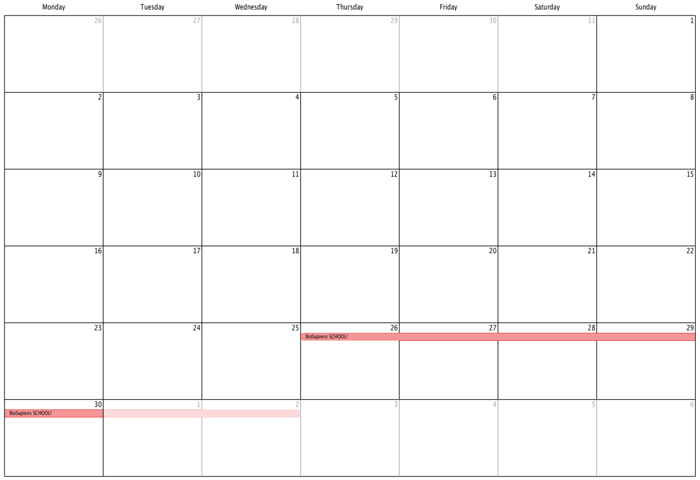
6th Permanent European School on Bioinformatics

http://bioinfo.cipf.es/6ESB/

Preliminary program for the BioSapiens 6th ESB

26th April to 30th of April 2007

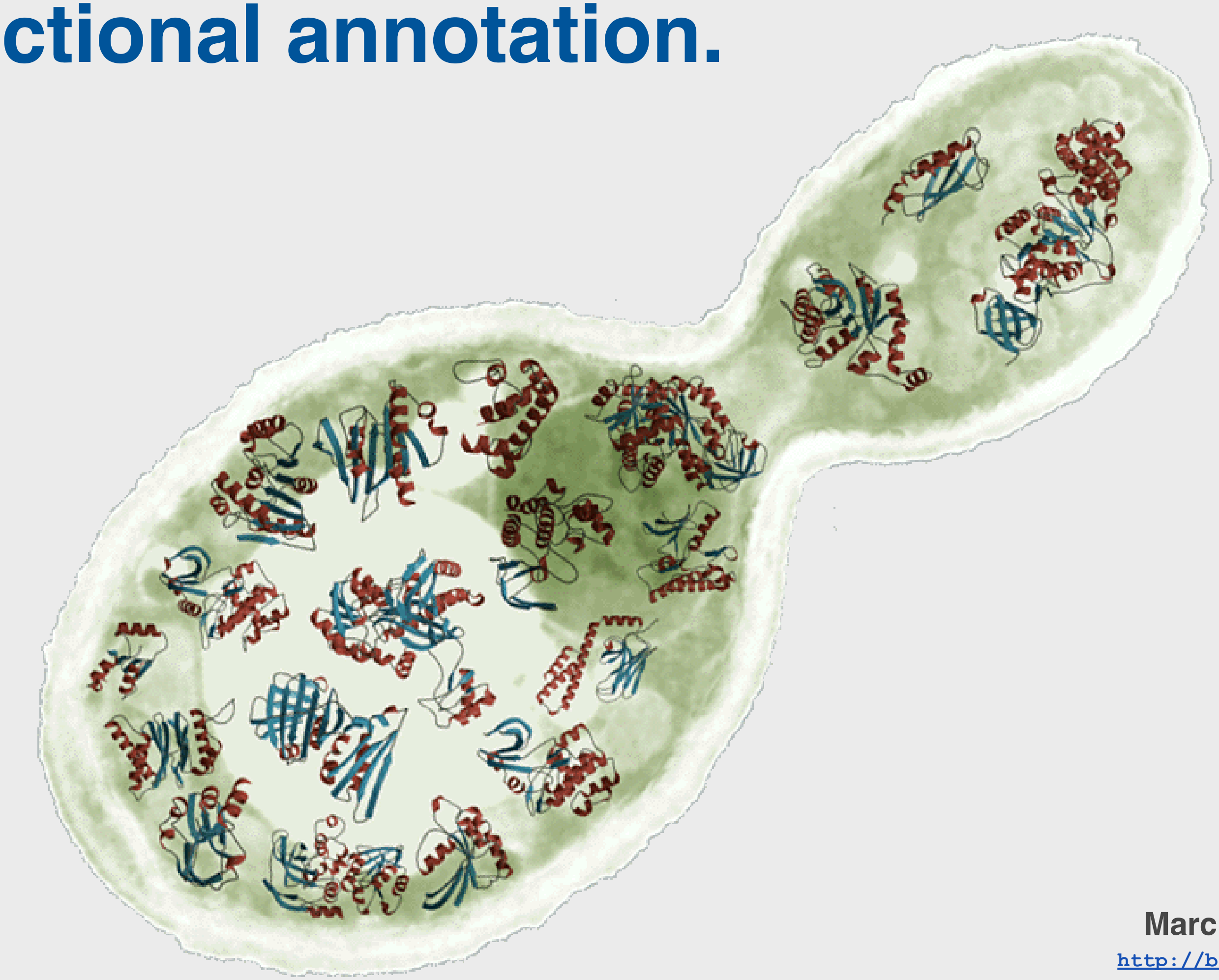


Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
26	27	28	29	30	31	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26 BioSapiens SCHOOL!	27	28	29
30 BioSapiens SCHOOL!	1	2	3	4	5	6

- **Day 1 (April 26th) Databases**
  Trainees from the European Bioinformatics Institute (EBI).
- **Day 2 (April 27th) Analysis of microarray data**
  Trainees to be confirmed (Brazma's group)
- **Day 3 (April 28th) Proteins and protein families**
  Trainee Dr. Yaniv Lowenstein (Linial's group)
- **Day 4 (April 29th) Protein structure prediction**
  Trainees to be confirmed (Marti-Renom's group)
- **Day 5 (April 30th) Systems biology**
  Trainee Dr. Ildefonso Cases (Valencia's groups)

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

Comparative protein structure models for functional annotation.



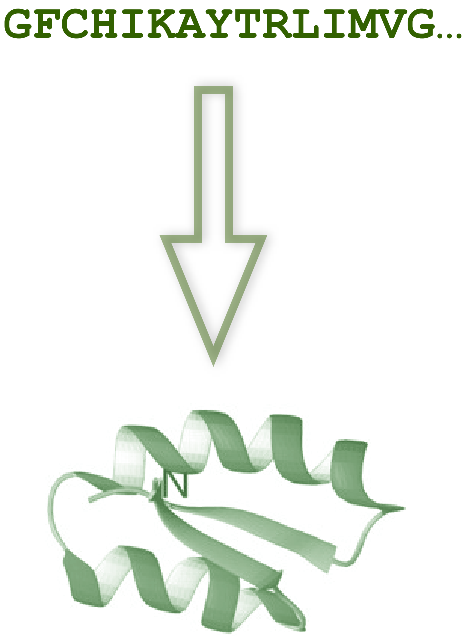
Marc A. Marti-Renom

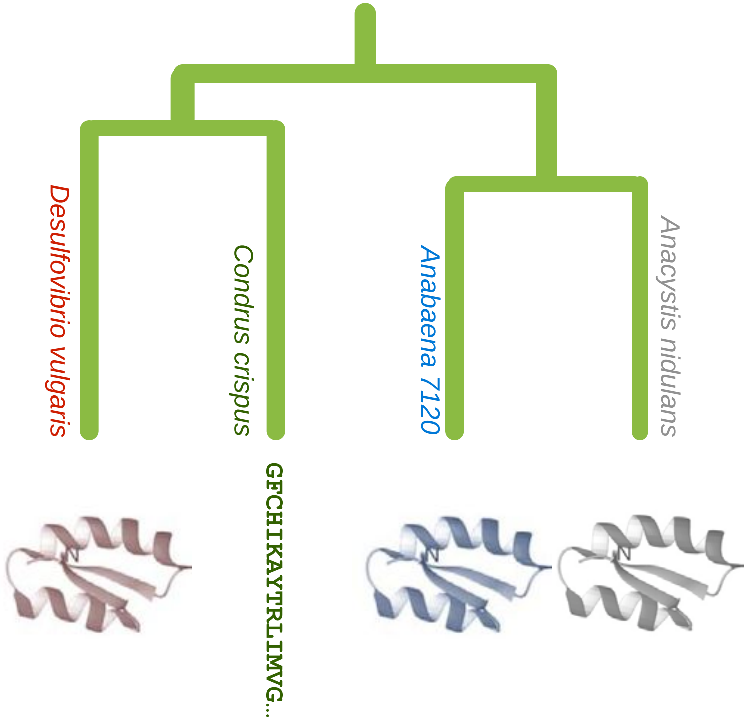
http://bioinfo.cipf.es/squ/

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

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Principles of protein structure





Folding (physics)

Ab initio prediction

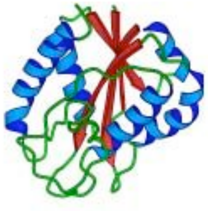
Evolution (rules)

Threading

Comparative Modeling

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

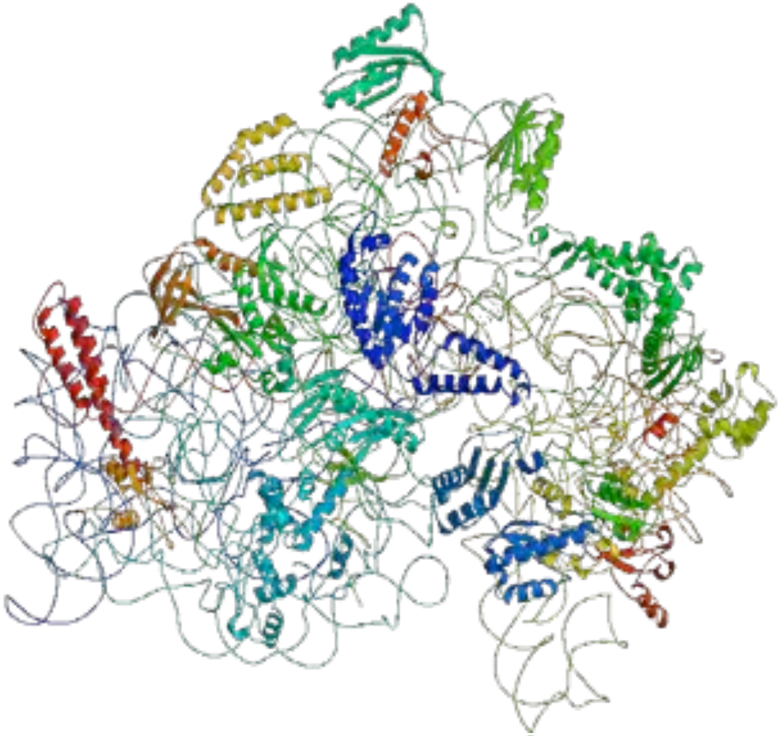
From domains to assemblies



domains

proteins

assemblies

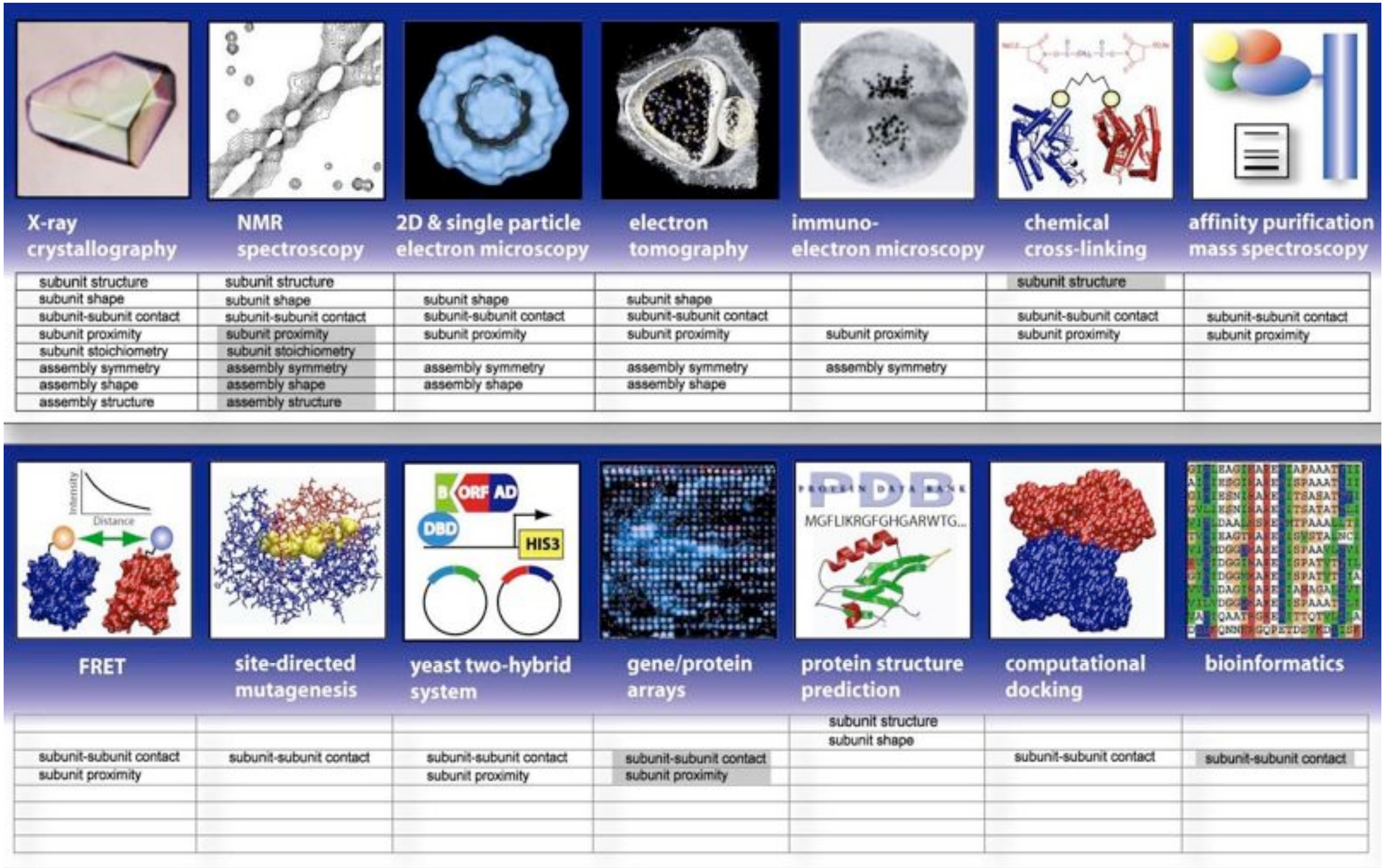


~2.5 domains in a protein
a few domain partners per domain

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

Determining the structures of proteins and assemblies

Use structural information from any
source: measurement, first principles, rules,
resolution: low or high resolution
to obtain the set of all models that are consistent with it.



X-ray crystallography	NMR spectroscopy	2D & single particle electron microscopy	electron tomography	immuno-electron microscopy	chemical cross-linking	affinity purification mass spectroscopy
subunit structure	subunit structure				subunit structure	
subunit shape	subunit shape	subunit shape	subunit shape			
subunit-subunit contact	subunit-subunit contact	subunit-subunit contact	subunit-subunit contact		subunit-subunit contact	subunit-subunit contact
subunit proximity	subunit proximity	subunit proximity	subunit proximity	subunit proximity	subunit proximity	subunit proximity
subunit stoichiometry	subunit stoichiometry					
assembly symmetry	assembly symmetry	assembly symmetry	assembly symmetry	assembly symmetry		
assembly shape	assembly shape	assembly shape	assembly shape			
assembly structure	assembly structure					

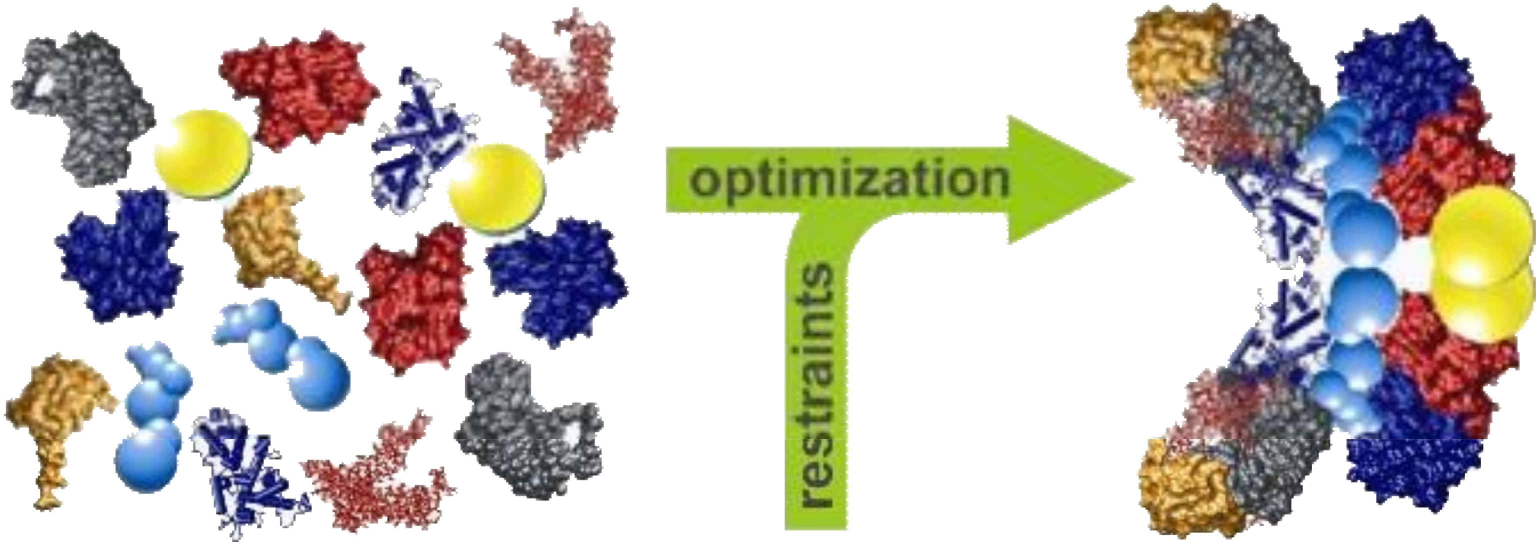
FRET	site-directed mutagenesis	yeast two-hybrid system	gene/protein arrays	protein structure prediction	computational docking	bioinformatics
				subunit structure		
				subunit shape		
subunit-subunit contact	subunit-subunit contact	subunit-subunit contact	subunit-subunit contact		subunit-subunit contact	subunit-subunit contact
subunit proximity		subunit proximity	subunit proximity			

Sali et al. Nature, 422, 216 (2003)

Modeling by satisfaction of spatial restraints

1) Representation of a system.
2) Scoring function (spatial restraints).
3) Optimization.

