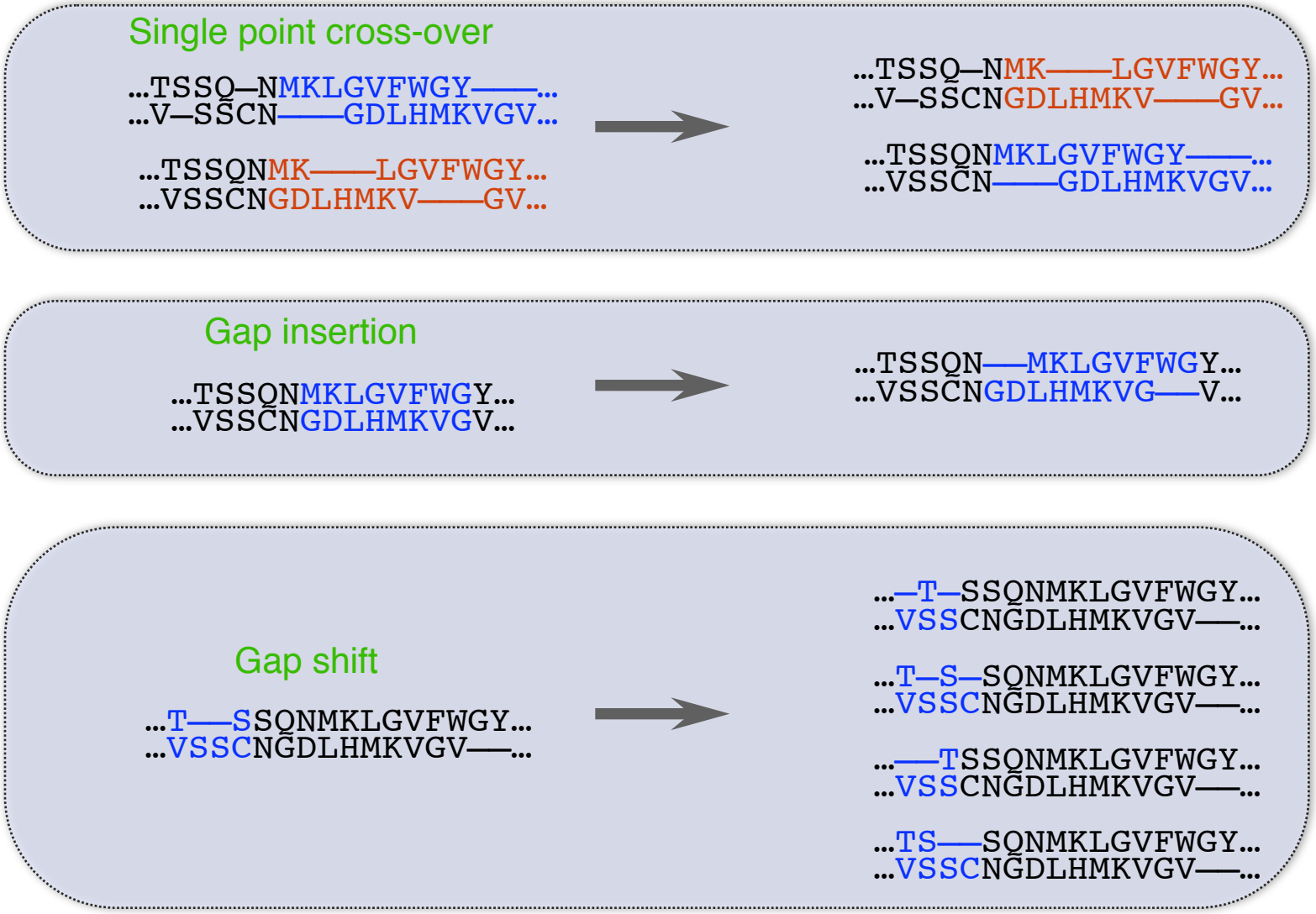
John, Sali (2003). NAR pp31 3982

Moulding: iterative alignment, model building, model assessment





Genetic algorithm operators



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).

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

$Z(score) = (score - \mu)/\sigma$

μ … average score of all models

σ … standard deviation of the scores

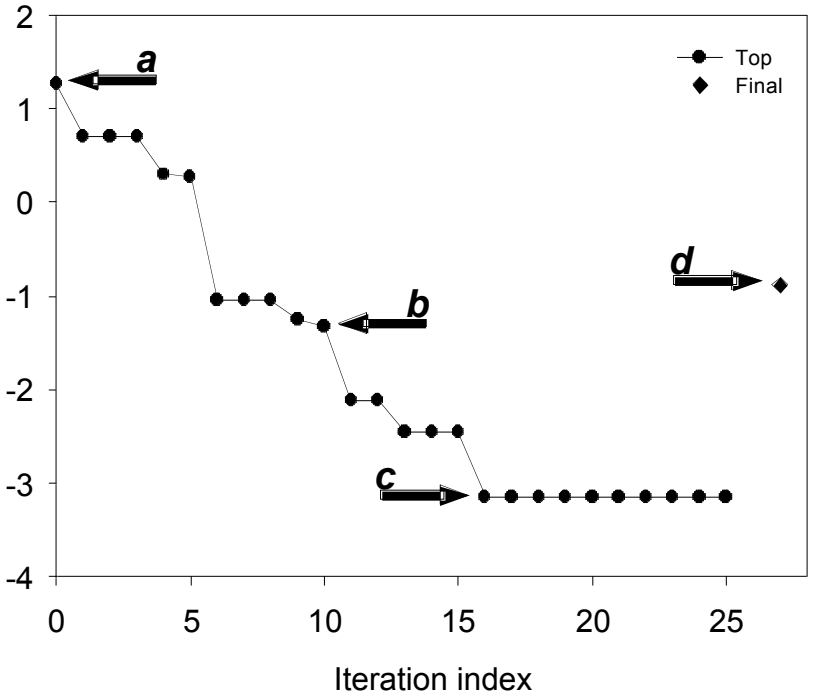
Benchmark with the “very difficult” test set

