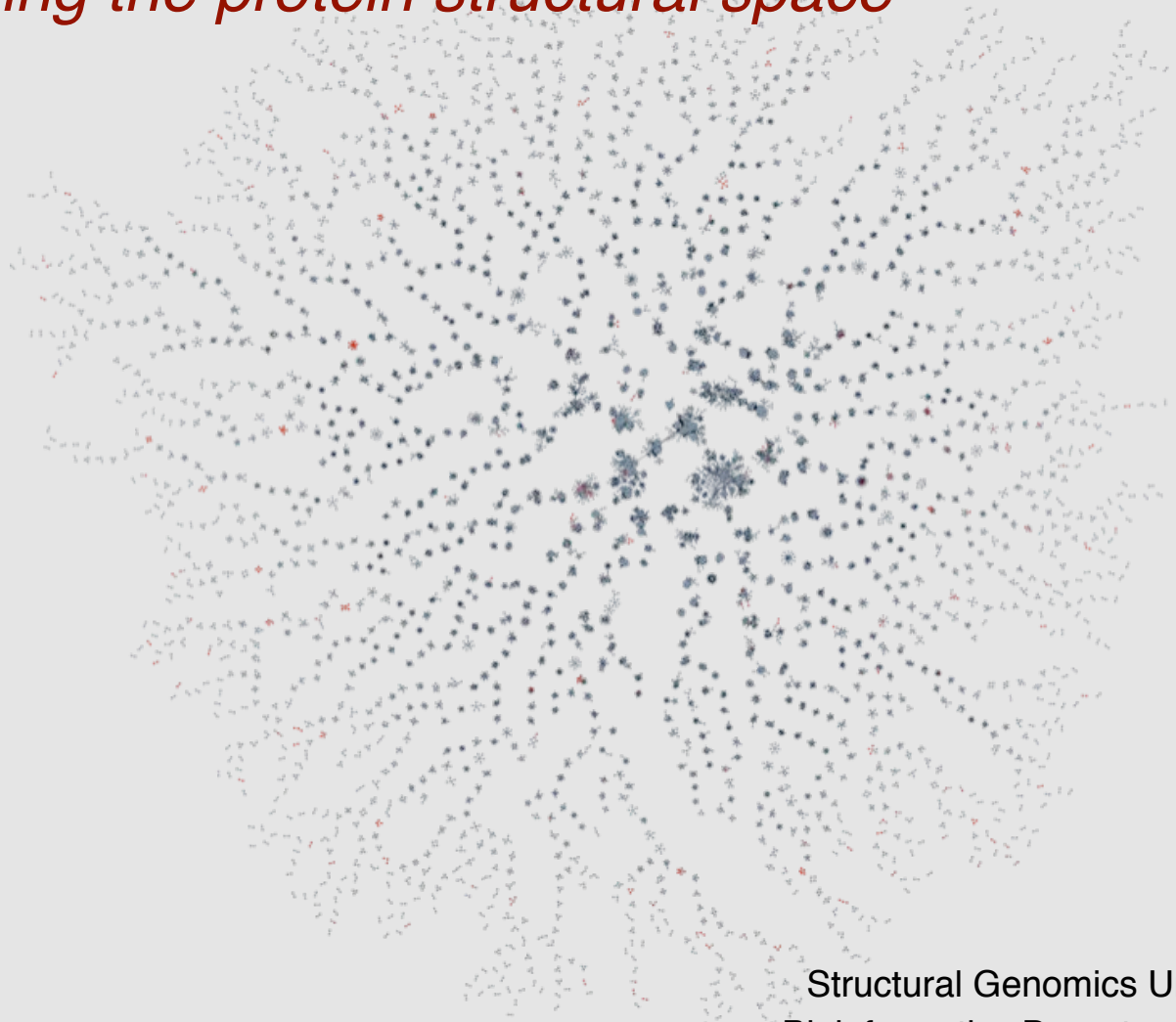
Comparative STRUCTURE Biology

Mining the protein structural space



Marc A. Marti-Renom

http://sgu.bioinfo.cipf.es

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

Bioinformatics Department http://bioinfo.cipf.es

Functional Genomics
Dr. Joaquín Dopazo
jdopazo@cipf.es

GEPAS

Pupa Suite

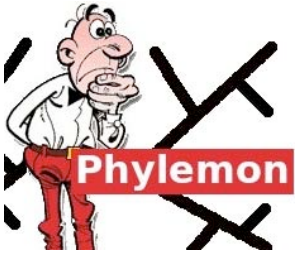
BABELOMICS

http://gepas.bioinfo.cipf.es