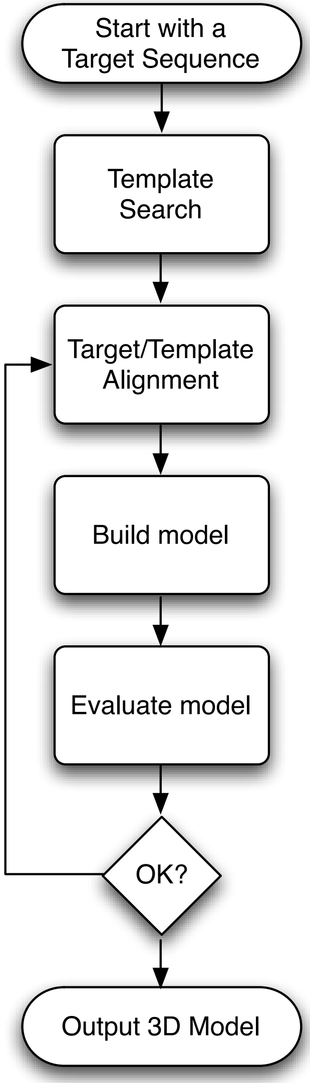
There is nothing but points and restraints on them.



Alber et al. Structure, 13, 435 (2005)

Comparative modeling by satisfaction of spatial restraints

MODELLER



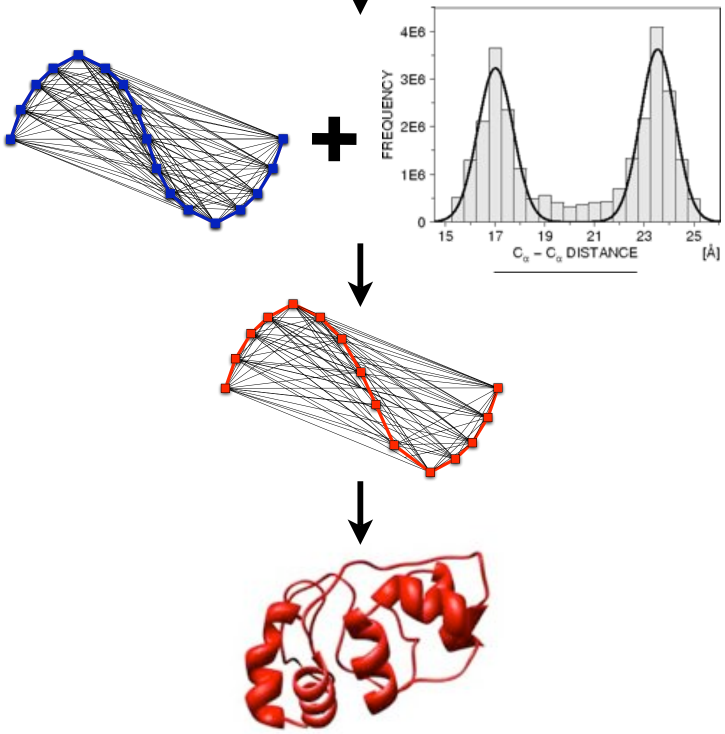
Given an alignment...

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

apply these features as restraints on your target sequence

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

```
MSVIPKR--GNCEQTSE
ASILPKRLFGNCEQTSD
```

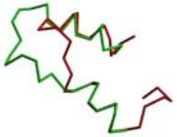


A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.

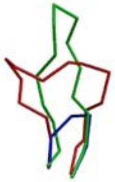
Comparative modeling by satisfaction of spatial restraints

Types of errors and their impact

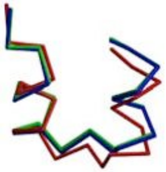
Wrong fold



Miss alignments

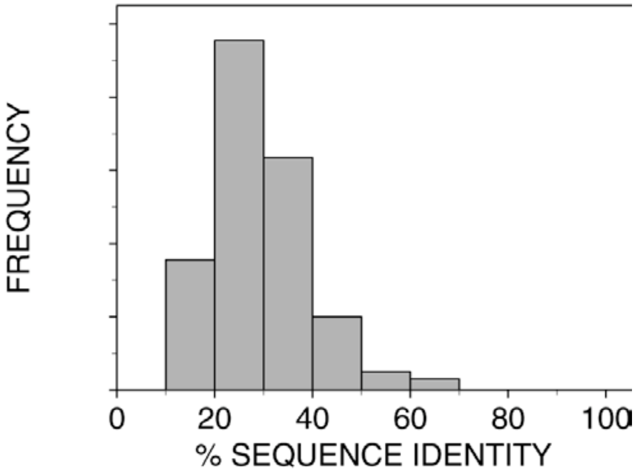


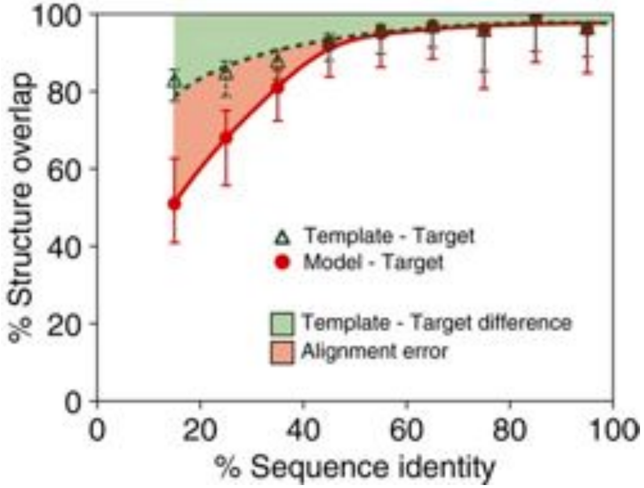
Loop regions



Rigid body distortions

Side-chain packing





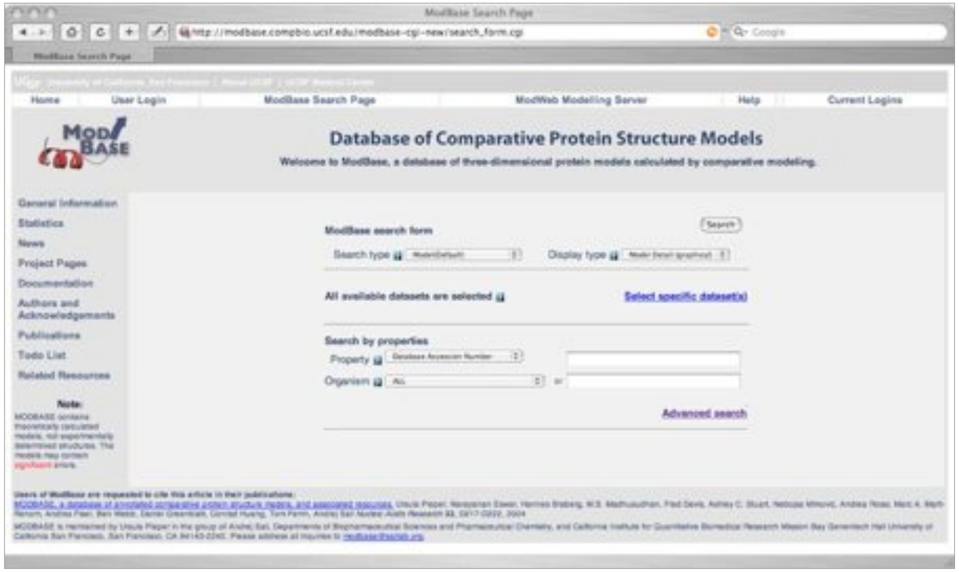
Marti-Renom etal. Ann Rev Biophys Biomol Struct (2000) 29, 291

ModBase Statistics

Large-scale modeling of the TrEMBL-SWISSPROT databases

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

Sequences (total)	1,930,692
Sequences (modeled)	1,084,784
Models	3,094,542



Database of Comparative Protein Structure Models

Welcome to ModBase, a database of three-dimensional protein models calculated by comparative modeling.

ModBase search form

Search type: Model(Default)

Display type: Model Detail (graphical)

All available datasets are selected

Search by properties

Property: Database Accession Number

Organism: All

Note: MODBASE contains theoretically calculated models, not experimentally determined structures. The models may contain significant errors.

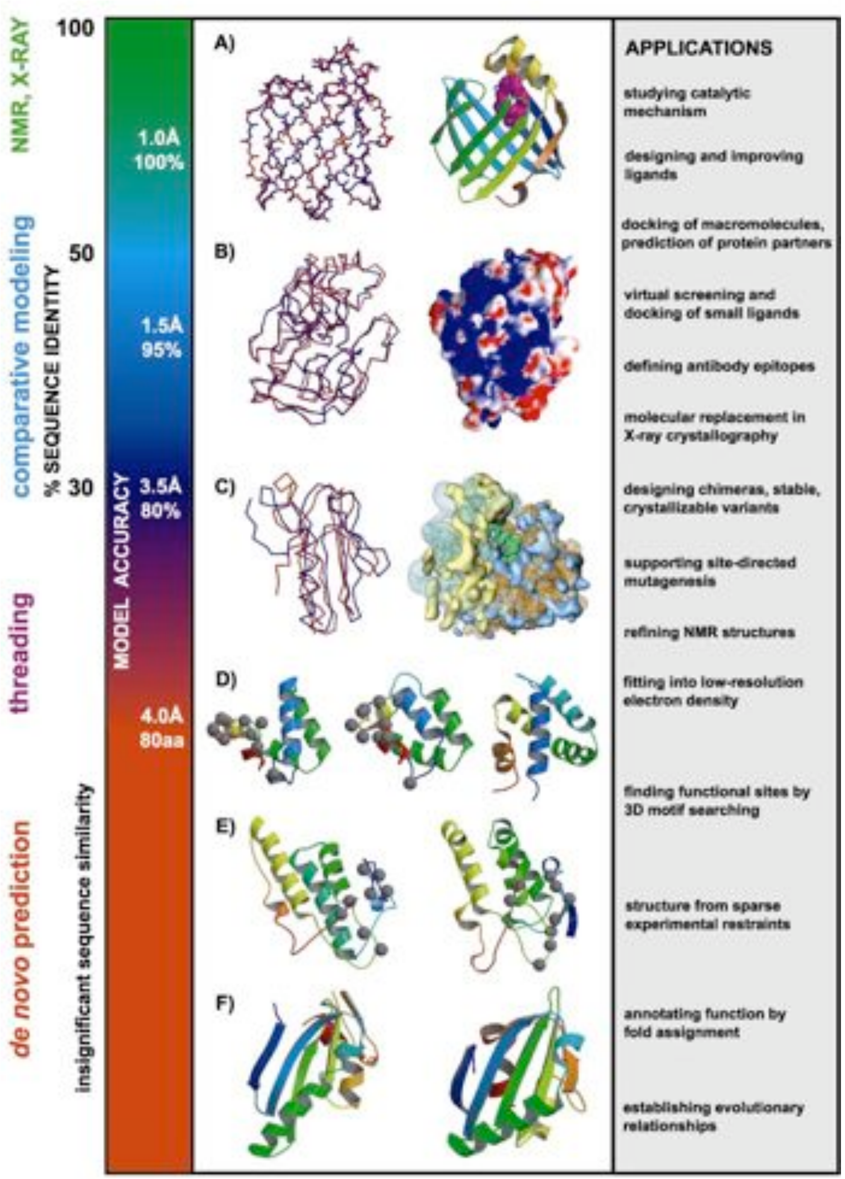
Users of ModBase are requested to cite this article in their publications:
MODBASE, a database of annotated comparative protein structure models, and associated resources. Ursula Pieper, Narayanan Eswar, Hannes Braberg, M.S. Madhusudhan, Fred Davis, Ashley C. Stuart, Nebojsa Mirkovic, Andrea Rossi, Marc A. Marti-Renom, Andras Fiser, Ben Webb, Daniel Greenbaum, Conrad Huang, Tom Ferrin, Andrej Sali Nucleic Acids Research 33, D217-D222, 2004

MODBASE is maintained by Ursula Pieper in the group of Andrej Sali, Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research Mission Bay Genentech Hall University of California San Francisco, San Francisco, CA 94143-2240. Please address all inquiries to modbase@salilab.org

UCSF
University of California
San Francisco

Pieper et al. NAR 34, D291 (2006)

Utility of protein structure models, despite errors


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

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

For many protein structures function is *unknown*

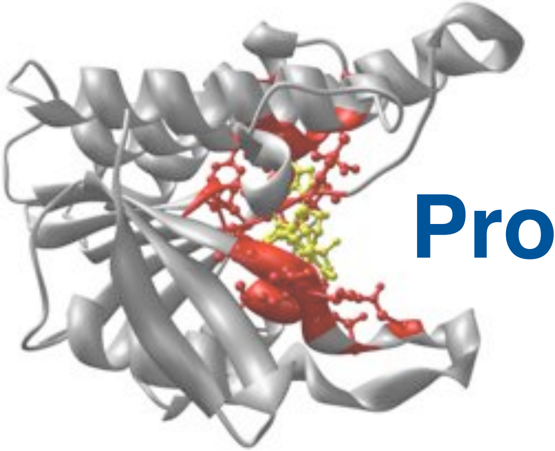
	Structural Genomics*	Traditional methods
Annotated**	654	28,342
Not Annotated	506 (43.6%)	6,815 (19,4%)
Total deposited	1,160	35,157