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

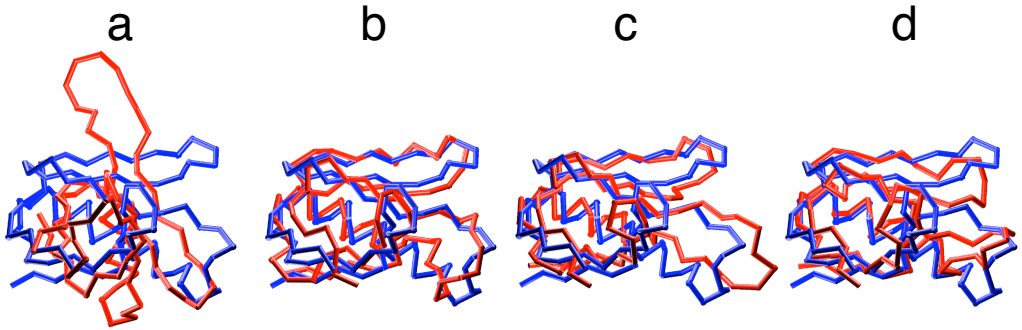
1BOV-1LTS



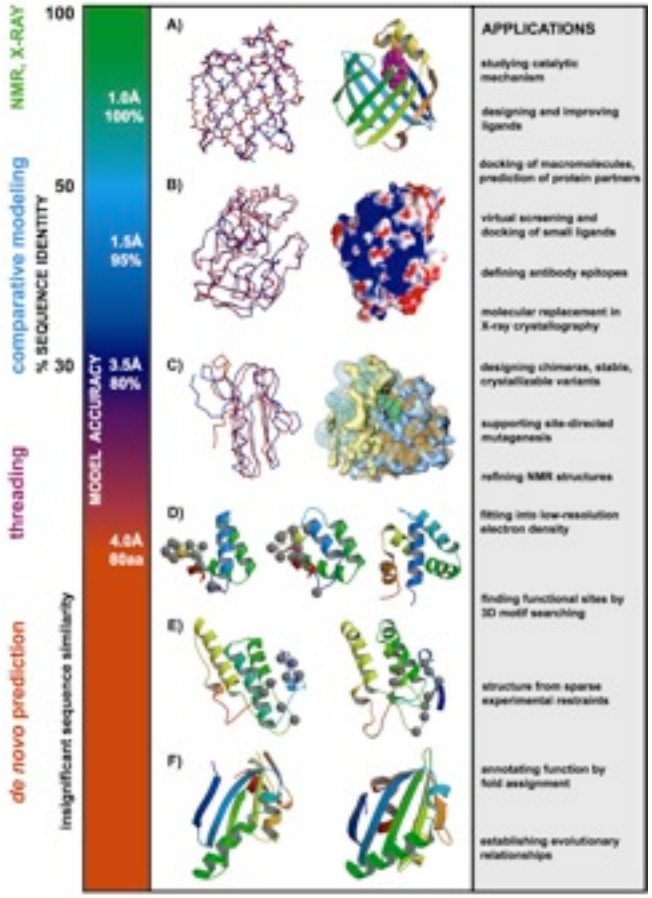
Sequence identity 4.4%

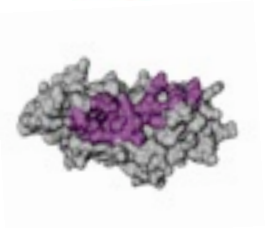
Initial model Cα RMSD 10.1Å

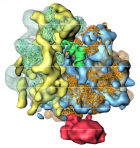
Final model Cα RMSD 3.6Å



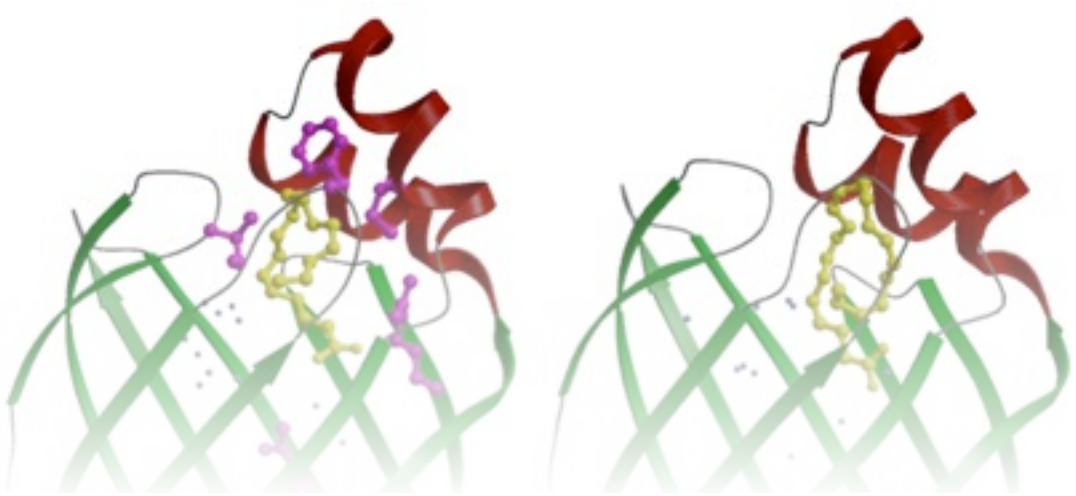
Can we use models to infer function?







T. cruzi



Modeling genes

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

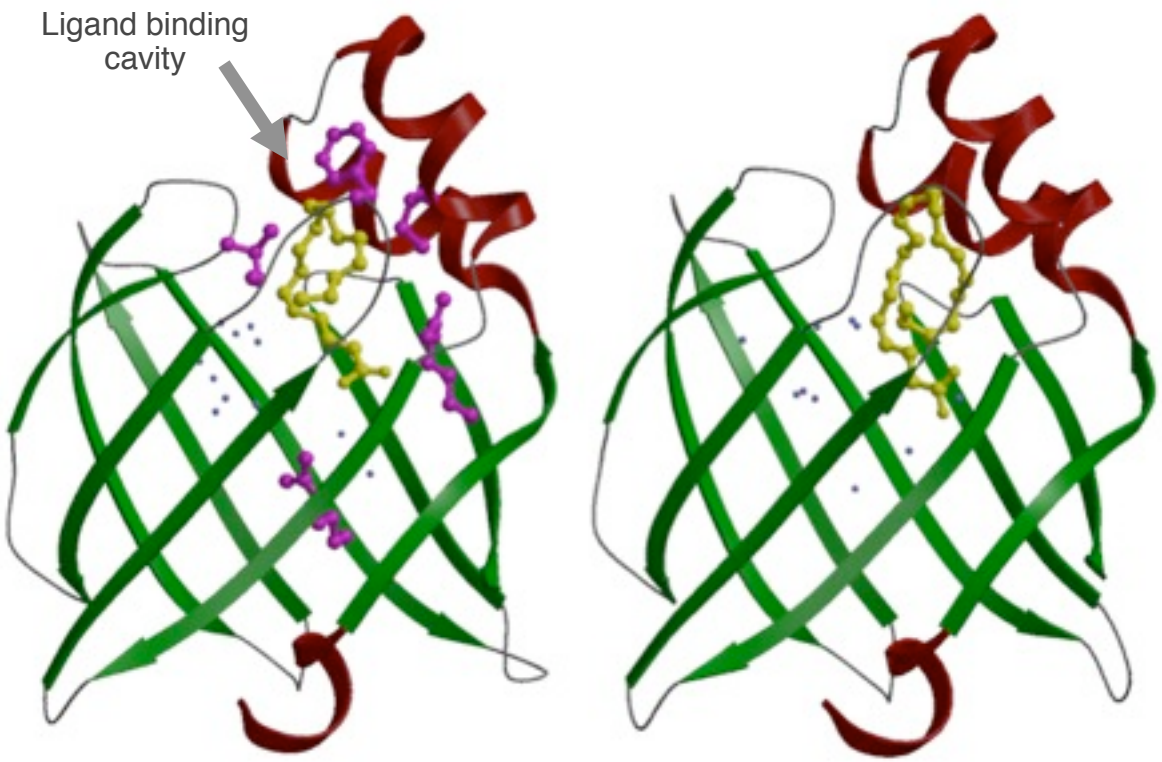
Predicting features of a model that are not present in the template

BLBP/oleic acid

Cavity is not filled

BLBP/docosahexaenoic acid

Cavity is filled



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

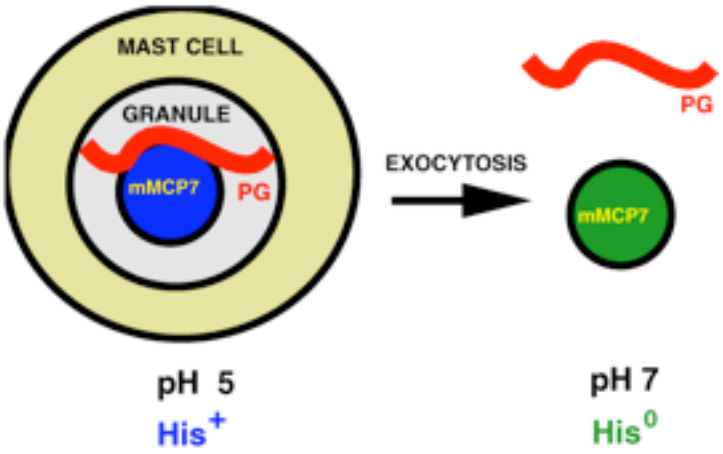
1. BLBP binds fatty acids.

2. Build a 3D model.

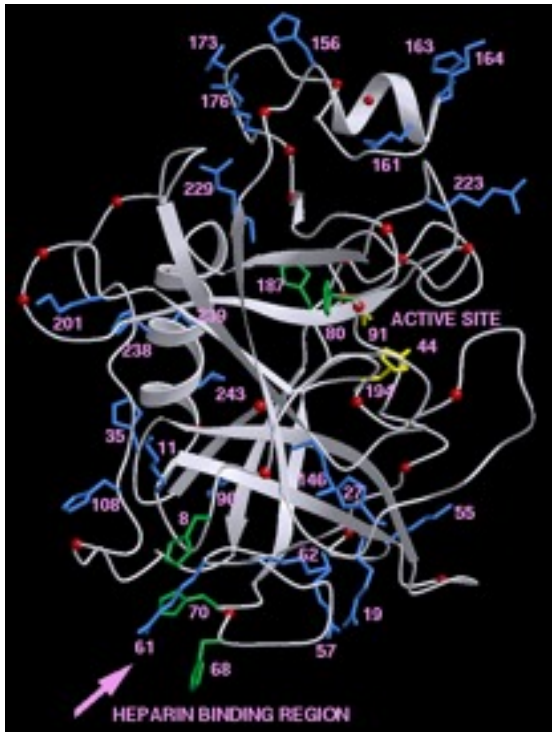
3. Find the fatty acid that fits most snuggly into the ligand binding cavity.

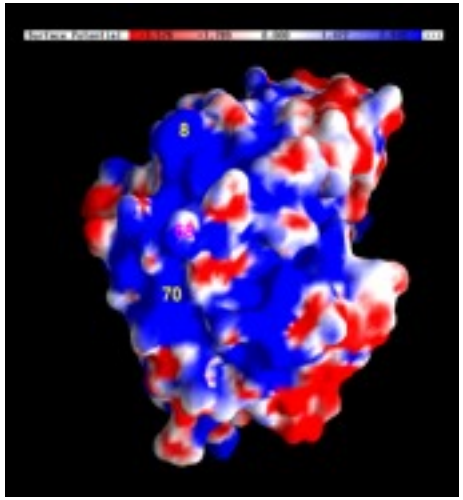
Predicting features of a model that are not present in the template

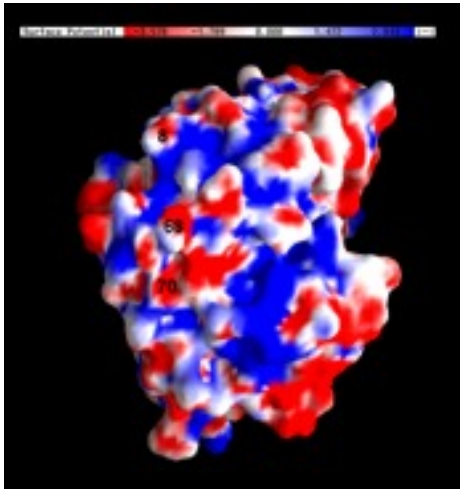
1. mMCPs bind negatively charged proteoglycans through electrostatic interactions
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.