** 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

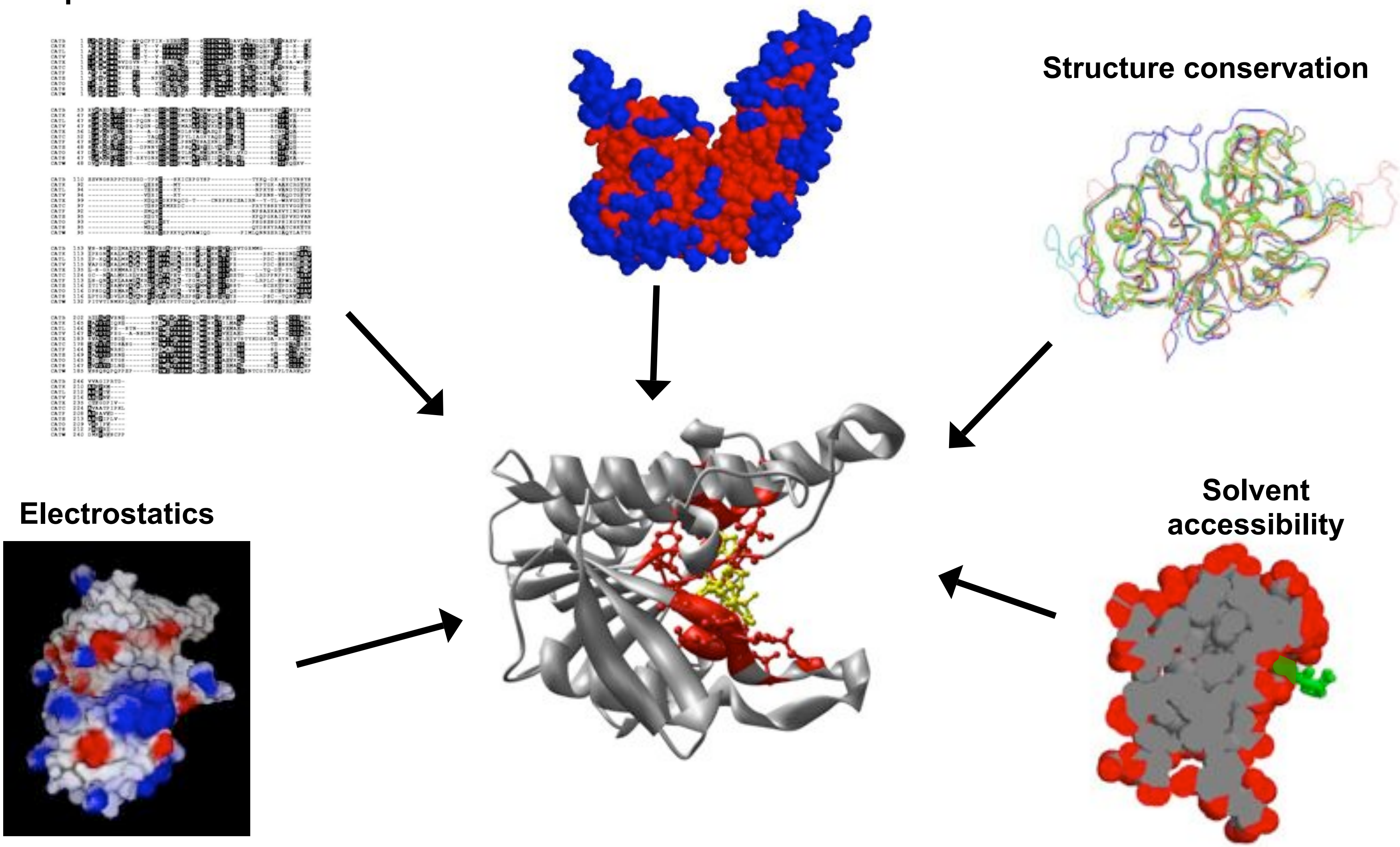


Protein function from structure

ab-initio *localization of binding sites*

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

Representation



Sequence conservation

Surface geometry

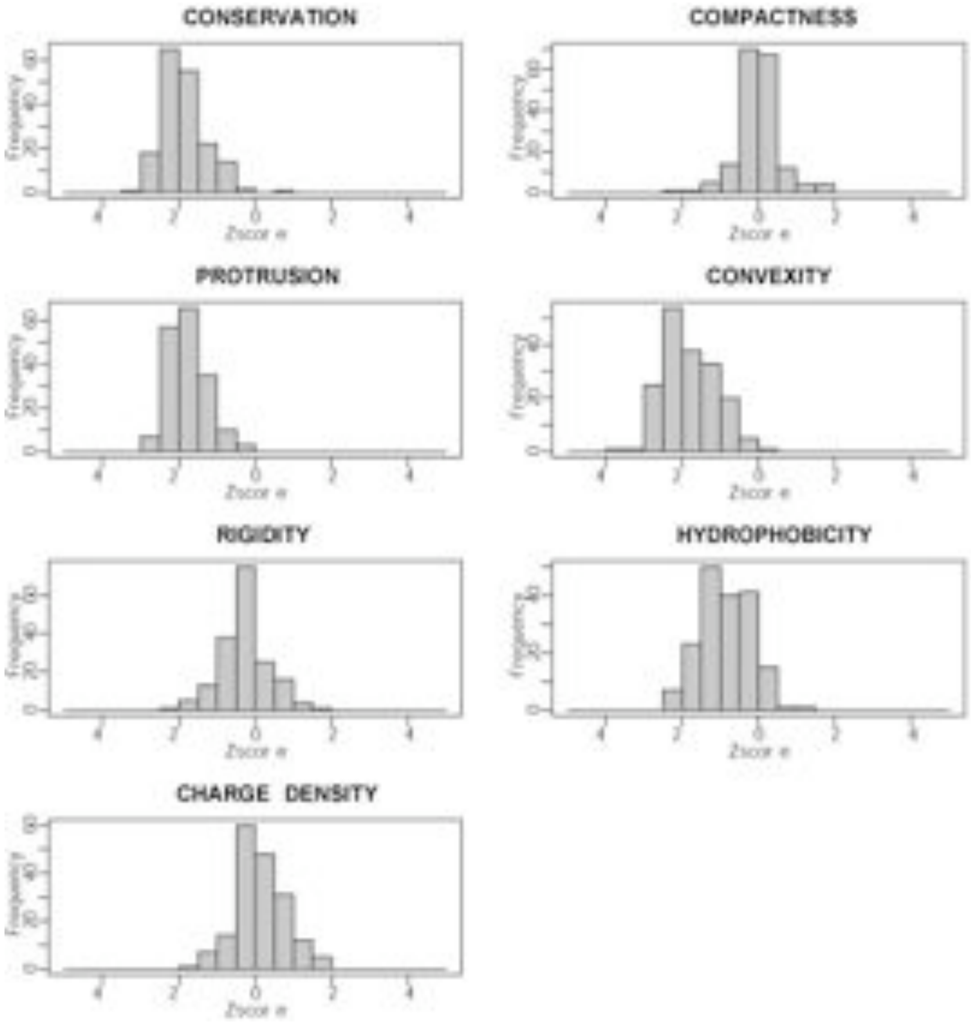
Structure conservation

Electrostatics

Solvent accessibility

Scoring

NAD



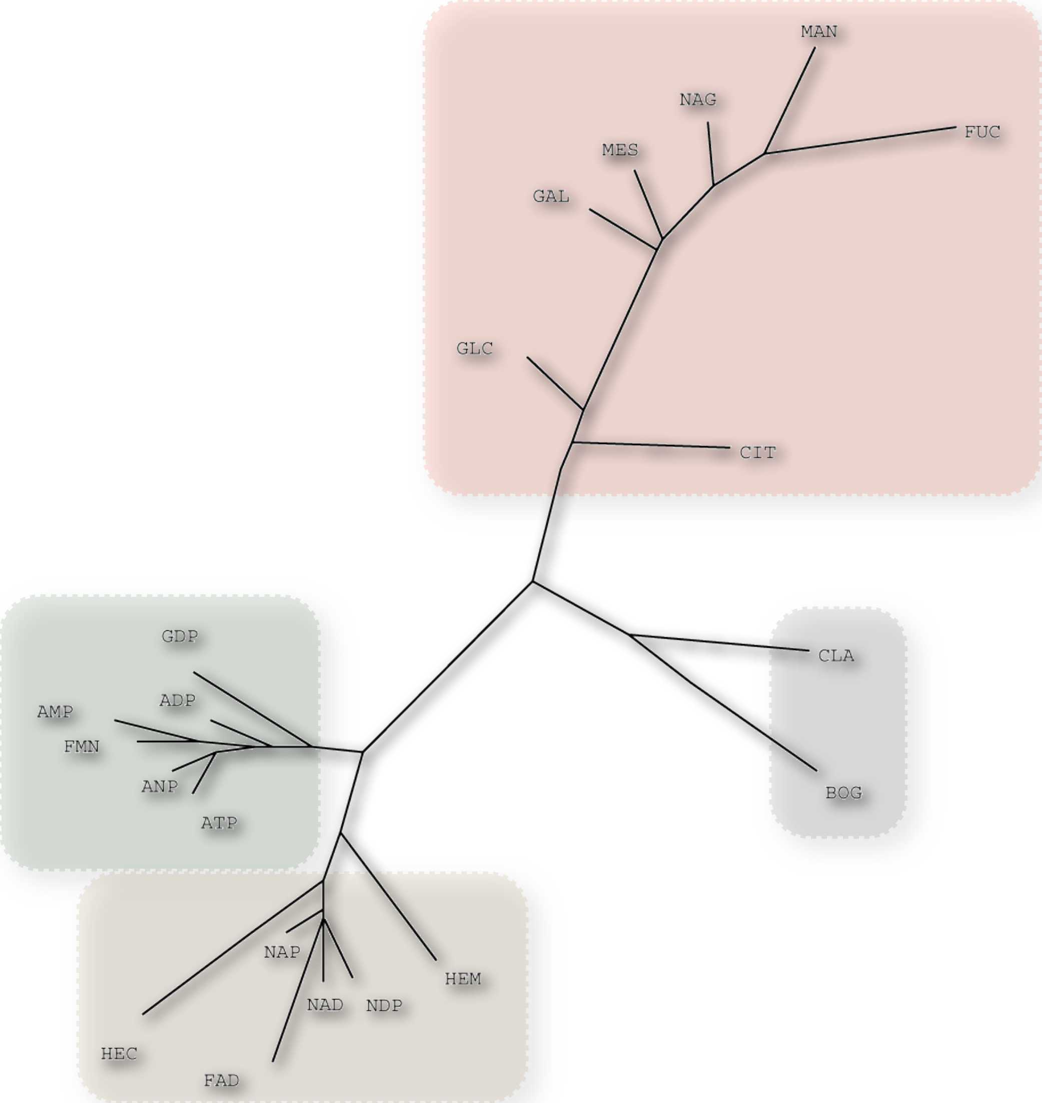
$$\rightarrow \quad w_k = \frac{1}{M}\sum_{\alpha=1}^{M} \tilde{f}_k^{(\alpha)}$$

M = number of proteins in training set

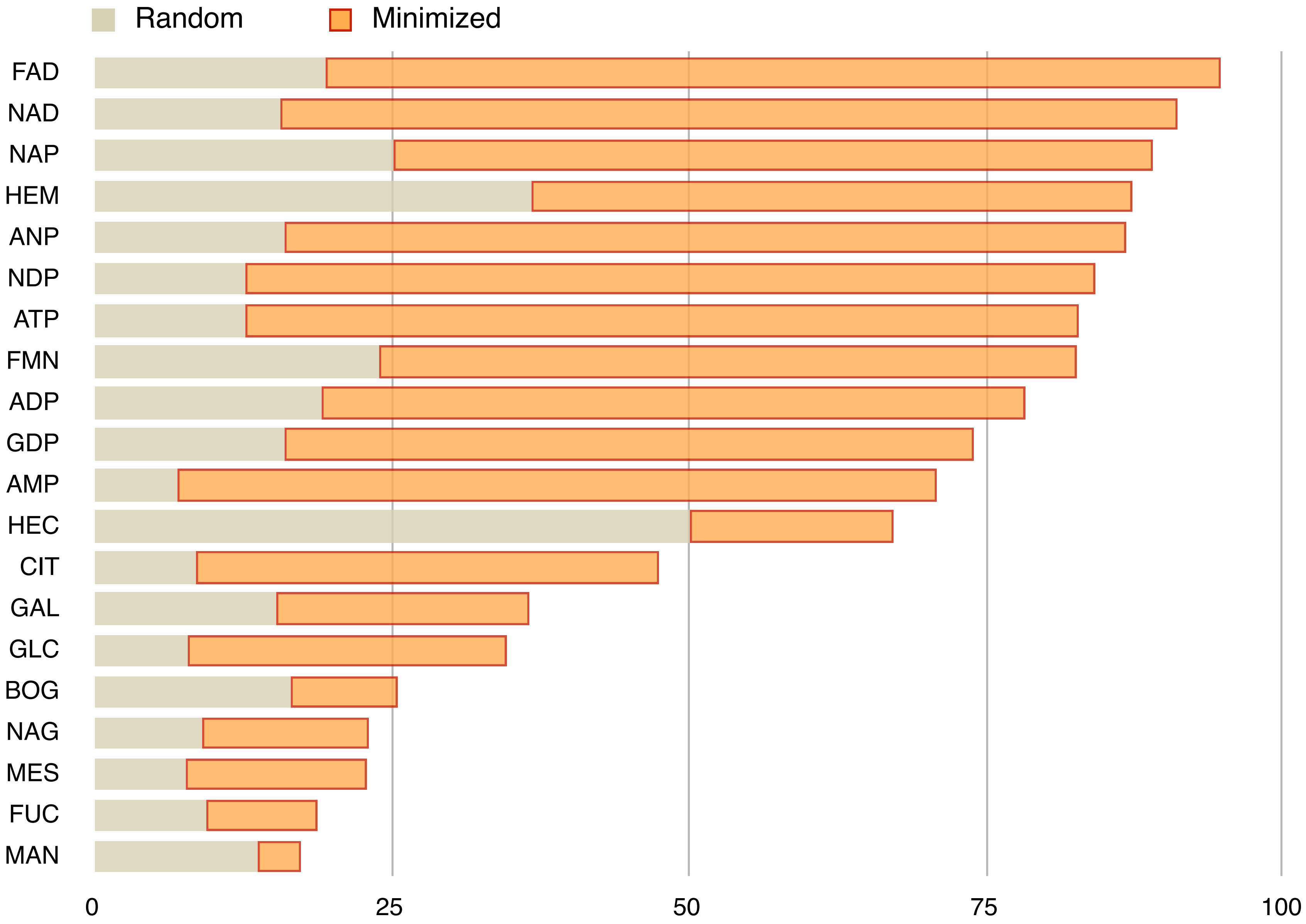
Ligand fingerprints

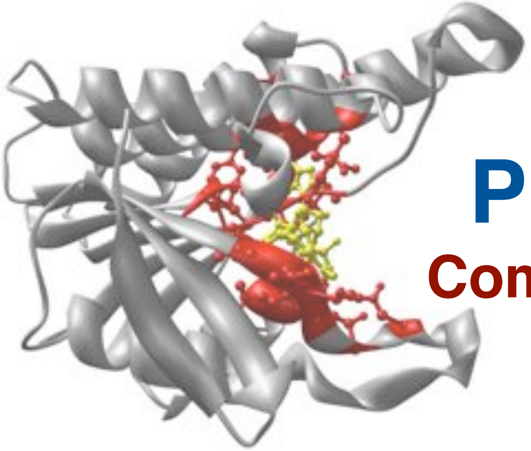
	Compactness	Conservation	Charge density	B-factor	Protrusion coefficient	Convexity score	Hydrophobicity
ADP	-1.266	-2.009	0.447	-0.414	-1.521	-1.388	-0.118
AMP	-1.62	-1.962	0.341	-0.381	-1.909	-1.944	-0.518
ANP	-1.007	-2.227	0.176	-0.392	-1.706	-1.595	-0.14
ATP	-1.122	-2.156	0.228	-0.274	-1.845	-1.768	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





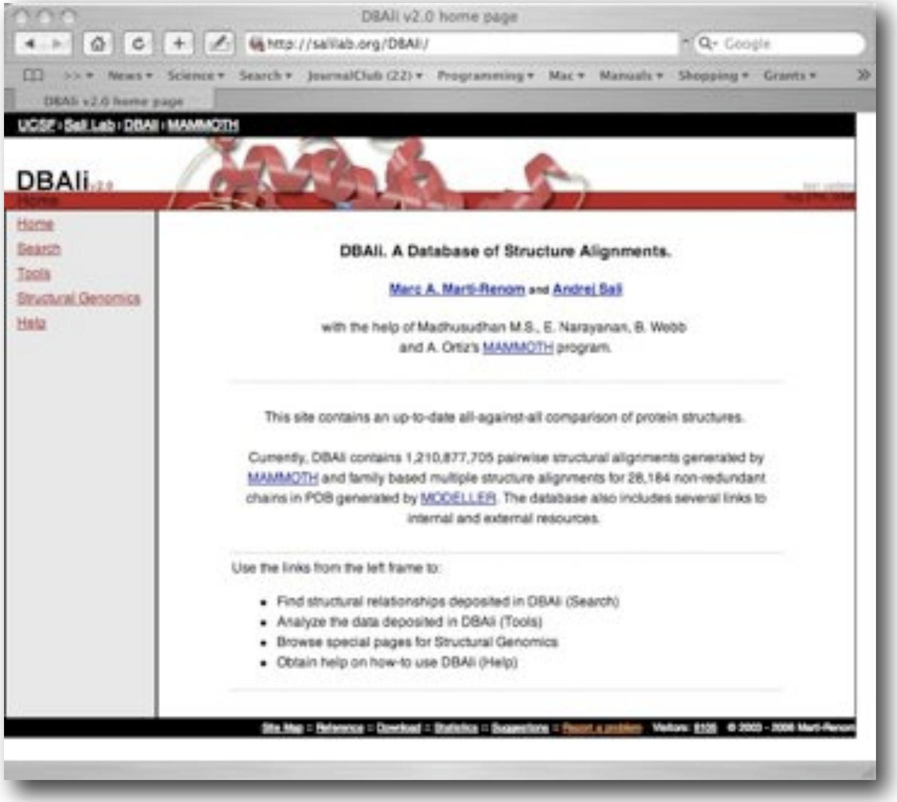
Protein function from structure

Comparative annotation. AnnoLite and AnnoLyze.

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

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

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


DBAli v2.0 home page
http://salilab.org/DBAli/
Google
News ▾ Science ▾ Search ▾ JournalClub (22) ▾ Programming ▾ Mac ▾ Manuals ▾ Shopping ▾ Grants ▾
DBAli v2.0 home page
UCSF › Sali Lab › DBAli › MAMMOTH
DBAli v2.0
Home
Home
Search
Tools
Structural Genomics
Help
Site Map :: Reference :: Download :: Statistics :: Suggestions :: Report a problem Visitors: 8105 © 2003 - 2006 Marti-Renom

DBAli. A Database of Structure Alignments.

Marc A. Marti-Renom and Andrej Sali

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

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

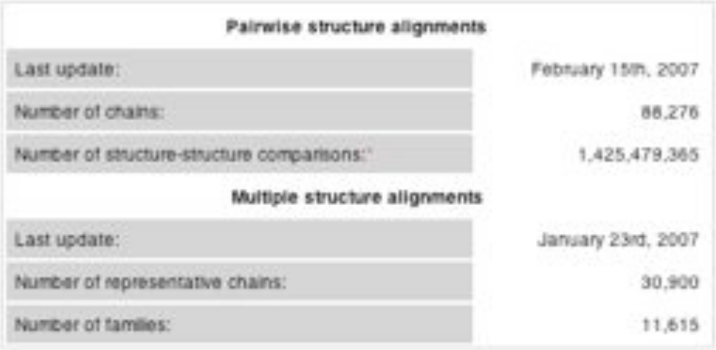
Currently, DBAli contains 1,210,877,705 pairwise structural alignments generated by MAMMOTH and family based multiple structure alignments for 28,184 non-redundant chains in PDB generated by MODELLER. The database also includes several links to internal and external resources.

Use the links from the left frame to:

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

✓ 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**

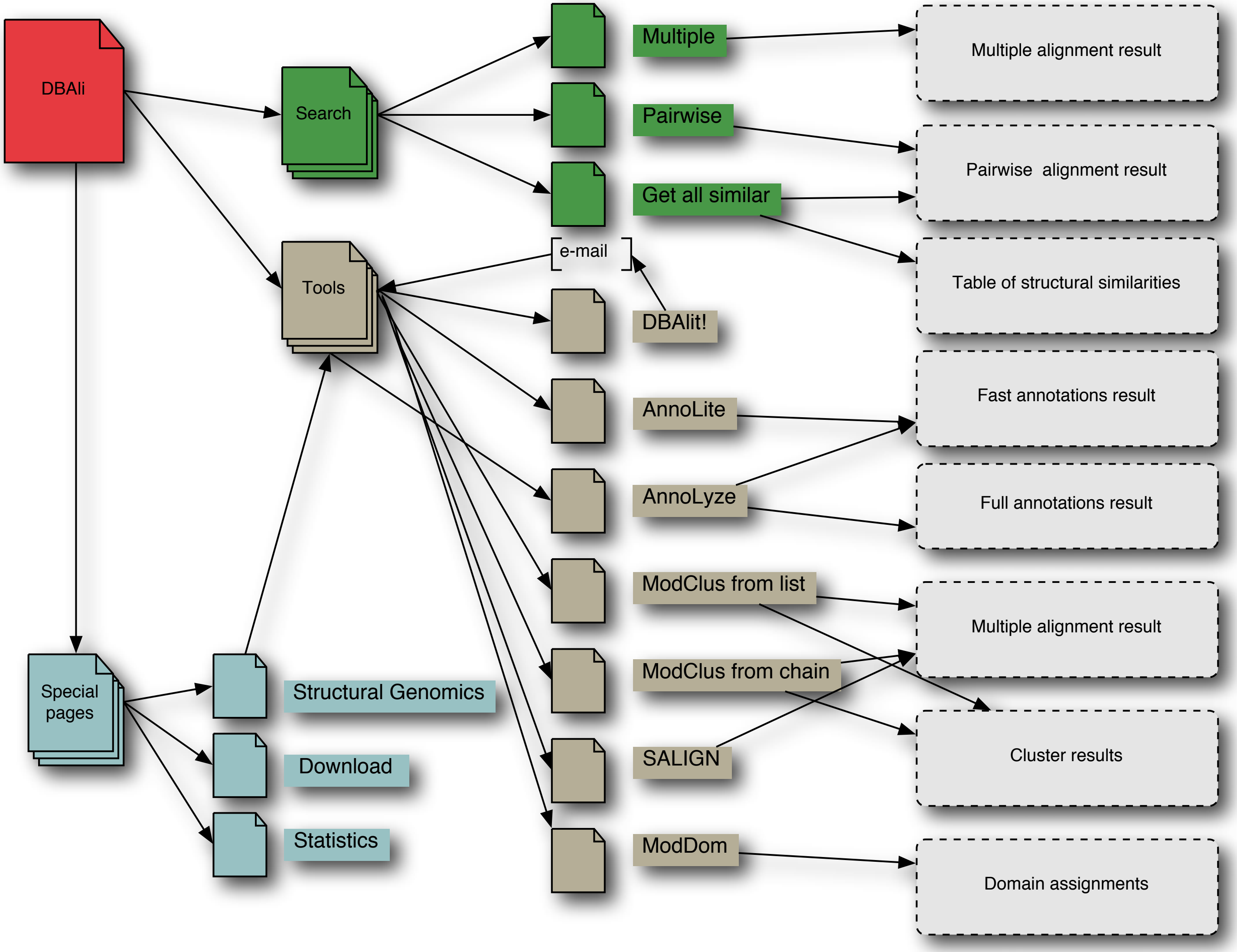
Ortiz AR, (2002) Protein Sci. 11 pp2606

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

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

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

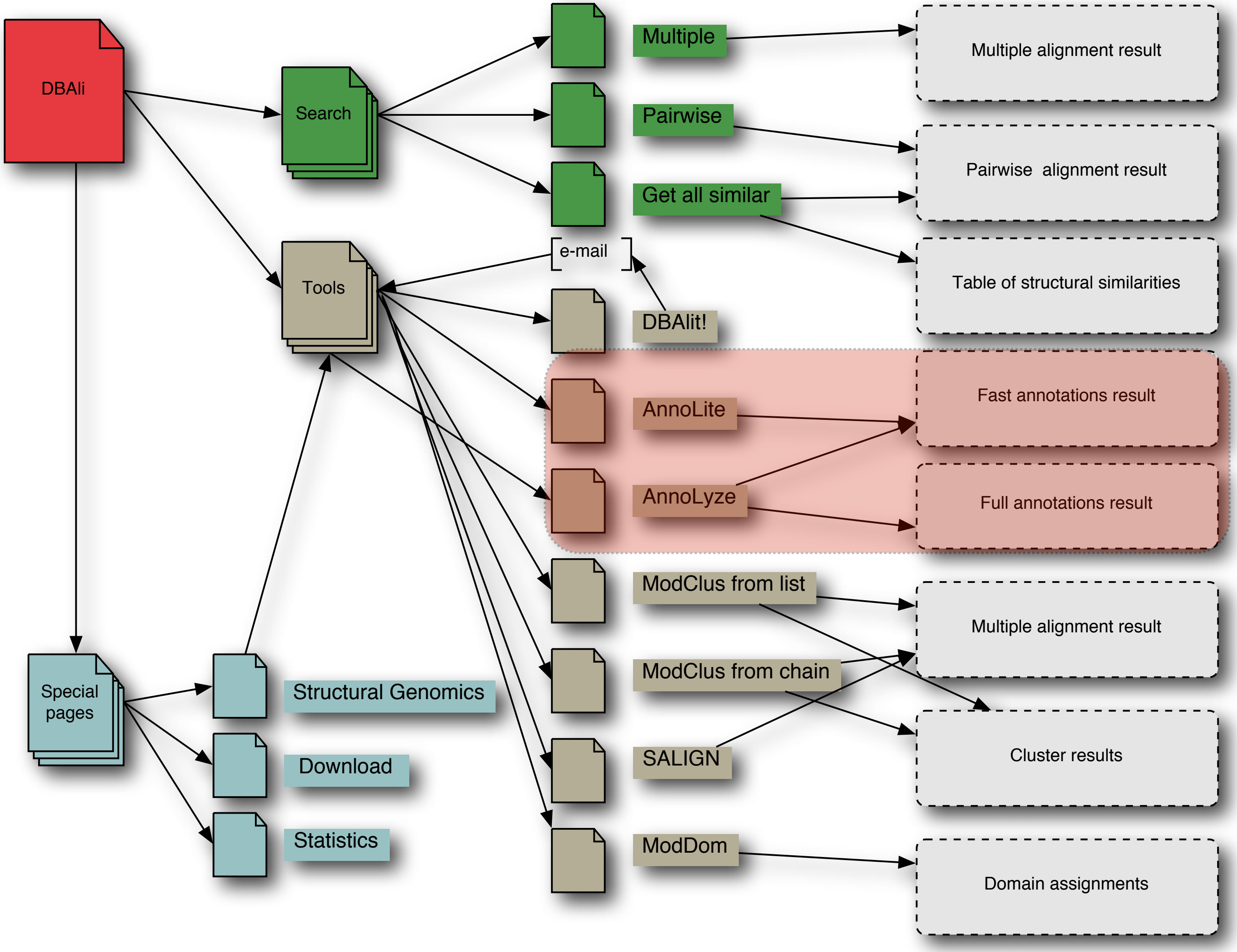
http://www.salilab.org/DBAli/



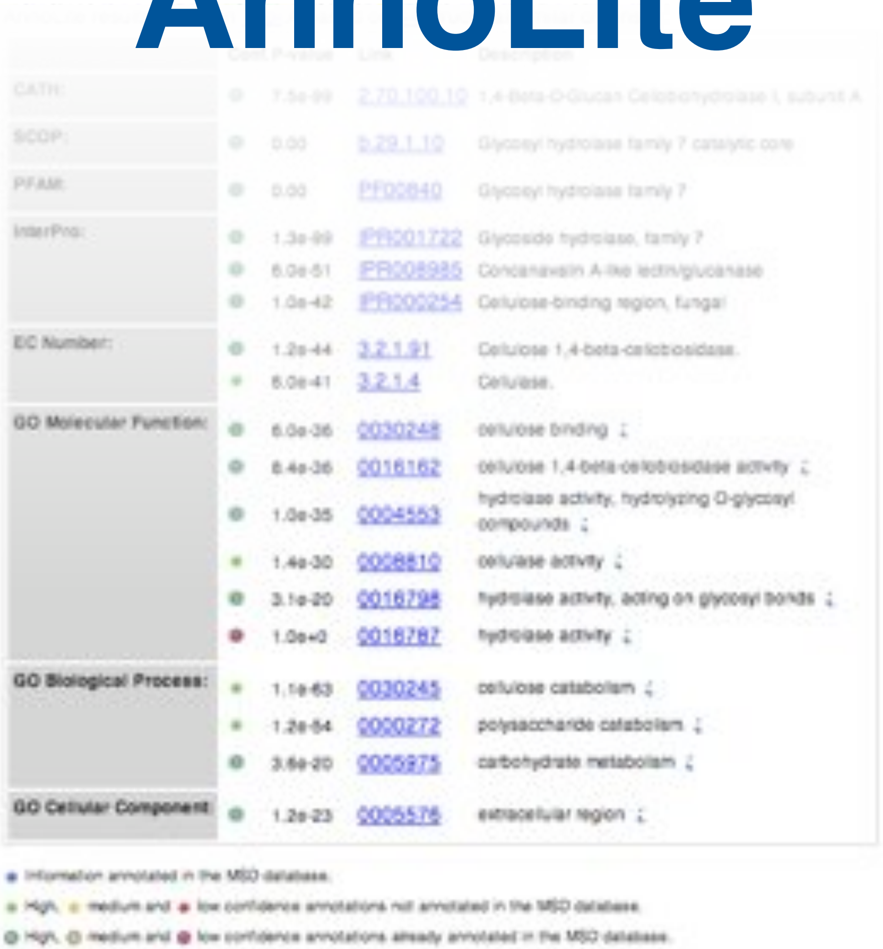
DBAli$_{v2.0}$ database

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

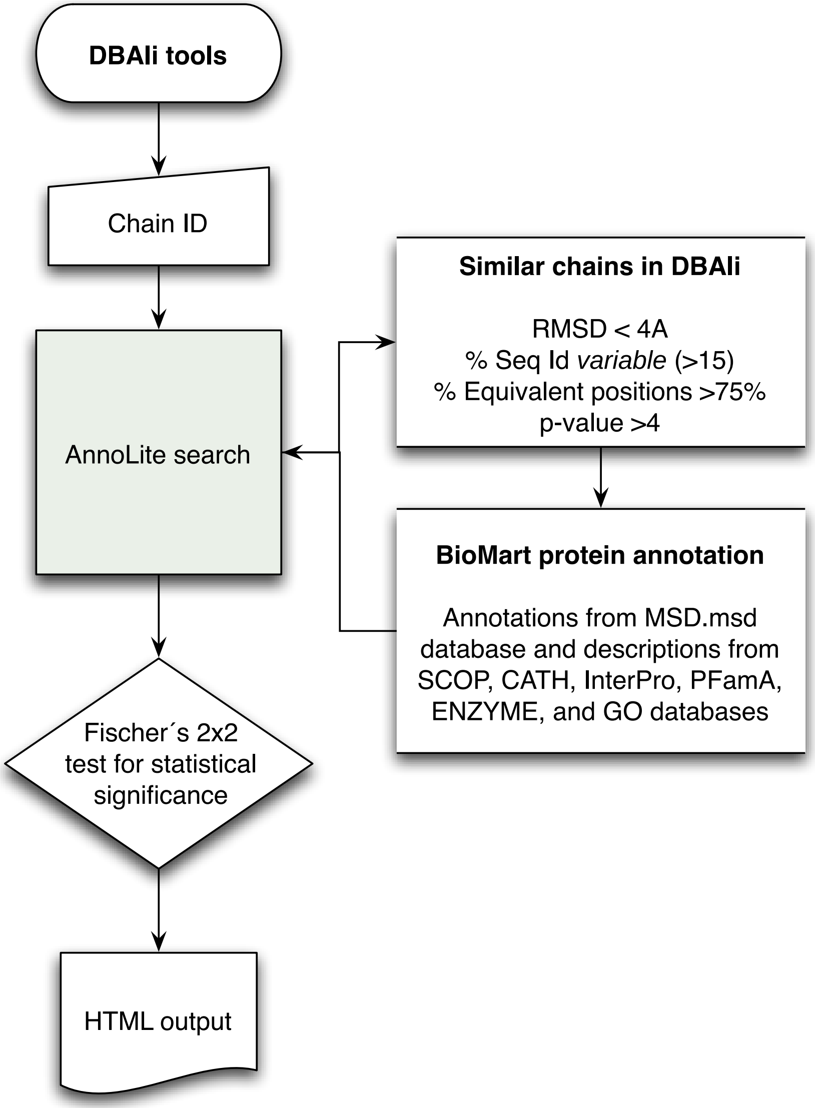
● High, ● medium and ● low confidence annotations already annotated in the MSD database.

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

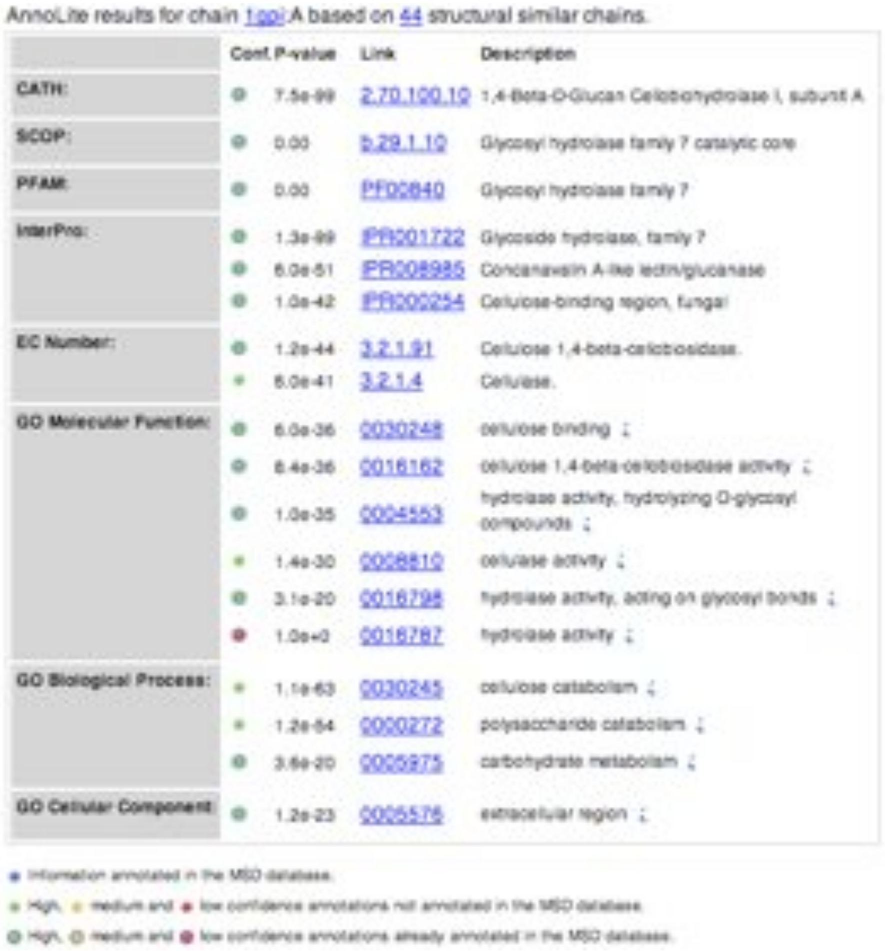


Similar chains in DBAli

RMSD < 4A
% Seq Id *variable* (>15)
% Equivalent positions >75%
p-value >4

BioMart protein annotation

Annotations from MSD.msd database and descriptions from SCOP, CATH, InterPro, PFamA, ENZYME, and GO databases



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
	●	6.4e-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

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

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

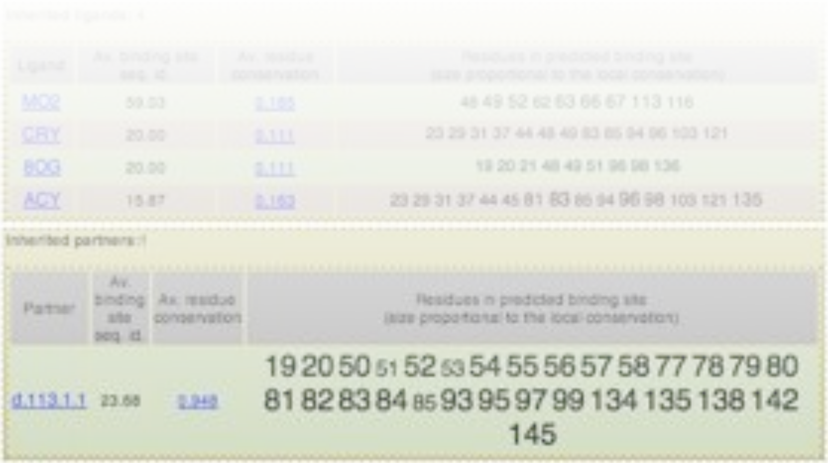
$$p = 1.78e^{-19}$$

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

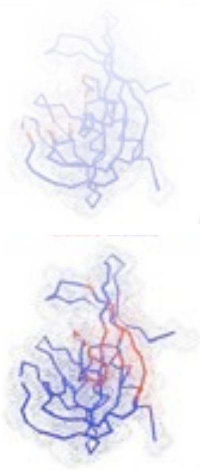


Inherited ligands: 4

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

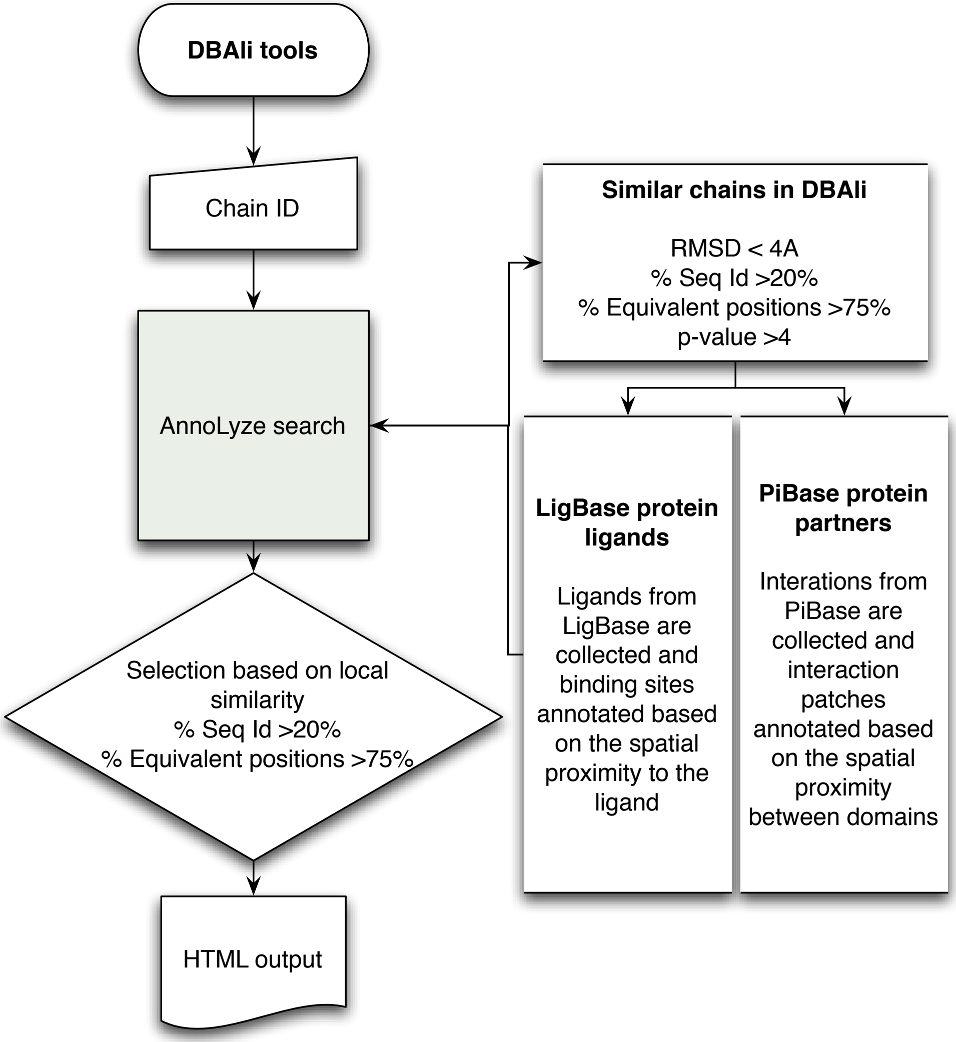
	Number of chains
Initial set*	78,167
LigBase**	30,126
Non-redundant set***	4,948 (8,846 ligands)