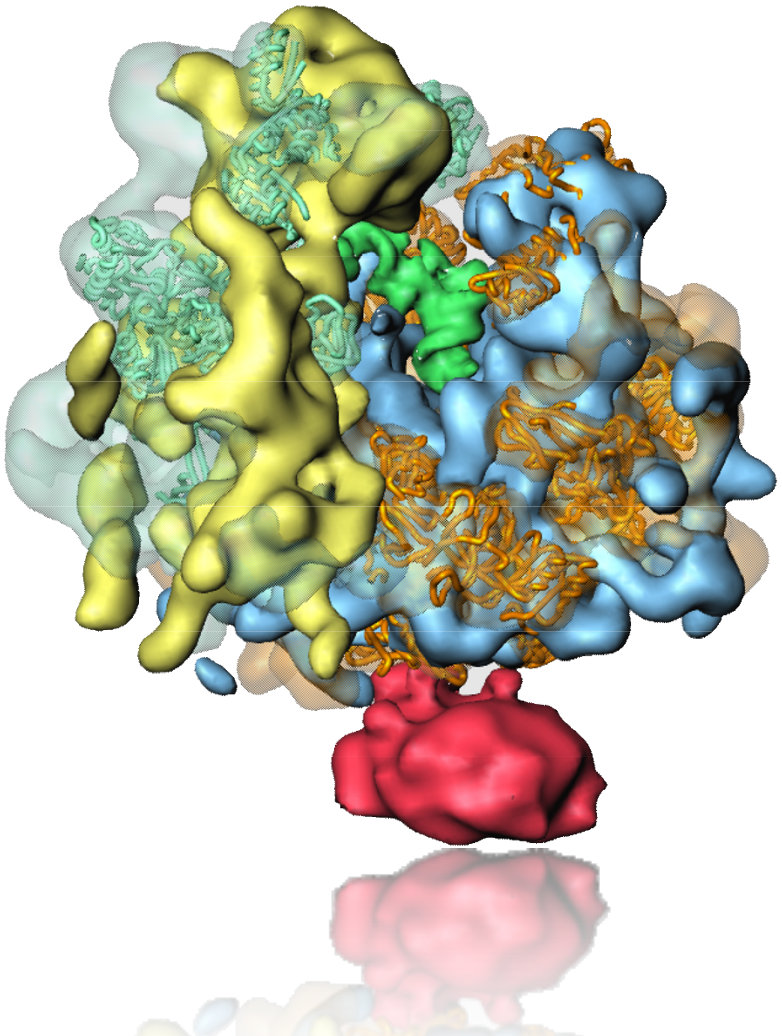
Huang *et al*. *J. Clin. Immunol*. **18**,169,1998.
Matsumoto *et al*. *J.Biol.Chem*. **270**,19524,1995.
Šali *et al*. *J. Biol. Chem*. **268**, 9023, 1993.







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%.

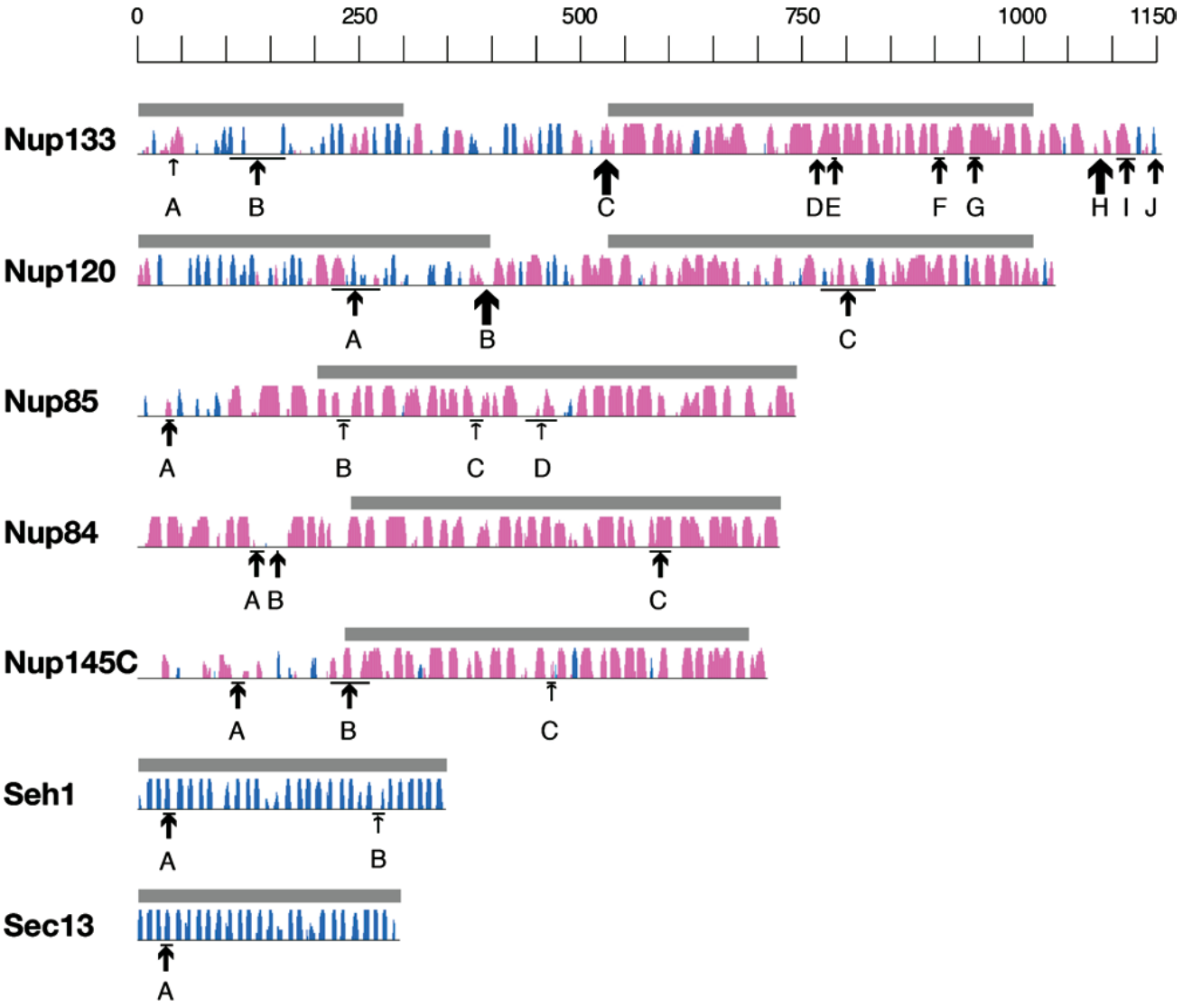
C. Spahn, R. Beckmann, N. Eswar, P. Penczek, A. Sali, G. Blobel, J. Frank. Cell 107, 361-372, 2001.