*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


DBAli tools

Chain ID

AnnoLyze search

Selection based on local similarity
% Seq Id >20%
% Equivalent positions >75%

HTML output

Similar chains in DBAli

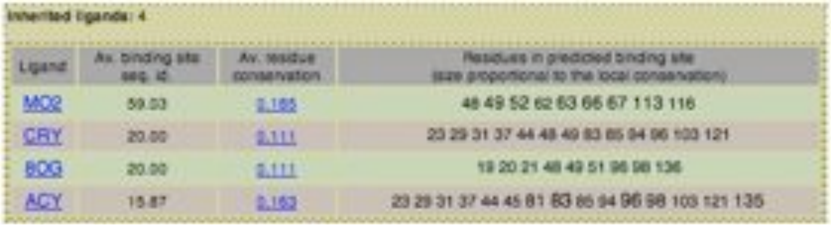
RMSD < 4A
% Seq Id >20%
% Equivalent positions >75%
p-value >4

LigBase protein ligands

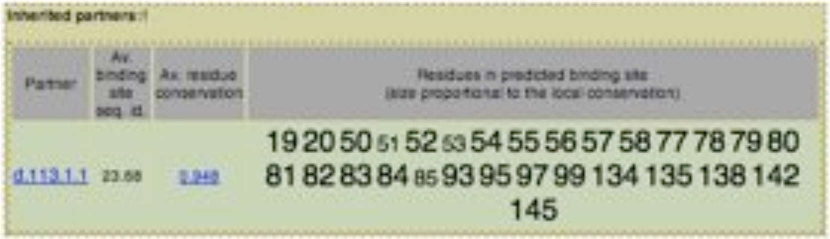
Ligands from LigBase are collected and binding sites annotated based on the spatial proximity to the ligand

PiBase protein partners

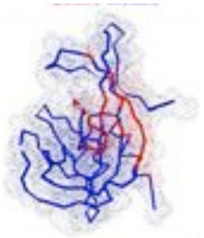
Interations from PiBase are collected and interaction patches annotated based on the spatial proximity between domains


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

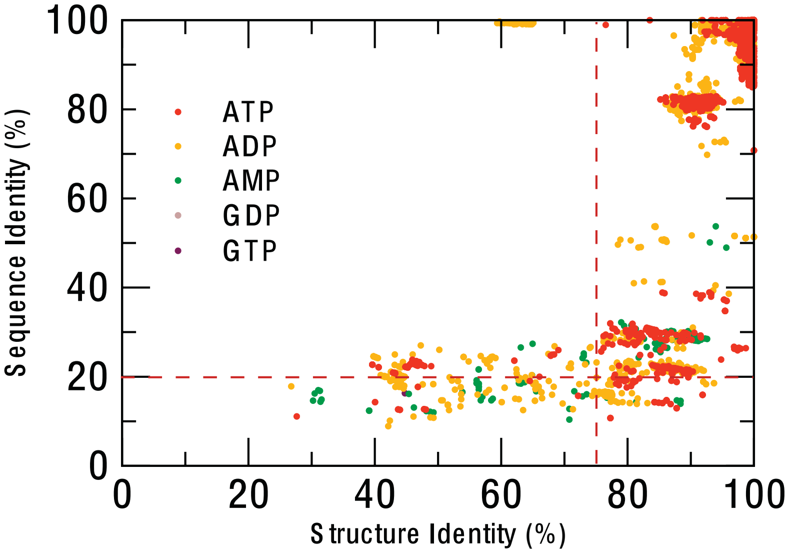

Inherited partners: 1

Partner	Av. binding site seq. id	Av. residue conservation	Residues n 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

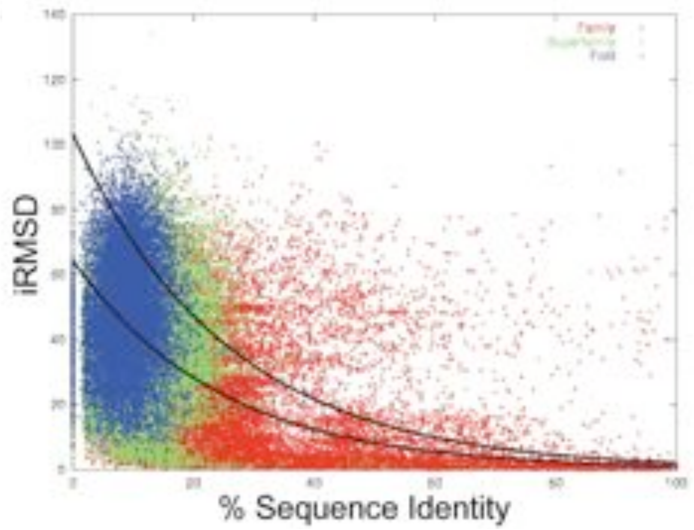


Scoring function

Ligands



Partners



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

Sensitivity .vs. Precision

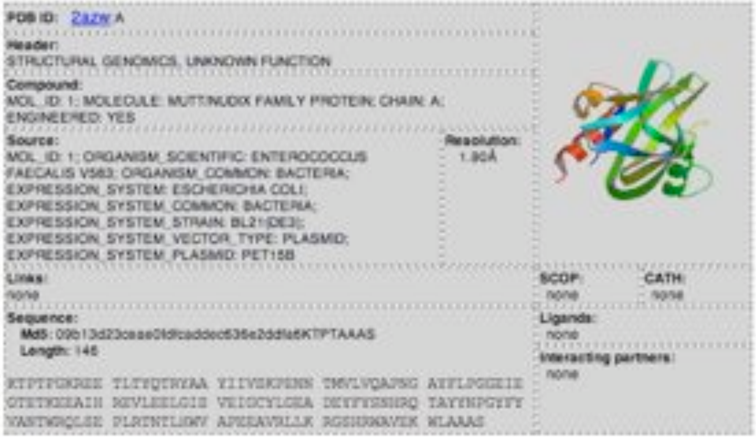
	Optimal cut-off	Sensitivity (%) Recall or TPR	Precision (%)
Ligands	30%	71.9	13.7
Partners	40%	72.9	55.7

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

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

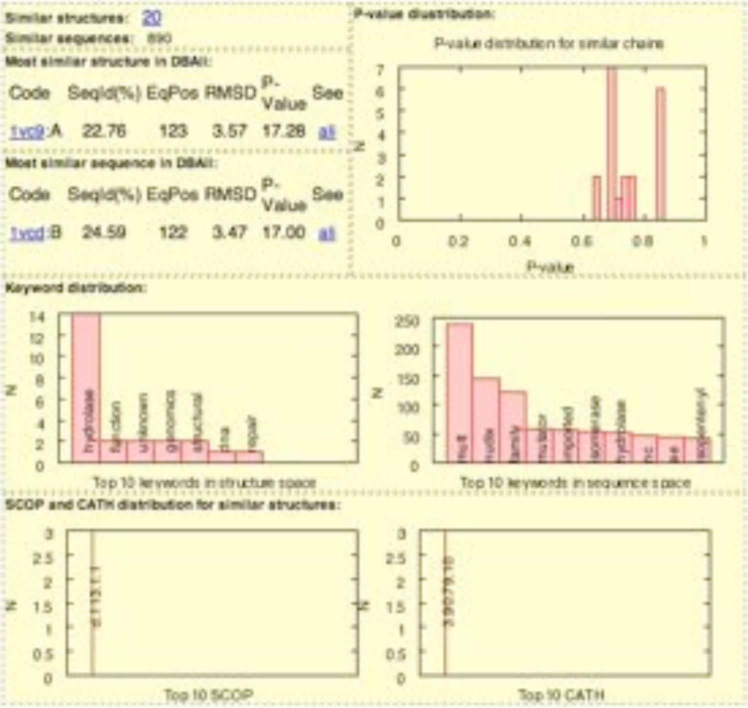
CATH:
none

Sequence:
Md5 : 09b13d23caae0fdfcaddec636e2dd1a6KTPTAAAS
Length: 146

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

Ligands:
none

Interacting partners:
none



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	all

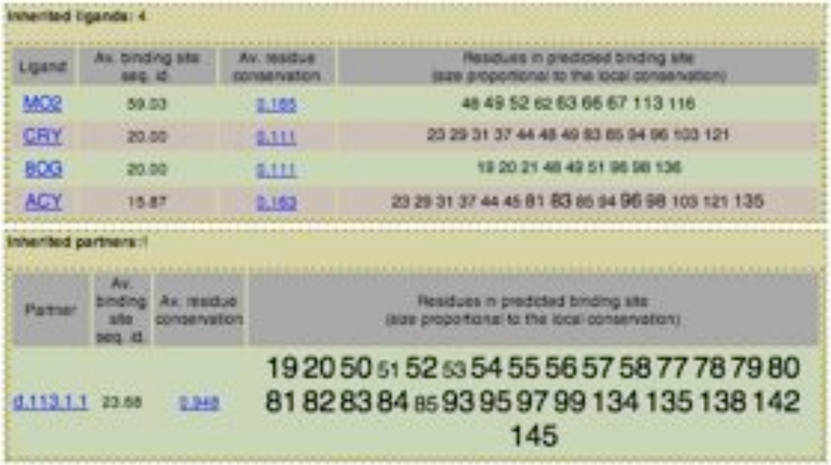
Most similar sequence in DBAli:

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

P-value distribution:

Keyword distribution:

SCOP and CATH distribution for similar structures:

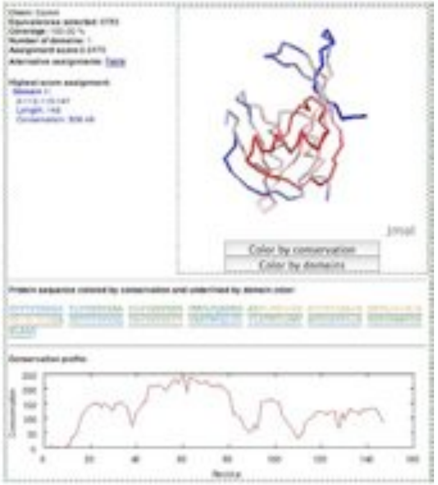


Inherited ligands: 4

Ligand	Av. binding site seq. id	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
MO2	59.03	0.165	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	0.949	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

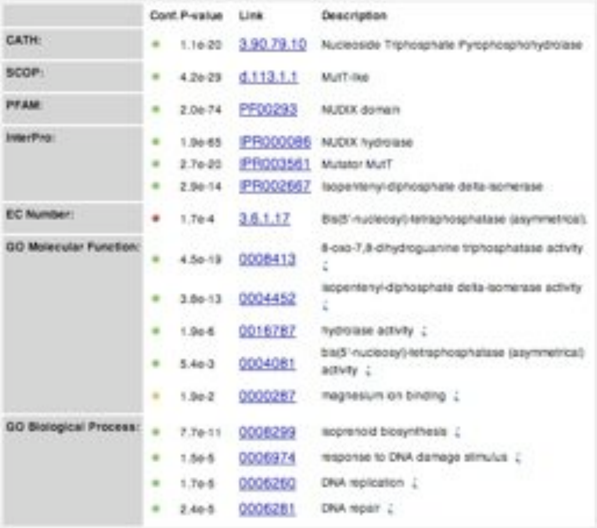


Chain: [illegible]
Equivalences selected: [illegible]
Coverage: 100.00 %
Number of domains: 1
Assignment score 0.[illegible]
Alternative assignments: here

Highest score assignment:
Domain 1:
d.113.1.1 [illegible]
Length: 146
Conservation: [illegible]

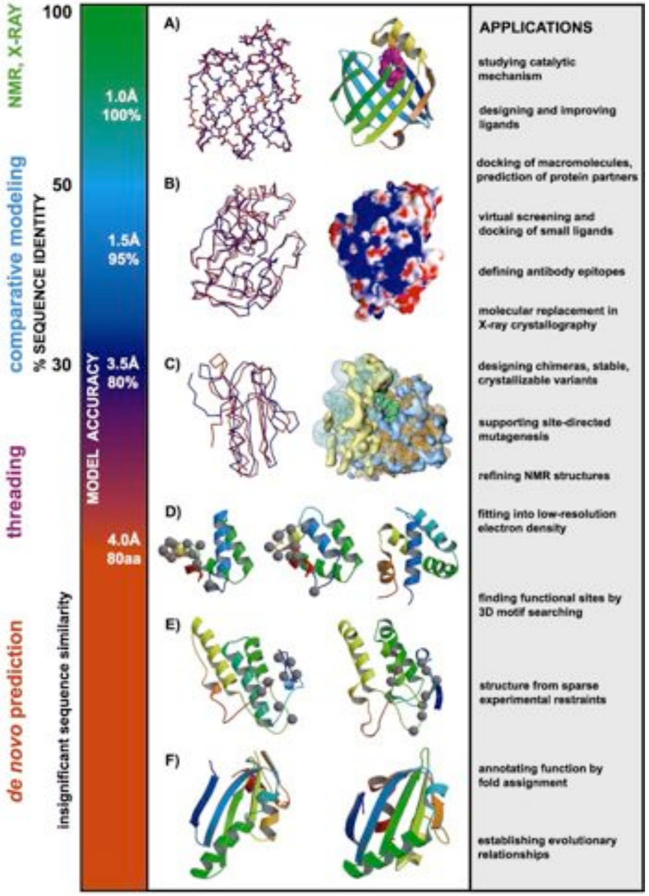
Protein sequence colored by conservation and underlined by domain color:

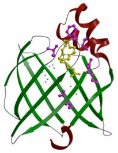
Conservation profile:

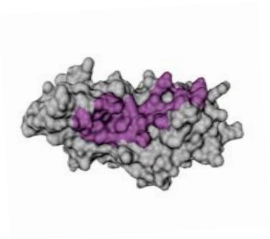


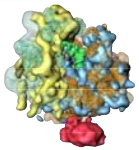
	Conf.	P-value	Link	Description
CATH:	●	1.1e-20	3.90.79.10	Nucleoside Triphosphate Pyrophosphohydrolase
SCOP:	●	4.2e-29	d.113.1.1	MutT-like
PFAM:	●	2.0e-74	PF00293	NUDIX domain
InterPro:	●	1.9e-65	IPR000086	NUDIX hydrolase
	●	2.7e-20	IPR003561	Mutator MutT
	●	2.9e-14	IPR002667	Isopentenyl-diphosphate 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 ¿

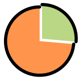
Can we use models to infer function?


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

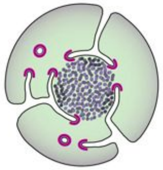








T. cruzi



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

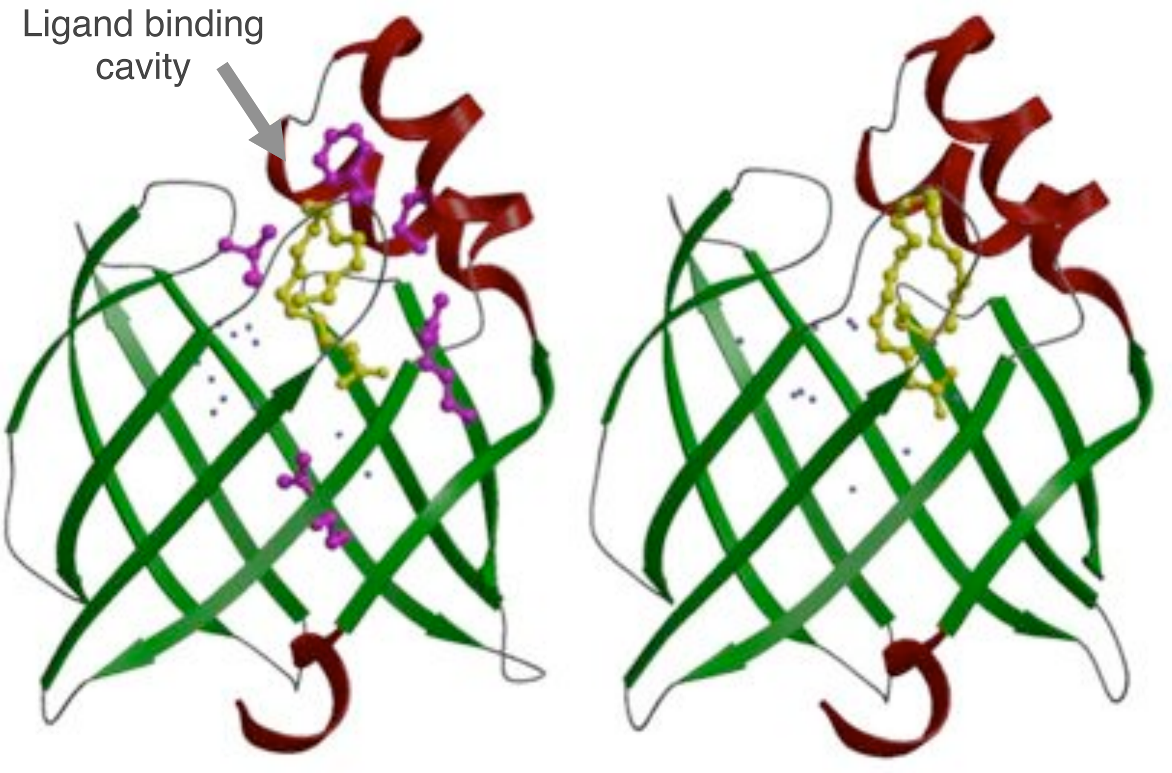
Predicting features of a model that are not present in the template

BLBP/oleic acid

Cavity is not filled

BLBP/docosahexaenoic acid

Cavity is filled



Ligand binding cavity

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

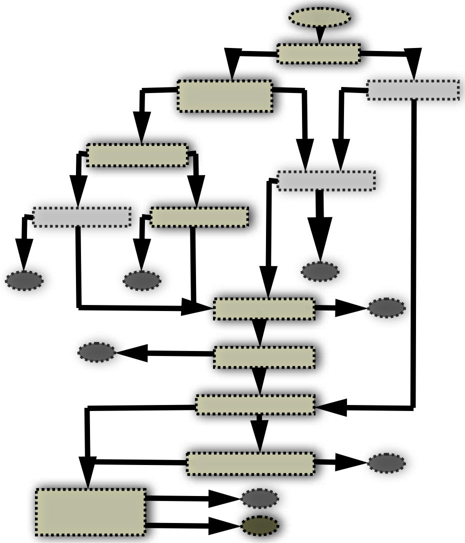
Nebojsa Mirkovic, Marc A. Marti-Renom, Barbara L. Weber, Andrej Sali and Alvaro N.A. Monteiro

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

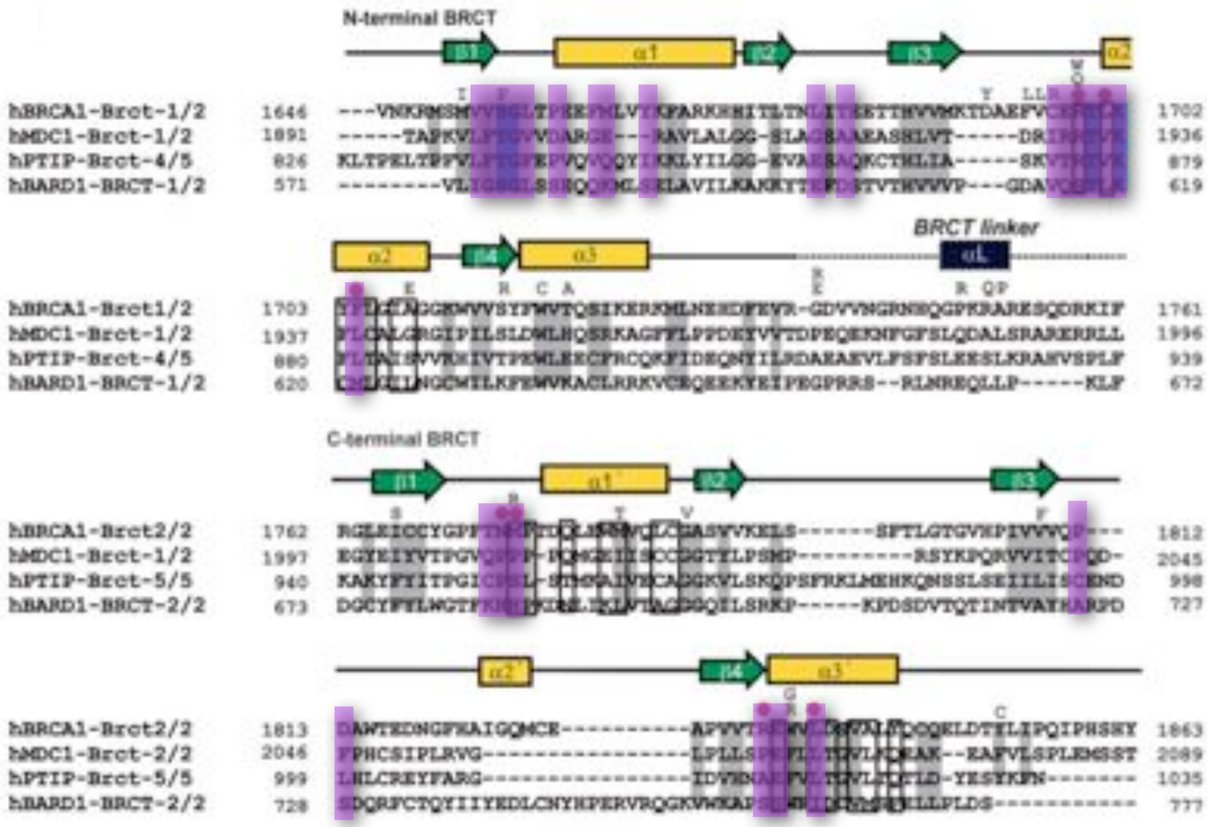
Cannot measure the functional impact of every possible SNP at all positions in each protein! Thus, prediction based on general principles of protein structure is needed.

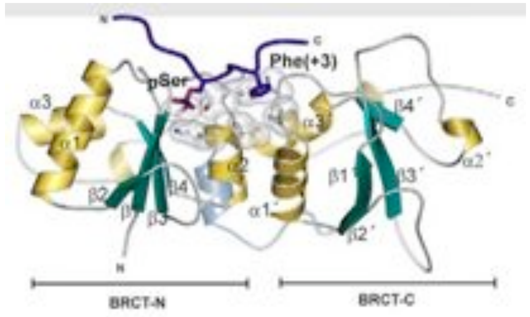
Missense mutations in BRCT domains by function

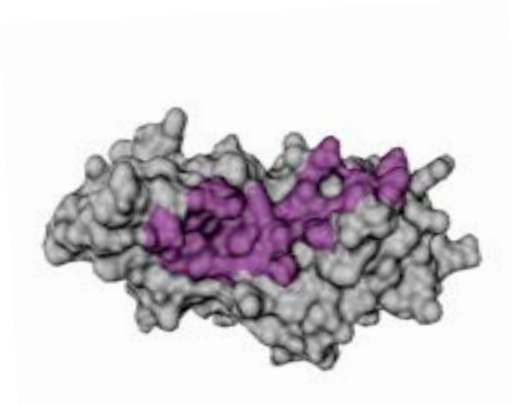
	cancer associate	not cancer associated	?
no transcription activation	C1697R R1699W A1708E S1715R P1749R M1775R		M1652K L1657P E1660G H1686Q R1699Q K1702E Y1703HF 1704S L1705PS 1715NS1 722FF17 34LG173 8EG1743 RA1752 PF1761I F1761S M1775E M1775K L1780P I1807S V1833E A1843T
transcription activation		M1652I 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





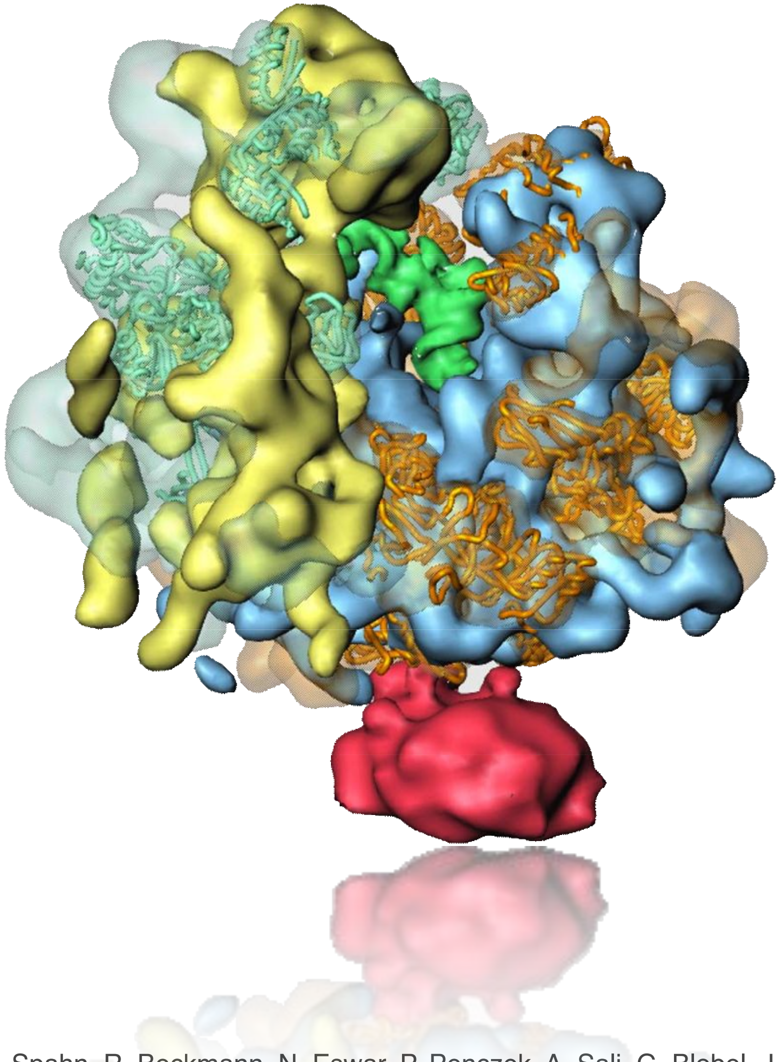


Putative binding site predicted in 2003
and accepted for publication on March 2004.

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

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

S. cerevisiae ribosome



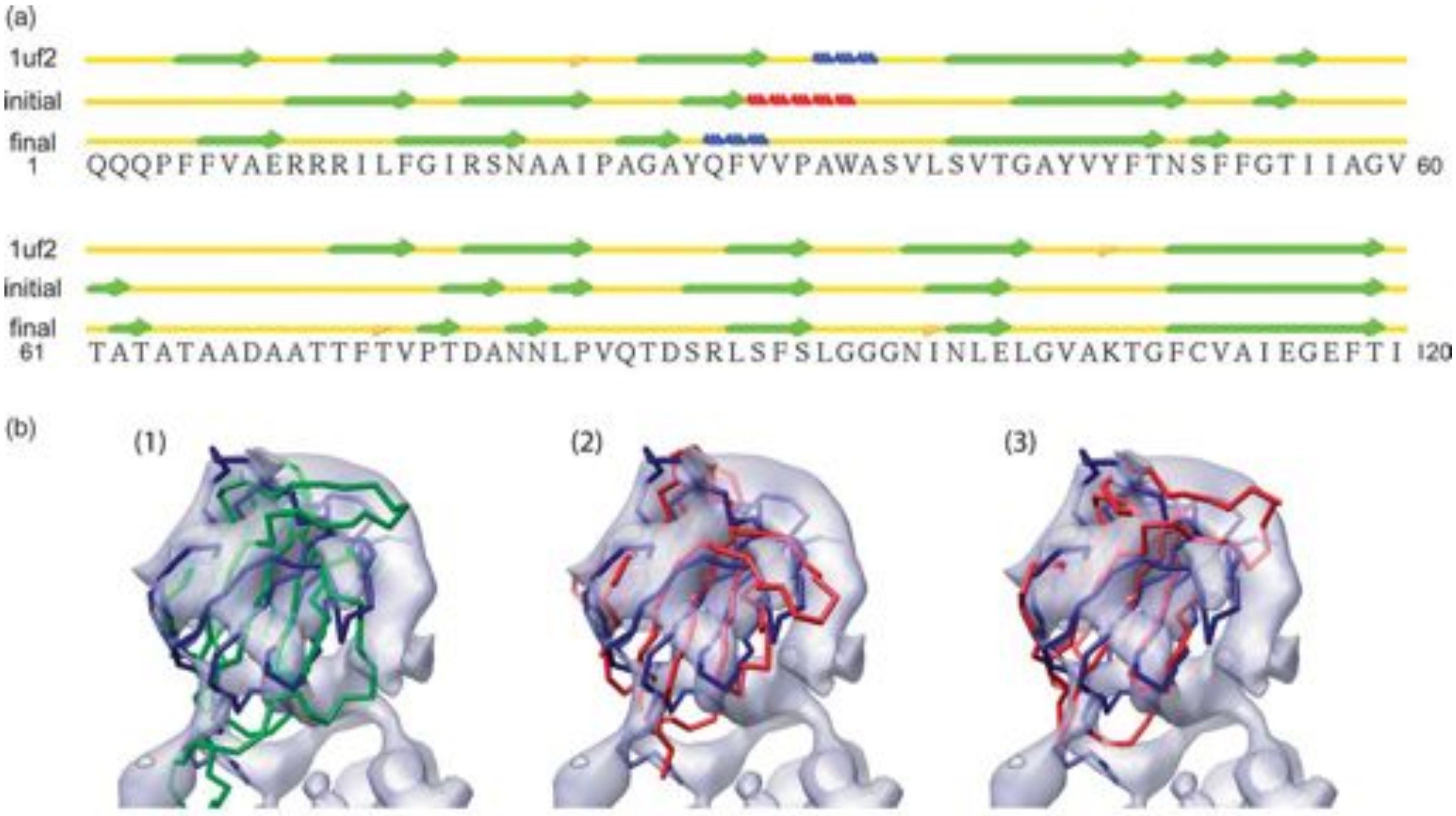
Fitting of comparative models into 15Å cryo-electron density map.

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.

Modeling & cryoEM


(a)
1uf2
initial
final
1
QQQPFFVAERRRILFGIRSNAAIPAGAYQFVVPAWASVLSVTGAYVYFTNSFFGTIIAGV 60
1uf2
initial
final
61
TATATAADAATTFTVPTDANNLPVQTDSRLSFSLGGGNINLELGVAKTGFCVAIEGEFTI 120
(b)
(1)
(2)
(3)

Template

NO = 70.8%

$Z_c = 1.2$

Structure-based alignment

NO = 84.2%

$Z_c = 1.0$

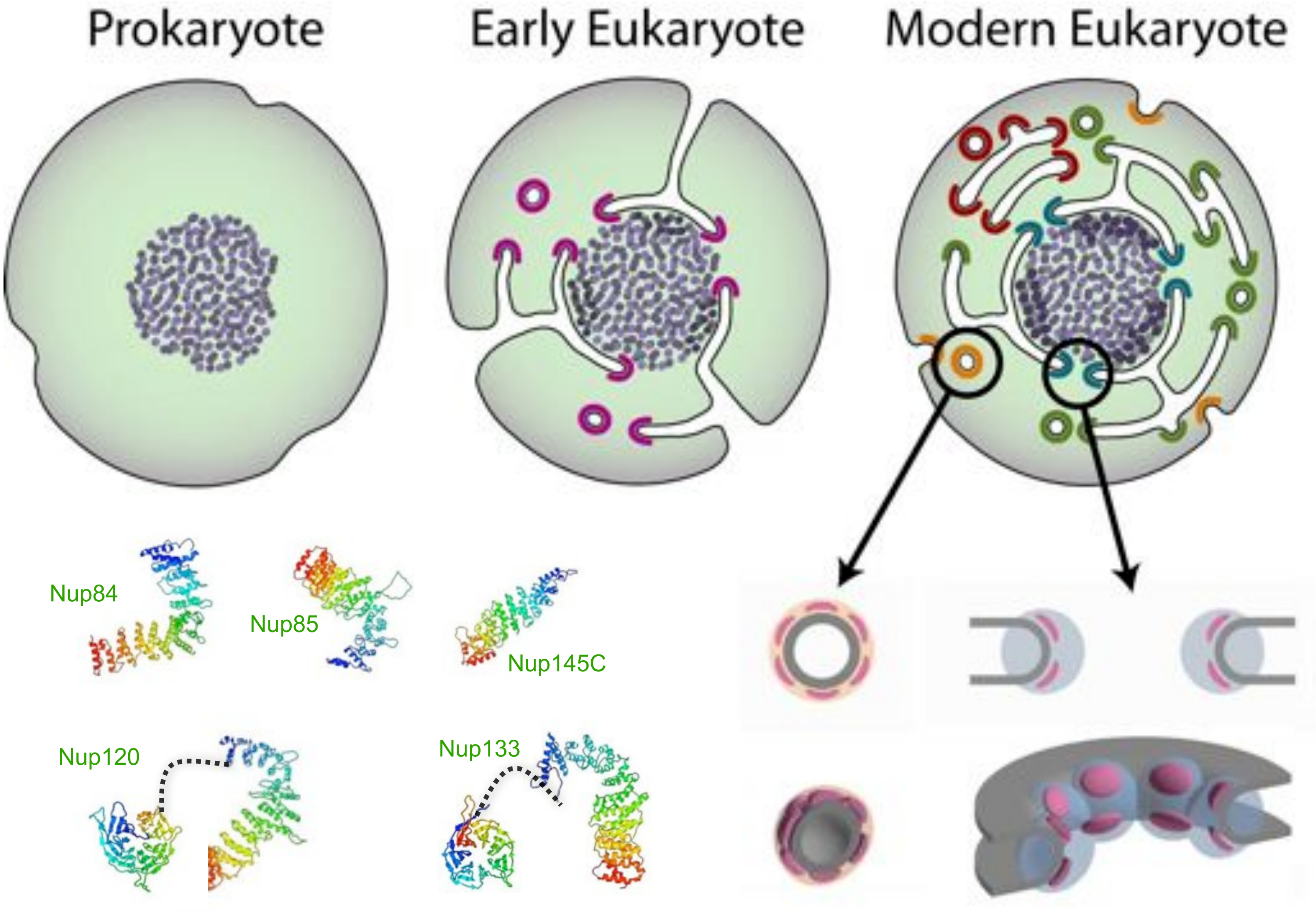
Final model

NO = 53.3%

$Z_c = 1.4$

Topf etal. JMB, 357, 1655 (2006)

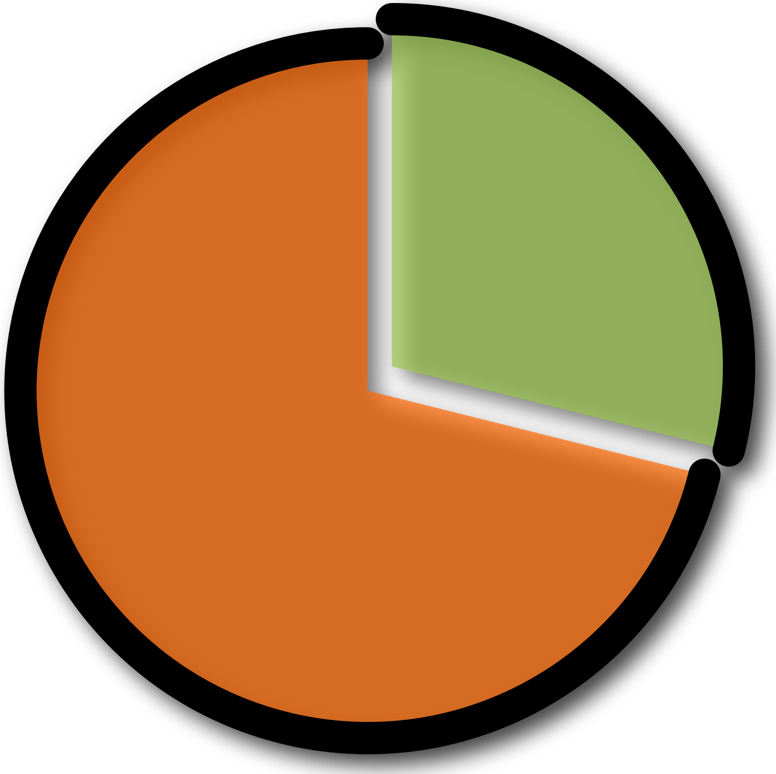
The Nucleopore complex
Cell evolution (?)