Common Evolutionary Origin of Coated Vesicles and Nuclear Pore Complexes

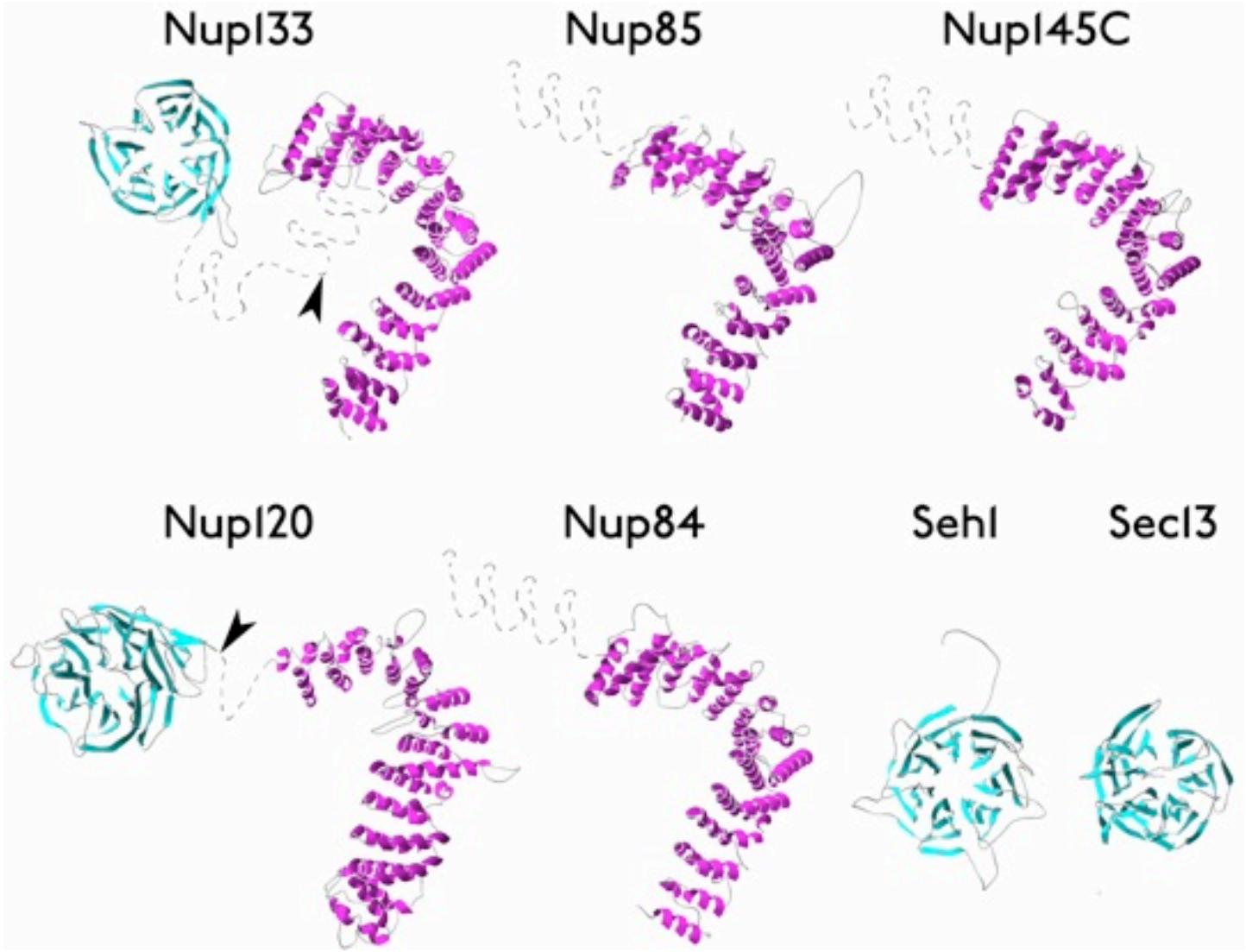
mGenThreader + SALIGN + MOULDER

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

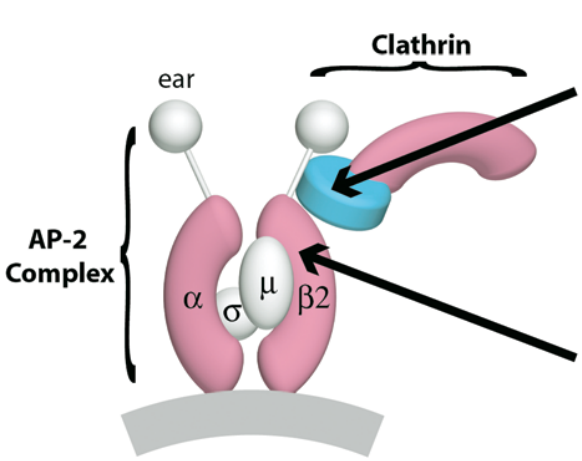
yNup84 complex proteins

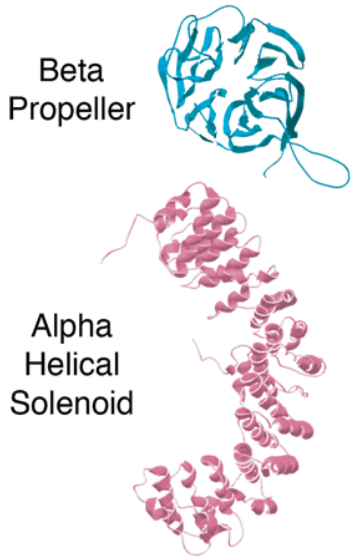


All Nucleoporins in the Nup84 Complex are Predicted to Contain β-Propeller and/or α-Solenoid Folds



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



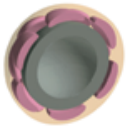


VARIANTS

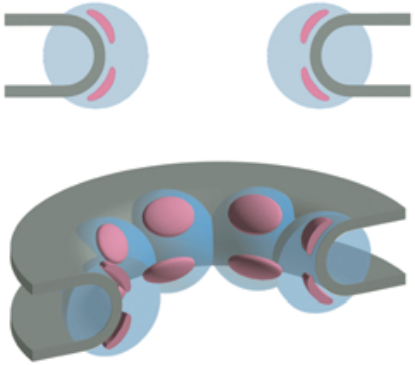
CCVs	*COPI*	*COPII*	*NPC*
Clathrin-N	α-subunit-N β'-subunit-N	Sec13 (Seh1) Sec31-N	Sec13 Seh1 Nup133-N Nup120-N
Clathrin-C α/γ/δ/ε-subunit β-subunits	α-subunit-C β'-subunit-C β-subunit γ-subunit	Sec31-C	Nup133-C Nup120-C Nup85 Nup84 Nup145C

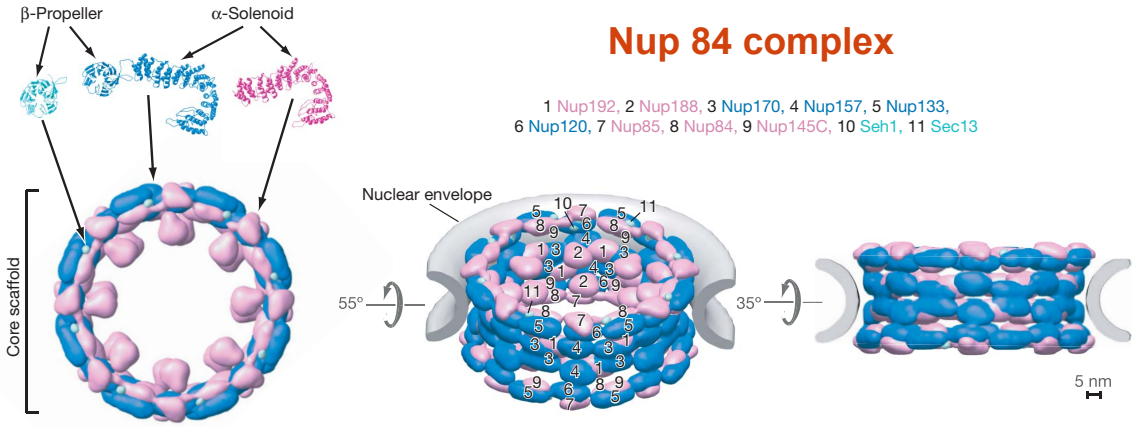
NPC and Coated Vesicles Both Associate with Membranes

Coated Vesicle



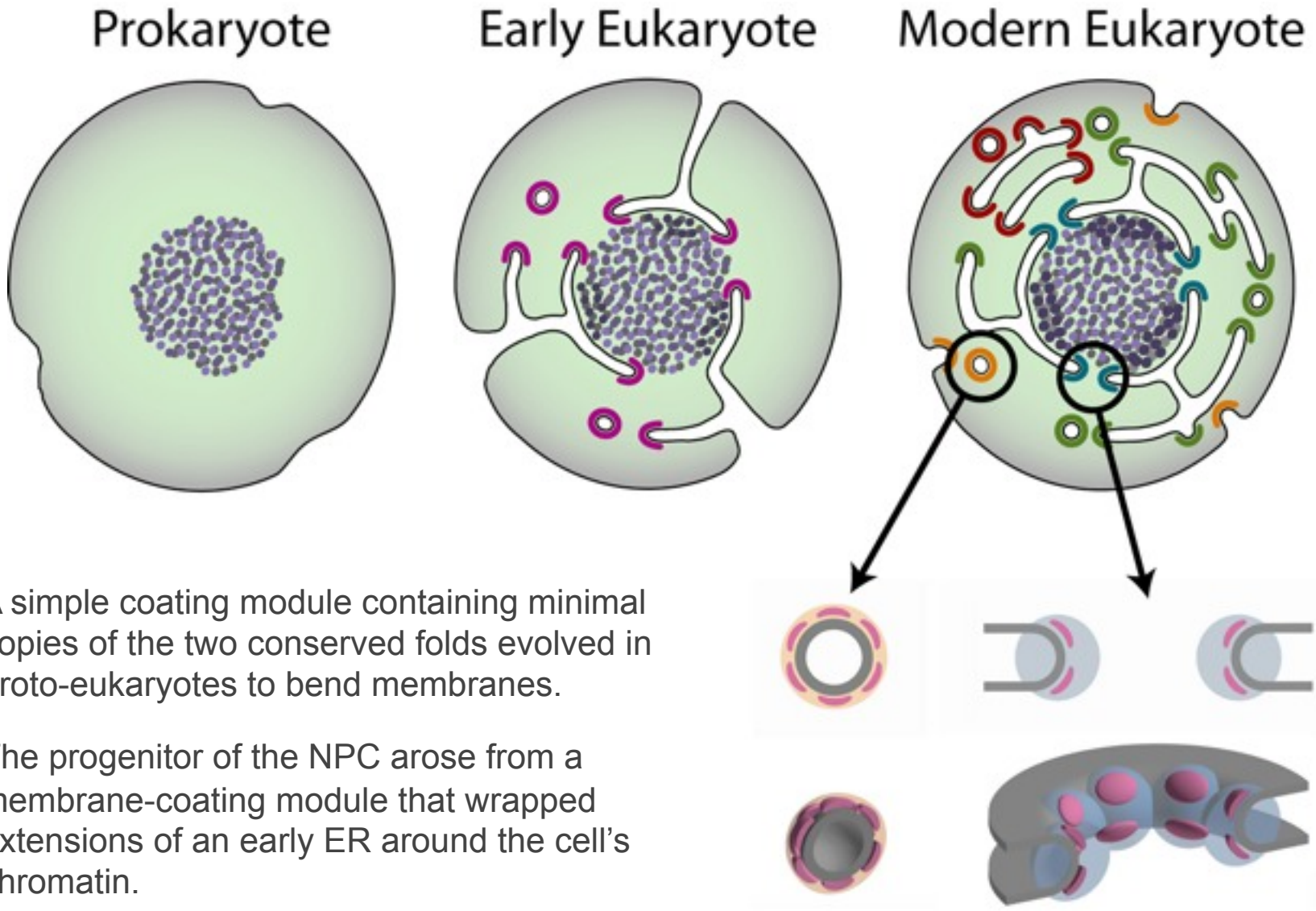
NPC model





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

A Common Evolutionary Origin for Nuclear Pore Complexes and Coated Vesicles?
The proto-coatomer hypothesis