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

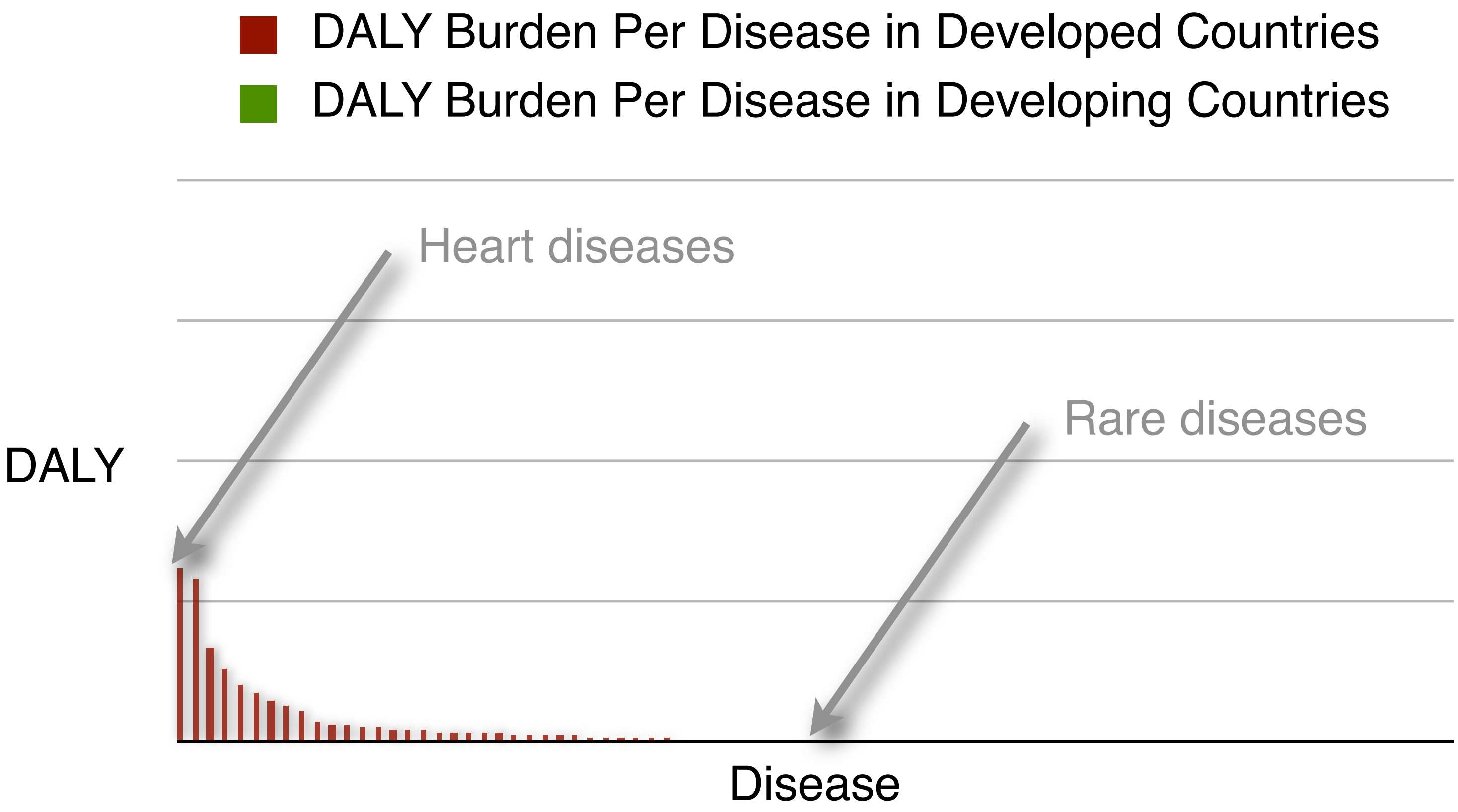
Tropical Disease Initiative (TDI)

Predicting binding sites in protein structure models.



http://www.tropicaldisease.org

Need is High in the Tail



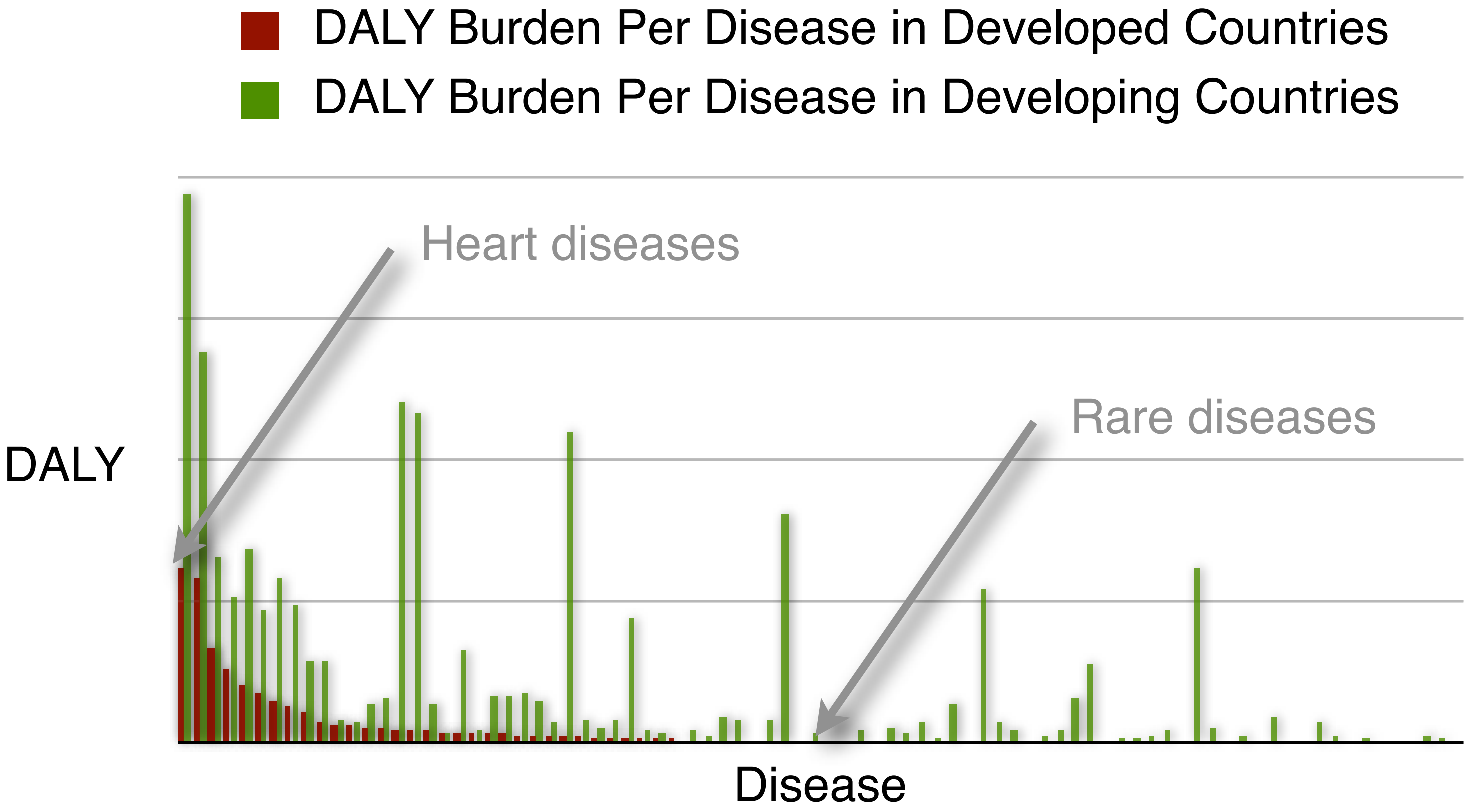
Disease data taken from WHO, *World Health Report 2004*

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)

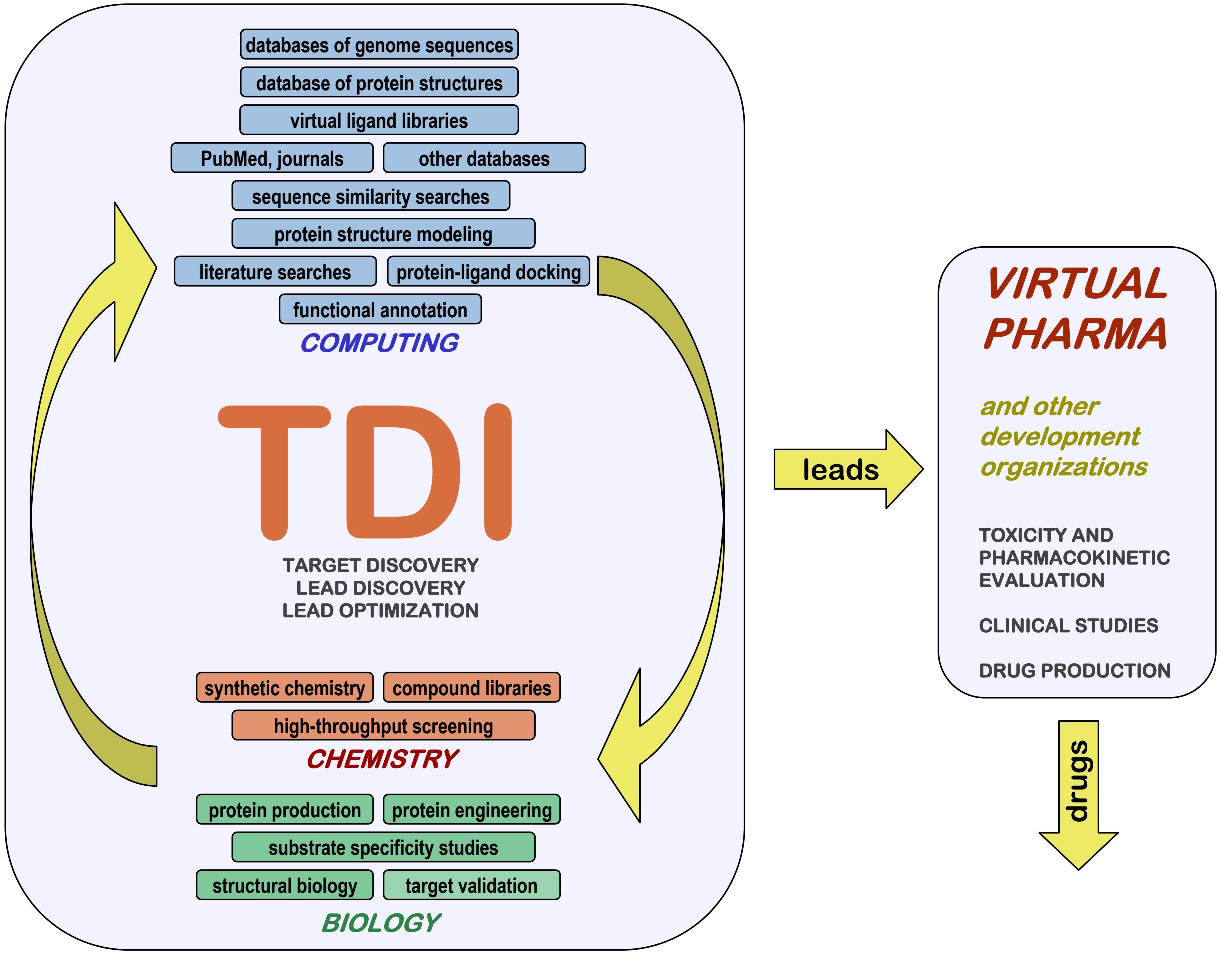
Malaria*	**46,486**
Tetanus	7,074
Lymphatic filariasis*	**5,777**
Syphilis	4,200
Trachoma	2,329
Leishmaniasis*	**2,090**
Ascariasis	1,817
Schistosomiasis*	**1,702**
Trypanosomiasis*	**1,525**

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.