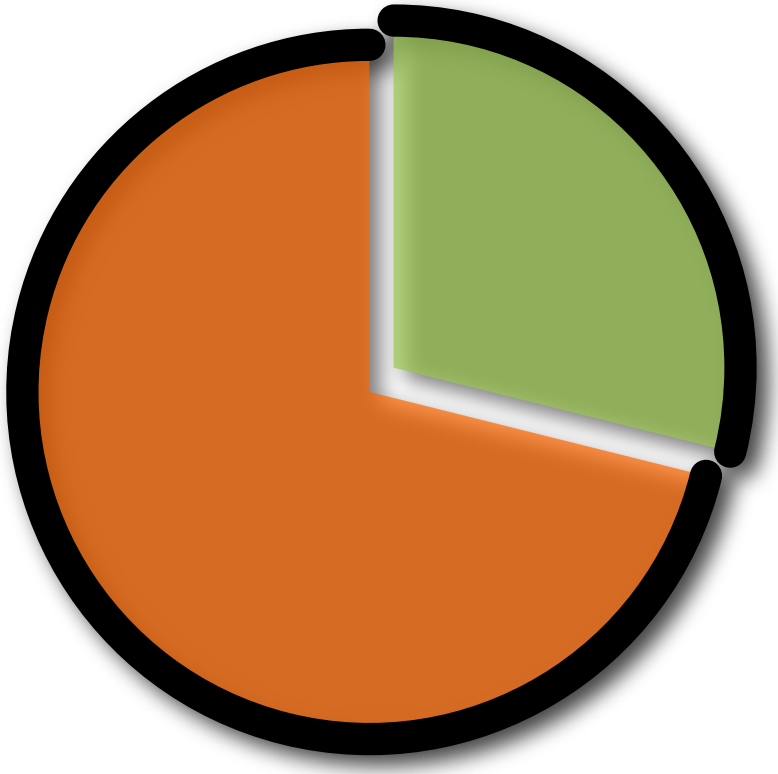


A simple coating module containing minimal copies of the two conserved folds evolved in proto-eukaryotes to bend membranes.

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

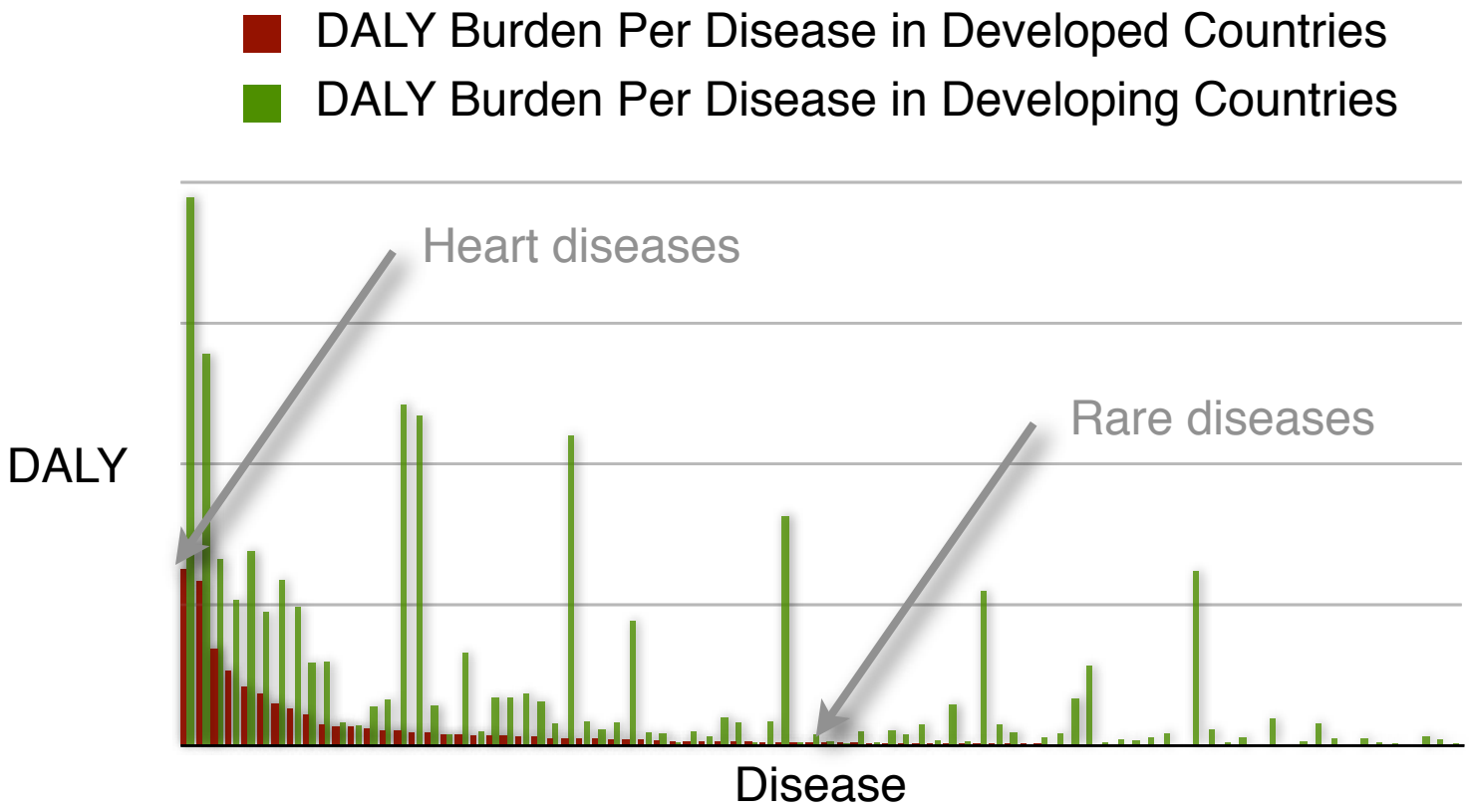
Tropical Disease Initiative (TDI)

Predicting binding sites in protein structure models.



http://www.tropicaldisease.org

Need is High in the Tail



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

DALY - Disability adjusted life years

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

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

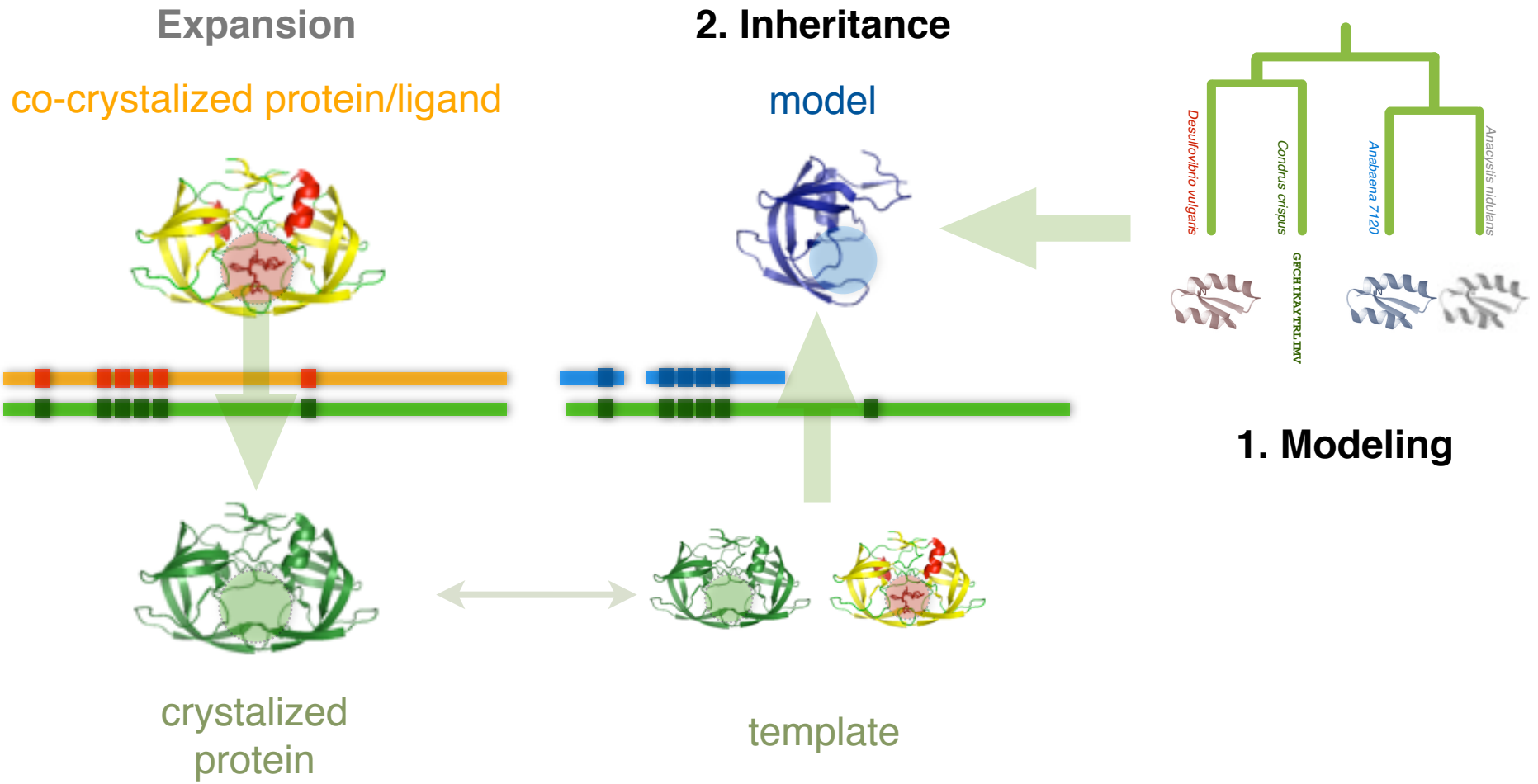
"Unprofitable" Diseases and Global DALY (in 1000's)

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

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

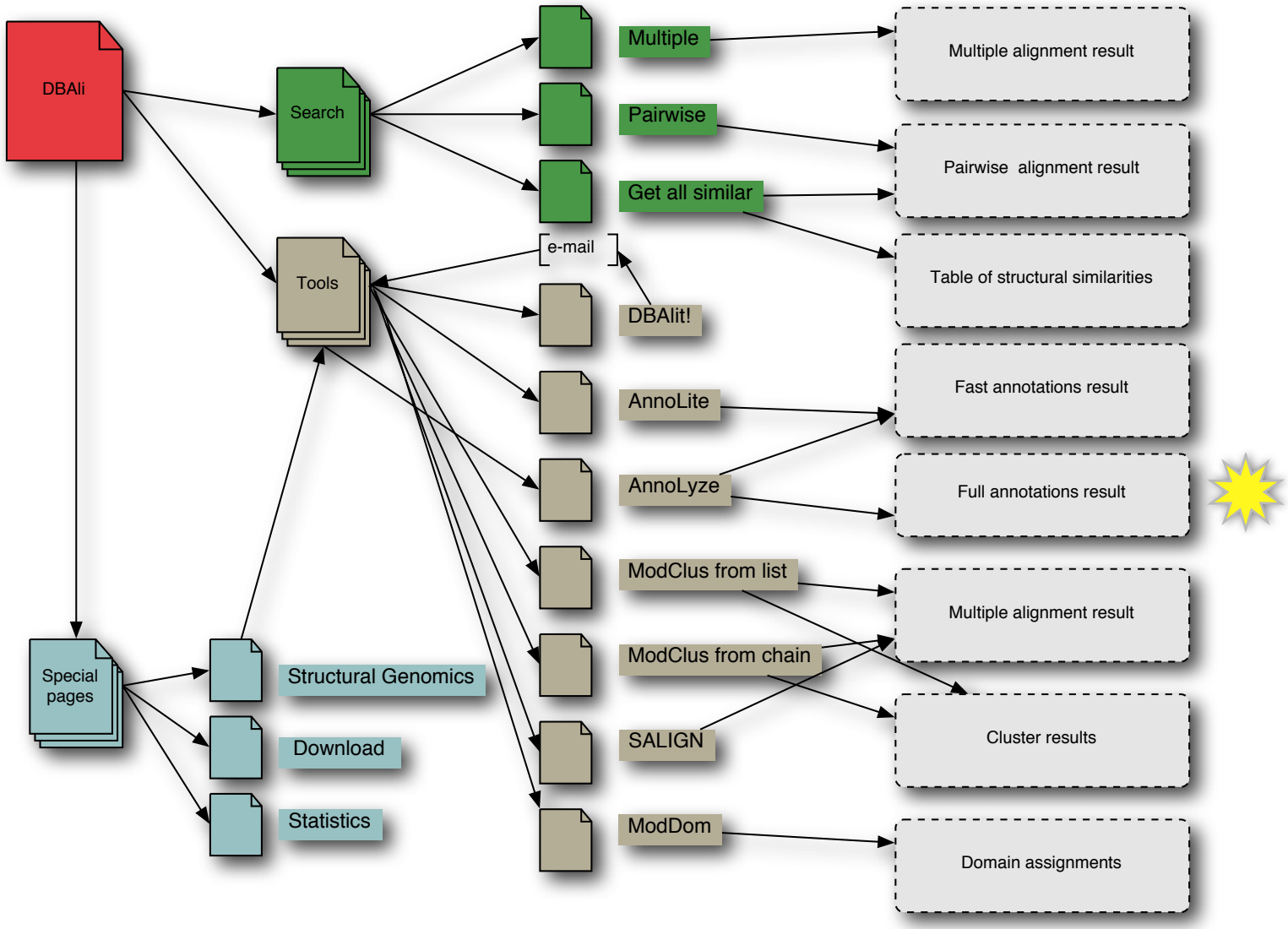
Disease data taken from WHO, *World Health Report 2004*
DALY - Disability adjusted life year in 1000's.
* Officially listed in the WHO Tropical Disease Research disease portfolio.