TDI flowchart

http://www.tropicaldisease.org



databases of genome sequences

database of protein structures

virtual ligand libraries

PubMed, journals

other databases

sequence similarity searches

protein structure modeling

literature searches

protein-ligand docking

functional annotation

COMPUTING

TDI

TARGET DISCOVERY
LEAD DISCOVERY
LEAD OPTIMIZATION

synthetic chemistry

compound libraries

high-throughput screening

CHEMISTRY

protein production

protein engineering

substrate specificity studies

structural biology

target validation

BIOLOGY

leads

VIRTUAL PHARMA

and other development organizations

TOXICITY AND PHARMACOKINETIC EVALUATION

CLINICAL STUDIES

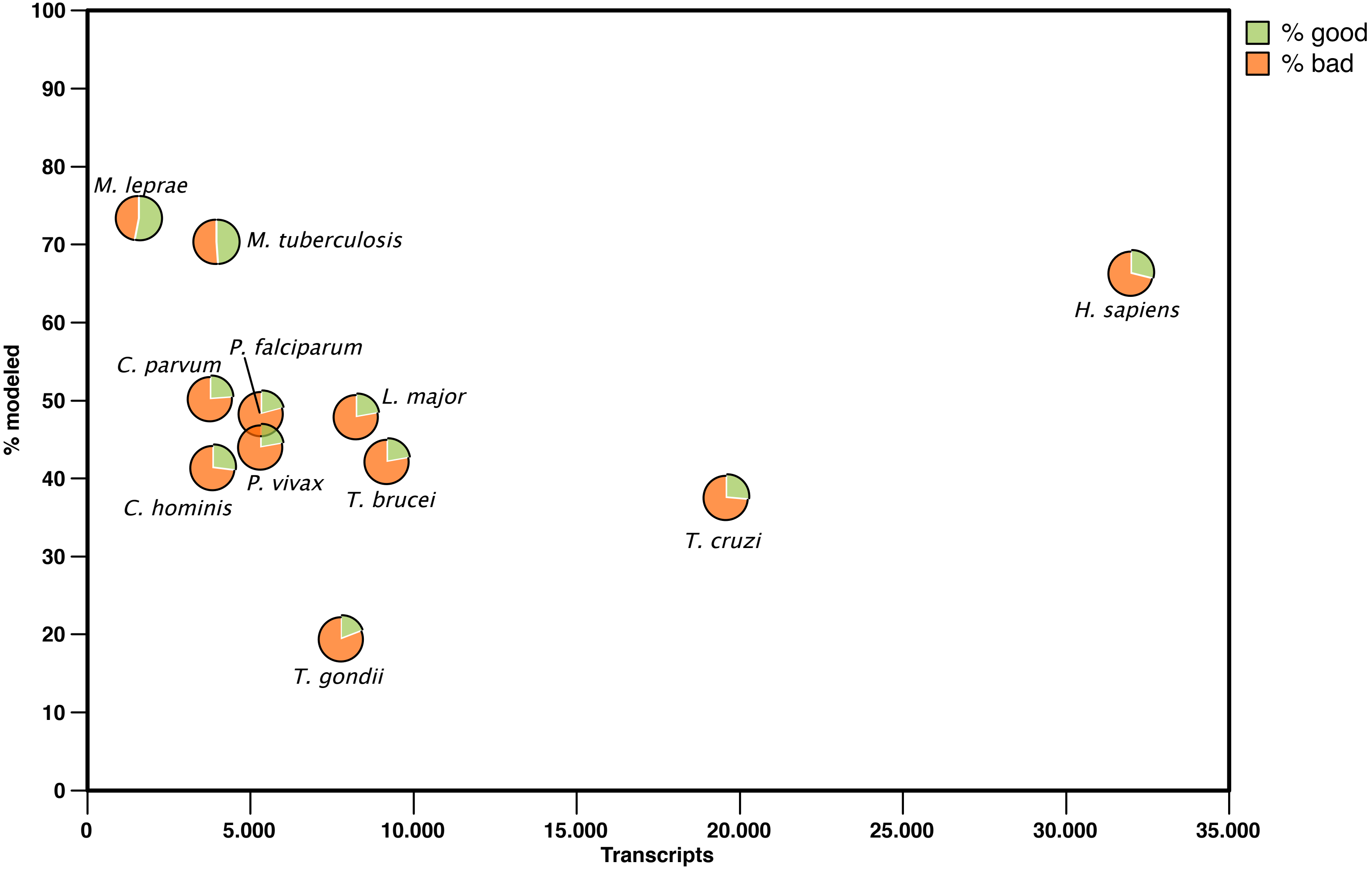
DRUG PRODUCTION

drugs

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

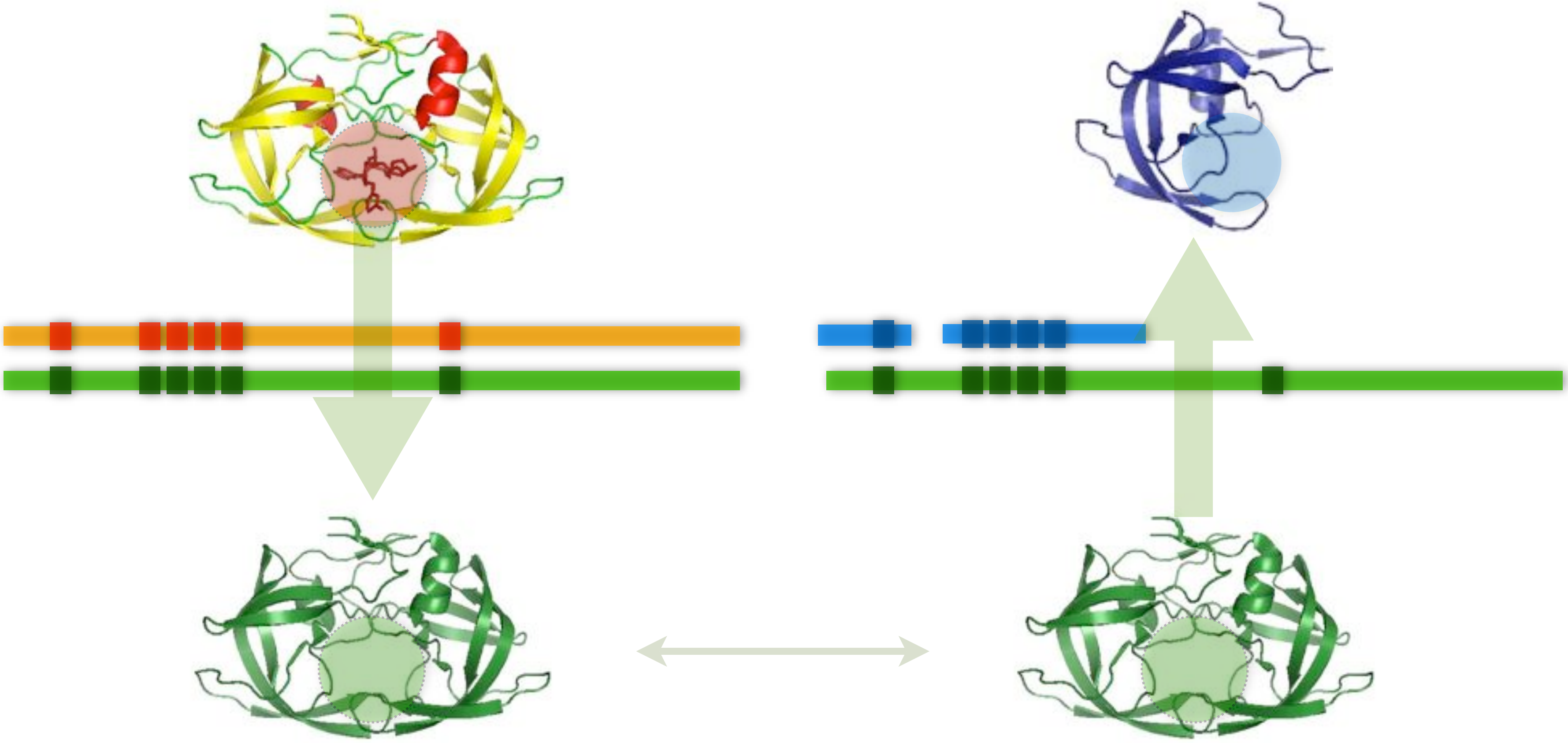
Modeling Genomes

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



A good model has MPQS of 1.1 or higher

Comparative docking



1. Expansion

co-crystalized protein/ligand

crystalized protein

2. Inheritance

model

template

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

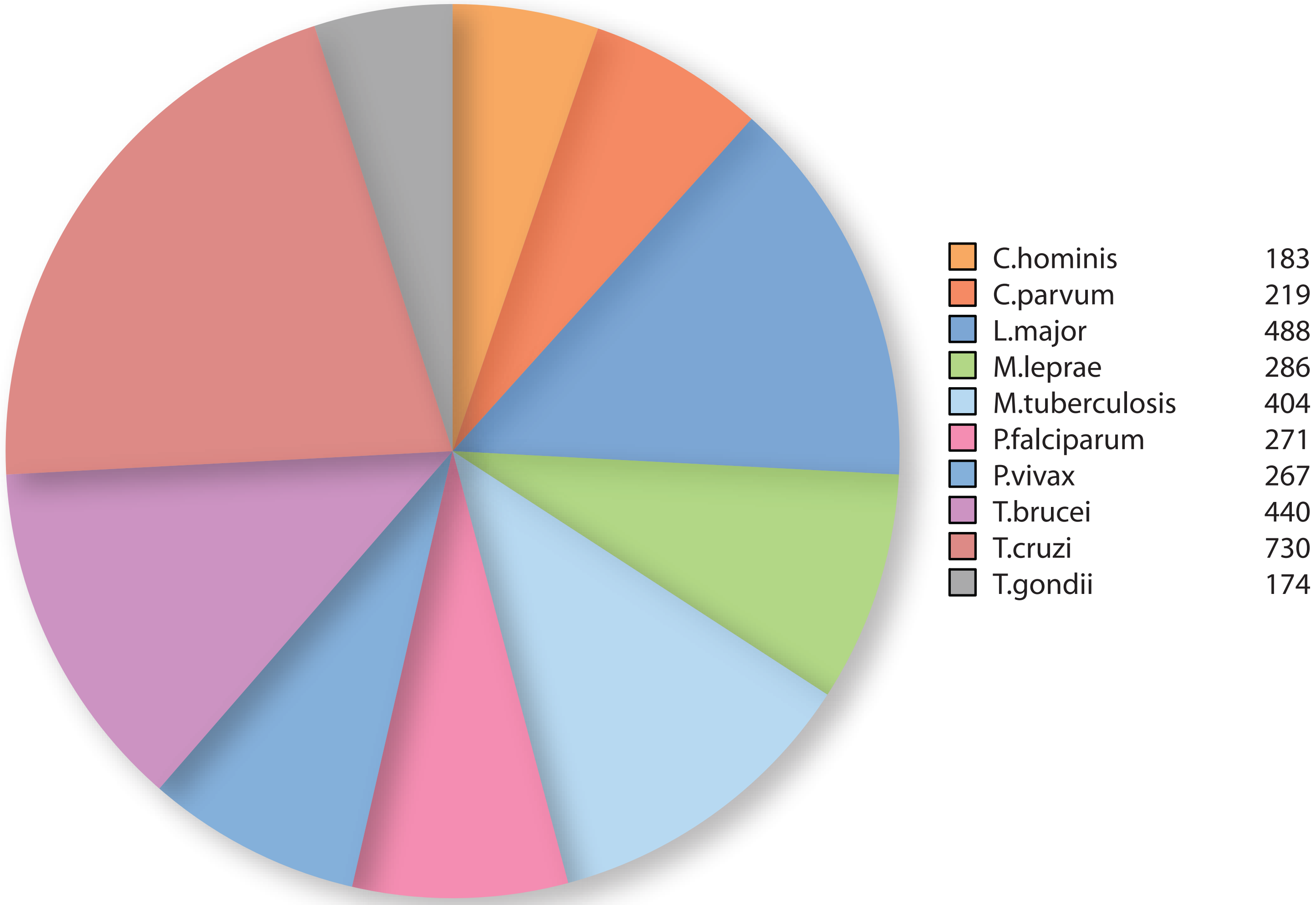
second cut-offs

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.

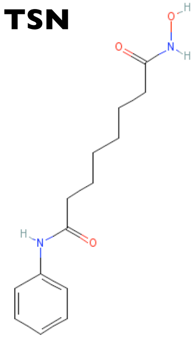
	Origen	Formula	Name	Cov.	Seq, Id. (%)
ZN	X-ray	Zn^{2+}	Zinc ion	--	--
NA	X-ray	Na^{+}	Sodium ion	--	--
CA	X-ray	Ca^{2+}	Calcium ion	--	--
TSN	X-ray	$C_{17} H_{22} N_2 O_3$	Trichostatin A	--	--
SHH	Expanded	$C_{14} H_{20} N_2 O_3$	Octadenioic acid hudroxyamide phenylamide	100.00	83.8

Example of inheritance (inheritance)

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

	Formula	Name	Cov.	Seq, Id. (%)	Residues
TSN	$C_{17} H_{22} N_2 O_3$	Trichostatin A	100.00	90.9	90 131 132 140 141 167 169 256 263 293 295
SHH	$C_{14} H_{20} N_2 O_3$	Octadenioic acid hudroxyamide phenylamide	100.00	90.9	

TSN



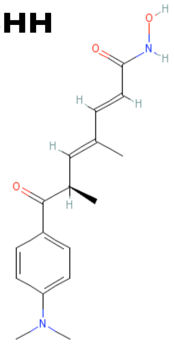
suberoylanilide hydroxamic acid

Pharmacological Action:

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

Inhibits histone deacetylase I and 3

SHH



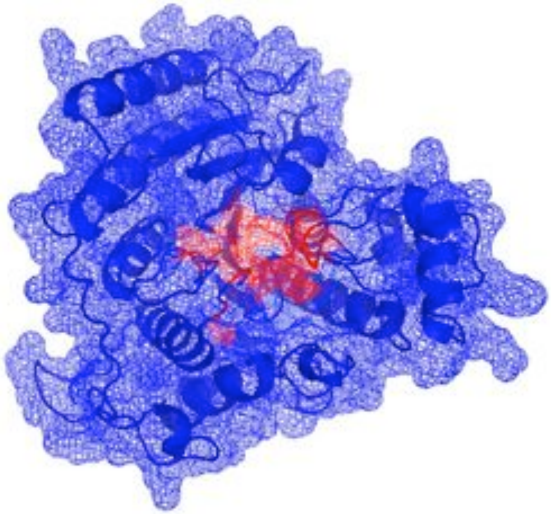
trichostatin A

Pharmacological Action:

- Antibiotics, Antifungal
- Enzyme Inhibitors
- Protein Synthesis Inhibitors

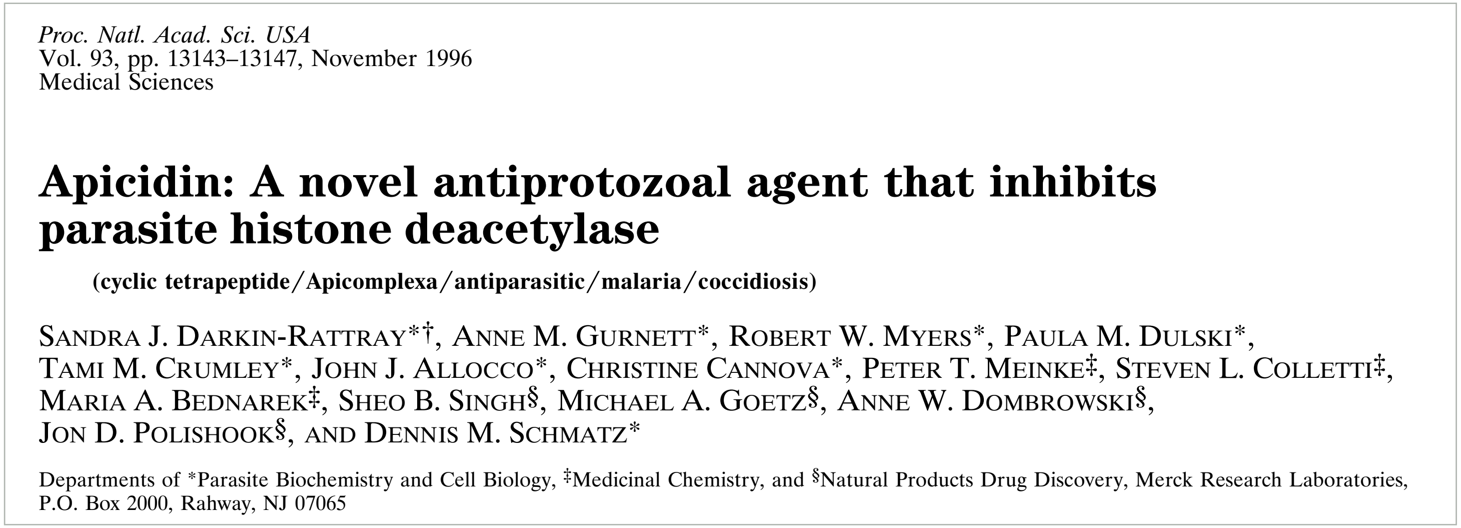
chelates zinc ion in the active site of histone deacetylases, resulting in preventing histone unpacking so DNA is less available for transcription

	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)*



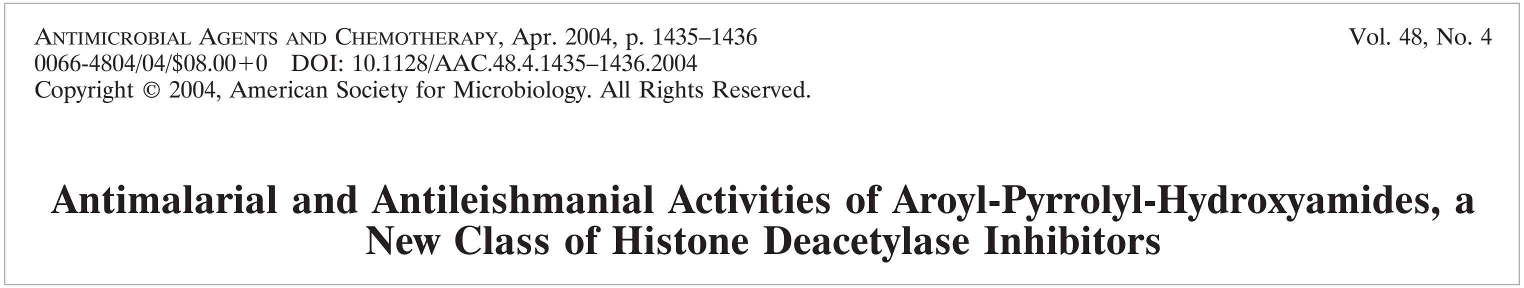
Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 13143–13147, November 1996
Medical Sciences

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[*]

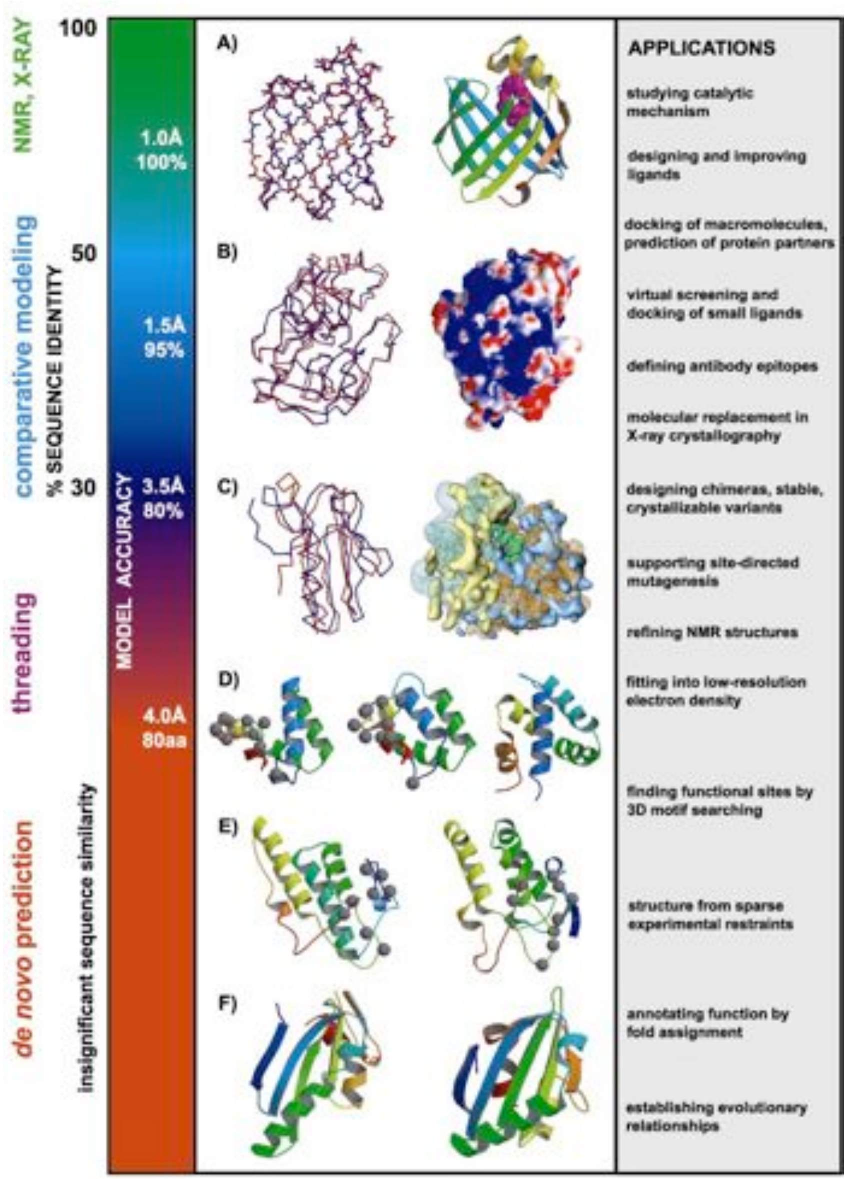
Departments of [*]Parasite Biochemistry and Cell Biology, [‡]Medicinal Chemistry, and [§]Natural Products Drug Discovery, Merck Research Laboratories, P.O. Box 2000, Rahway, NJ 07065



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


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

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

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)