Comparative docking



DBAli$_{v2.0}$ database

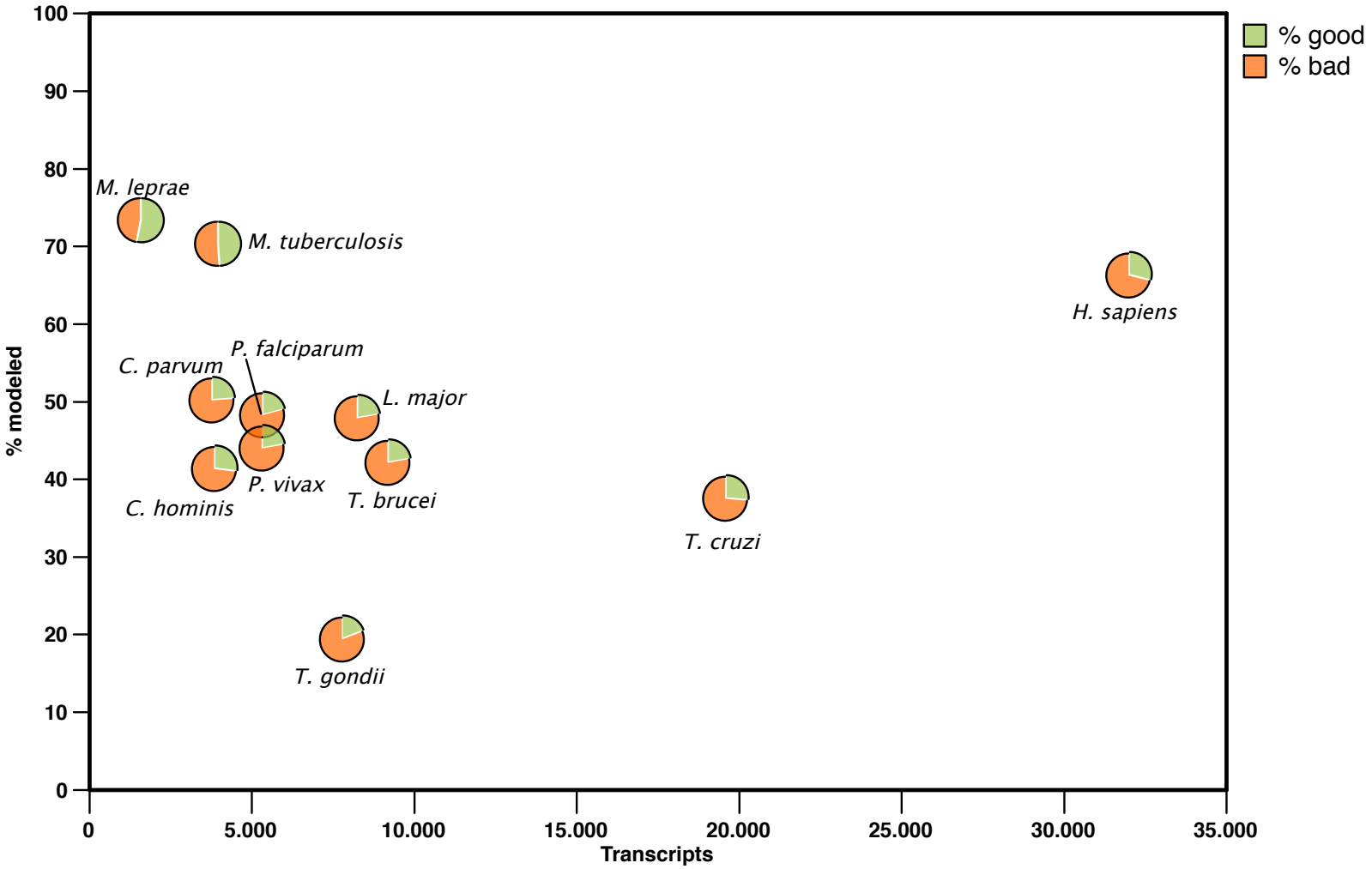
http://www.dbali.org



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

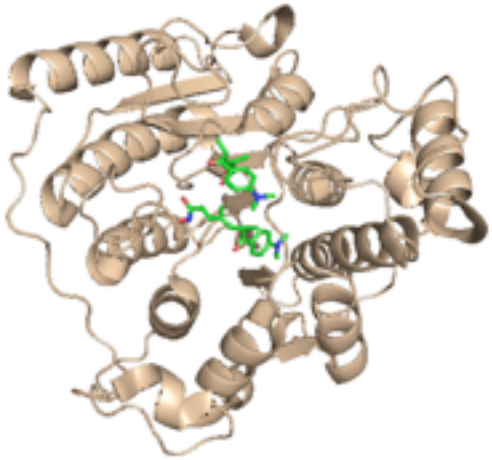
models with inherited ligands

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

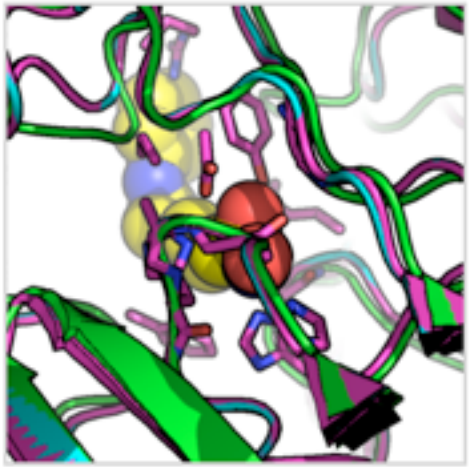
	Transcripts	Modeled targets	Selected models	Inherited ligands	Similar to a drug	Drugs
C. hominis	3,886	1,614	666	197	20	13
C. parvum	3,806	1,918	742	232	24	13
L. major	8,274	3,975	1,409	478	43	20
M. leprae	1,605	1,178	893	310	25	6
M. tuberculosis	3,991	2,808	1,608	365	30	10
P. falciparum	5,363	2,599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

L. major Histone deacetylase 2 + Vorinostat

Template 1t64A a human HDAC8 protein.



PDB	ID	Template	SIM	Model	G	Ligand	Exact	SupStr	SubStr	Similar
1c3sA	83.33/80.00	1t64A	36.00/1.47	LmjF21.0680.1.pdb	90.91/100.00	SHH	DB02546	DB02546	DB02546	DB02546



DB02546 Vorinostat

Small Molecule; Approved; Investigational

Drug categories:

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

Drug indication:

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

L. major Histone deacetylase 2 + Vorinostat

Literature

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

Apicidin: A novel antiprotozoal agent that inhibits parasite histone deacetylase

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

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

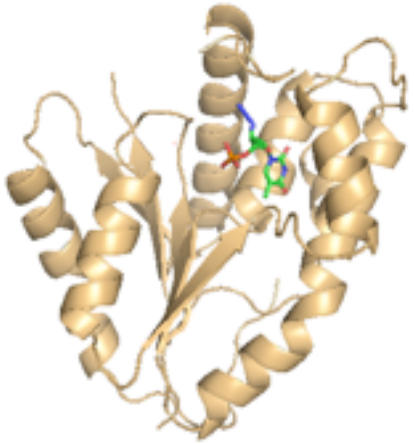
Departments of *Parasite Biochemistry and Cell Biology, ‡Medicinal Chemistry, and §Natural Products Drug Discovery, Merck Research Laboratories, P.O. Box 2000, Rahway, NJ 07065

ANTIMICROBIAL AGENTS AND CHEMOTHERAPY, Apr. 2004, p. 1435–1436 Vol. 48, No. 4
0066-4804/04/$08.00+0 DOI: 10.1128/AAC.48.4.1435–1436.2004

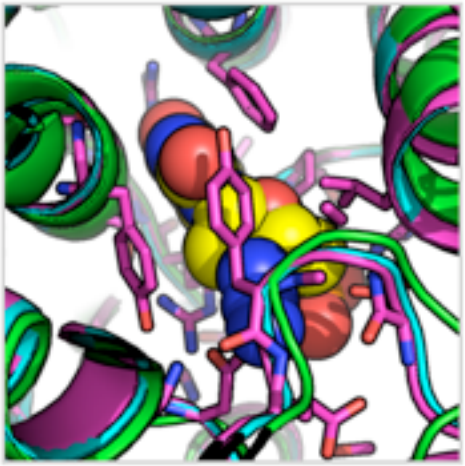
Antimalarial and Antileishmanial Activities of Aroyl-Pyrrolyl-Hydroxyamides, a New Class of Histone Deacetylase Inhibitors

P. falciparum tymidylate kinase + zidovudine

Template 3tmkA a yeast tymidylate kinase.



PDB		Template		Model		Ligand	Exact	SupStr	SubStr	Similar
2tmkB	100.00/100.00	3tmkA	41.00/1.49	PFL2465c.2.pdb	82.61/100.00	ATM		DB00495		DB00495



DB00495 **Zidovudine**

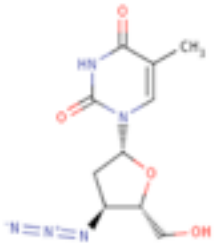
Small Molecule; Approved

Drug categories:

Anti-HIV Agents
Antimetabolites
Nucleoside and Nucleotide Reverse Transcriptase Inhibitors

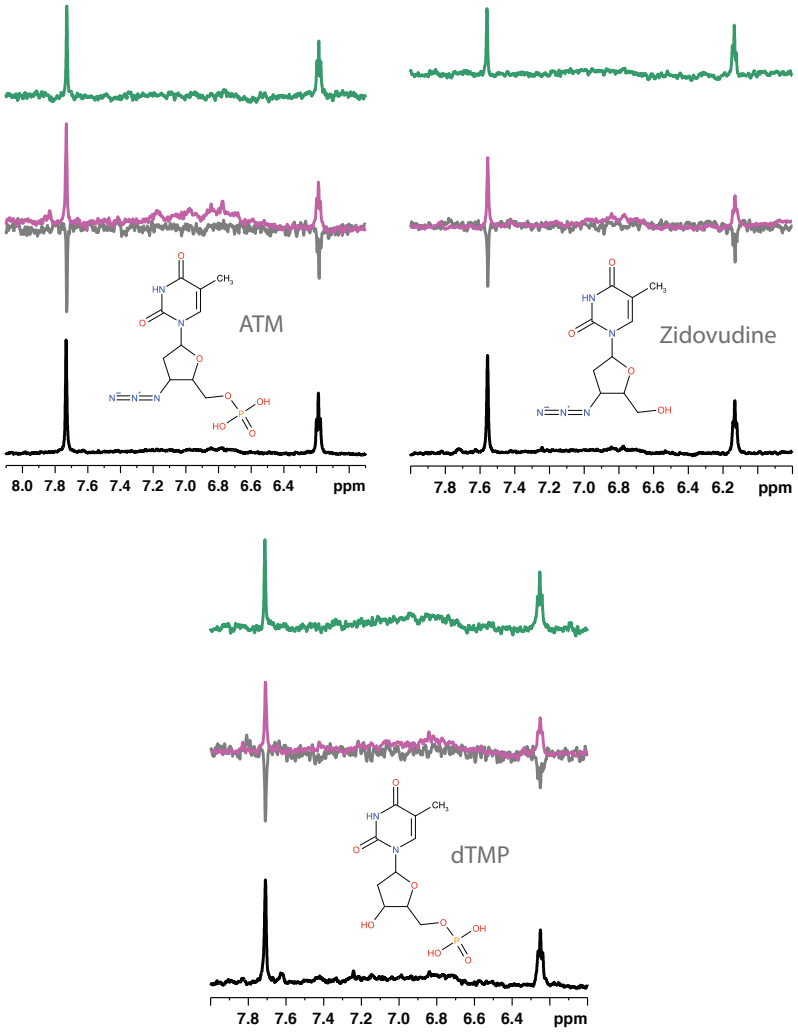
Drug Indication:

For the treatment of human immunovirus (HIV) infections.



P. falciparum tymydilate kinase + zidovudine

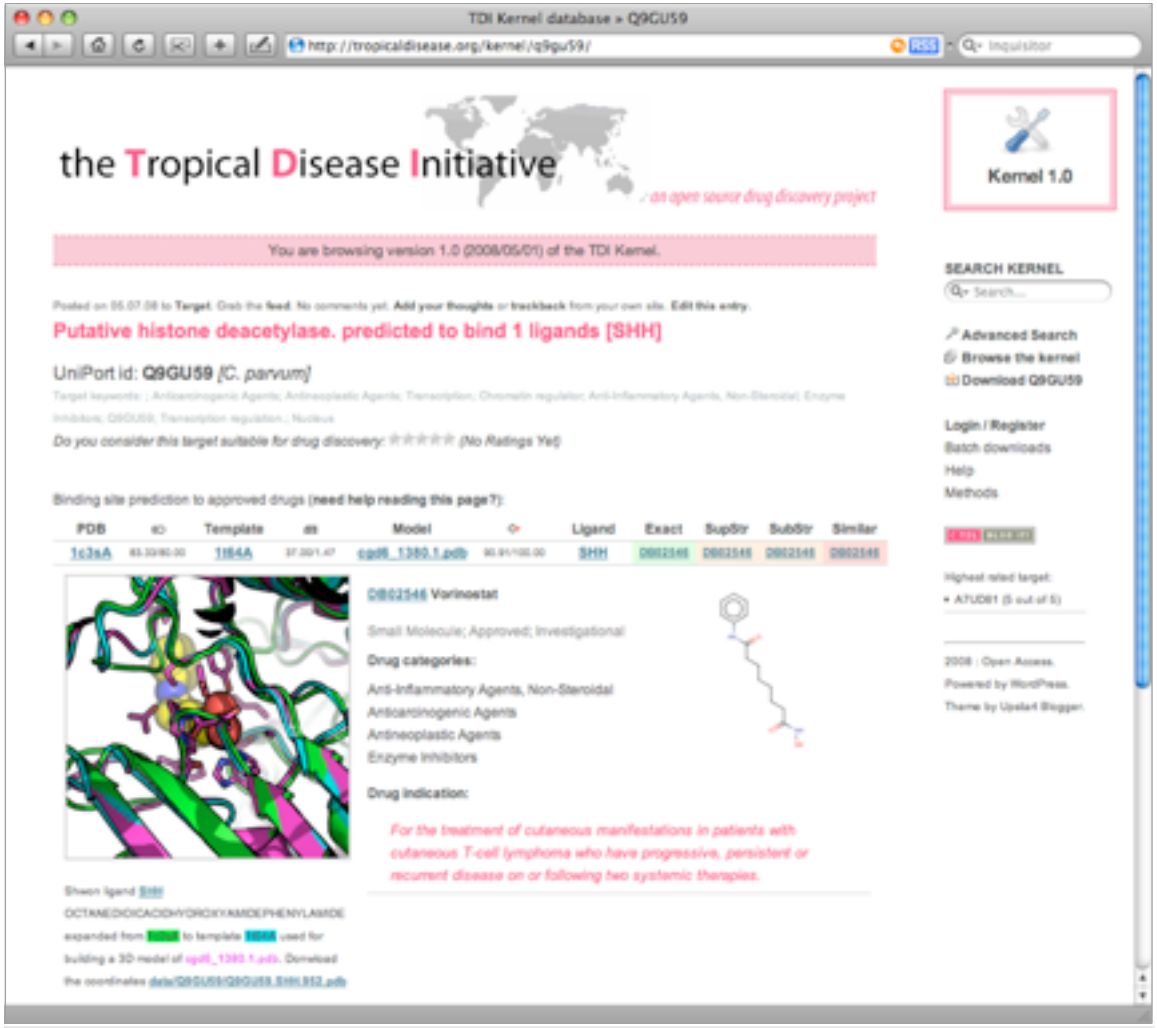
*NMR **Water-LOGSY** and **STD** experiments*



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

TDI's kernel

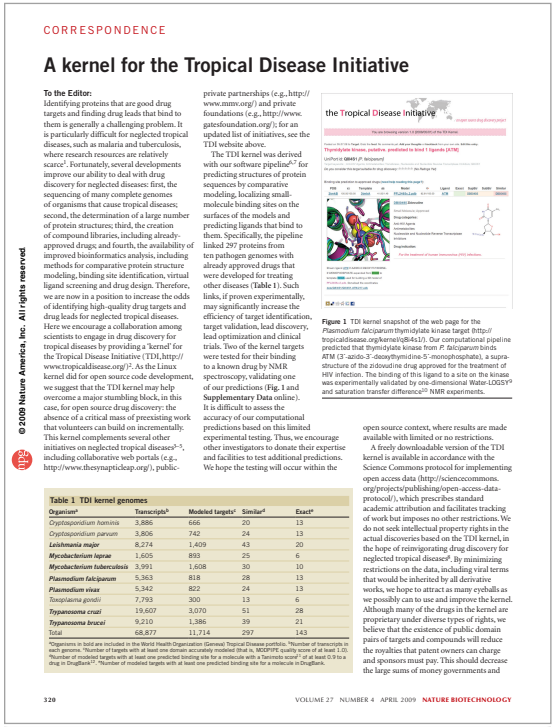
http://tropicaldisease.org/kernel



TDI's kernel

http://tropicaldisease.org/kernel

L. Orti *et al.*, *Nat Biotechnol* **27**, 320 (Apr, 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[1]. 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 comparative protein structure modeling, binding site 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/)[2]. 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[3–5], including collaborative web portals (e.g., http://www.thesynapticleap.org/), public-private partnerships (e.g., http://www.mmv.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 derived with our software pipeline[6,7] 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 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 reinvigorating drug discovery for neglected tropical diseases[8]. By minimizing restrictions on the data, including viral 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 pairs 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

Figure 1 TDI kernel snapshot of the web page for the *Plasmodium falciparum* thymidylate kinase target (http://tropicaldisease.org/kernel/q8k4s1/). Our computational pipeline predicted that thymidylate kinase from *P. falciparum* binds ATM (3′-azido-3′-deoxythymidine-5′-monophosphate), a supra-structure of the zidovudine 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[9] and saturation transfer difference[10] NMR experiments.

Table 1 TDI kernel genomes

Organism[a]	Transcripts[b]	Modeled targets[c]	Similar[d]	Exact[e]
Cryptosporidium hominis	3,886	666	20	13
Cryptosporidium parvum	3,806	742	24	13
Leishmania major	8,274	1,409	43	20
Mycobacterium leprae	1,605	893	25	6
Mycobacterium tuberculosis	3,991	1,608	30	10
Plasmodium falciparum	5,363	818	28	13
Plasmodium vivax	5,342	822	24	13
Toxoplasma gondii	7,793	300	13	6
Trypanosoma cruzi	19,607	3,070	51	28
Trypanosoma brucei	9,210	1,386	39	21
Total	68,877	11,714	297	143

[a]Organisms in bold are included in the World Health Organization (Geneva) Tropical Disease portfolio. [b]Number of transcripts in each genome. [c]Number of targets with at least one domain accurately modeled (that is, MODPIPE quality score of at least 1.0). [d]Number of modeled targets with at least one predicted binding site for a molecule with a Tanimoto score[11] of at least 0.70 to a drug in DrugBank[12]. [e]Number of modeled targets with at least one predicted binding site for a molecule in DrugBank.

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

OPEN ACCESS Freely available online

PLoS NEGLECTED TROPICAL DISEASES

A Kernel for Open Source Drug Discovery in Tropical Diseases

Leticia Ortí[1,2], Rodrigo J. Carbajo[2], Ursula Pieper[3], Narayanan Eswar[3¤], Stephen M. Maurer[4], Arti K. Rai[5], Ginger Taylor[6], Matthew H. Todd[7], Antonio Pineda-Lucena[2], Andrej Šali[2*], Marc A. Marti-Renom[1*]

1 Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain, **2** Structural Biology Laboratory, Medicinal Chemistry Department, Centro de Investigación Príncipe Felipe, Valencia, Spain, **3** 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, **4** Gould School of Law, University of Southern California, Los Angeles, California, United States of America, **5** School of Law, Duke University, Durham, North Carolina, United States of America, **6** The Synaptic Leap, San Ramon, California, United States of America, **7** 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: Ortí 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

Funding: MAM-R acknowledges the support from a Spanish Ministerio de Educación y Ciencia grant (BIO2007/66670). AS acknowledges the support from the Sandler Family Supporting Foundation and the National Institutes of Health (R01 GM54762, U54 GM074945, P01 AI035707, and P01 GM71790). AP-L acknowledges 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 Homer, 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.

* E-mail: sali@salilab.org (AS); mmarti@cipf.es (MAM-R)

¤ Current address: DuPont Knowledge Center, Hyderabad, India

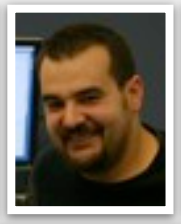
Introduction

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 published. 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 detectably 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

Acknowledgments

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


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

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

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)