Gene Expression Pattern Analysis Suite

http://pupasuite.bioinfo.cipf.es
SNP Analysis Suite

http://babelomics.bioinfo.cipf.es
Functional Profiling Analysis Suite

Comparative Genomics
Pharmacogenomics
Dr. Hernán Dopazo
hdopazo@cipf.es



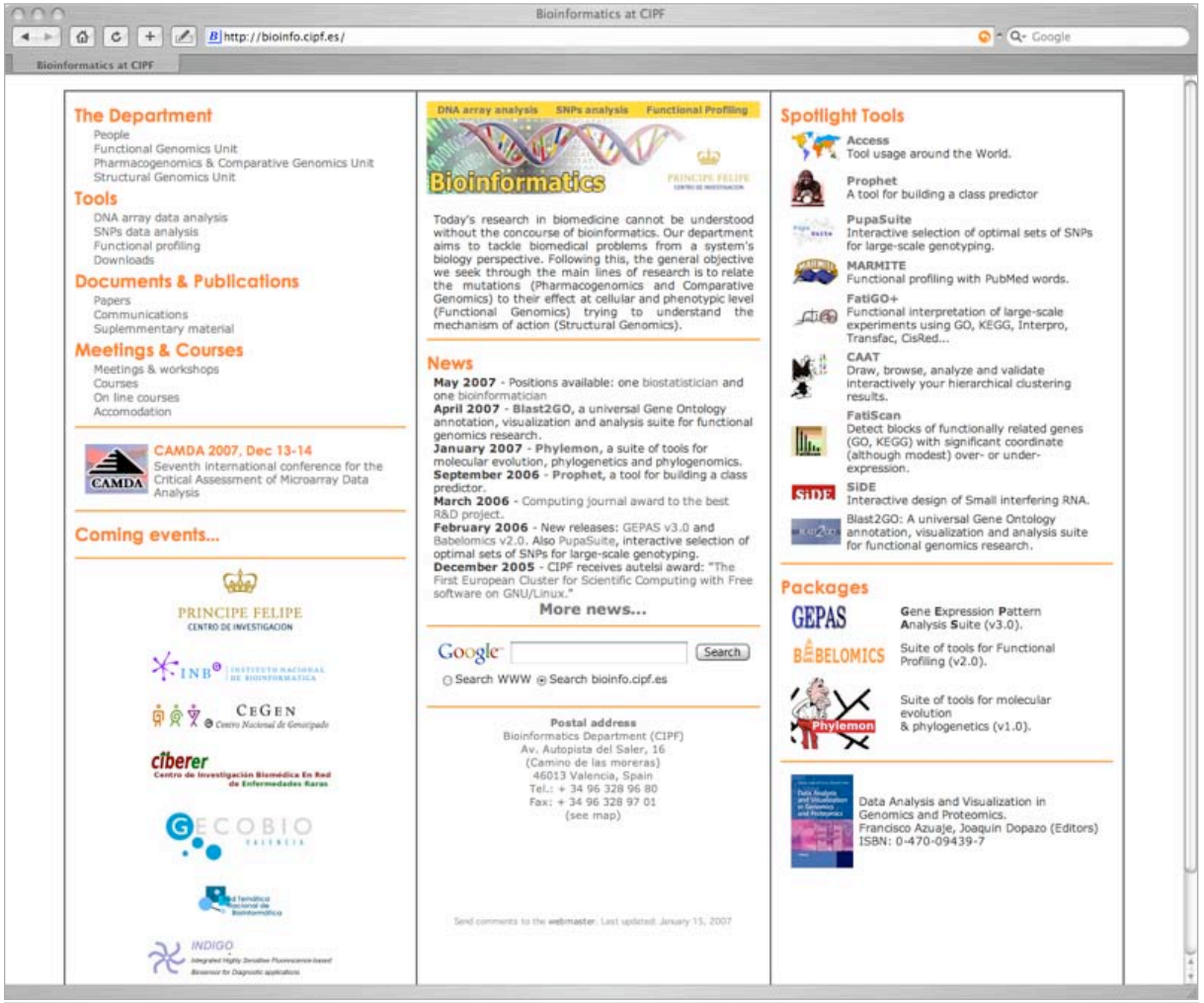
http://phylemon.bioinfo.cipf.es
Molecular Evolution Analysis Suite

Structural Genomics
Dr. Marc A. Marti-Renom
mmarti@cipf.es

DBAli v2.0
Home

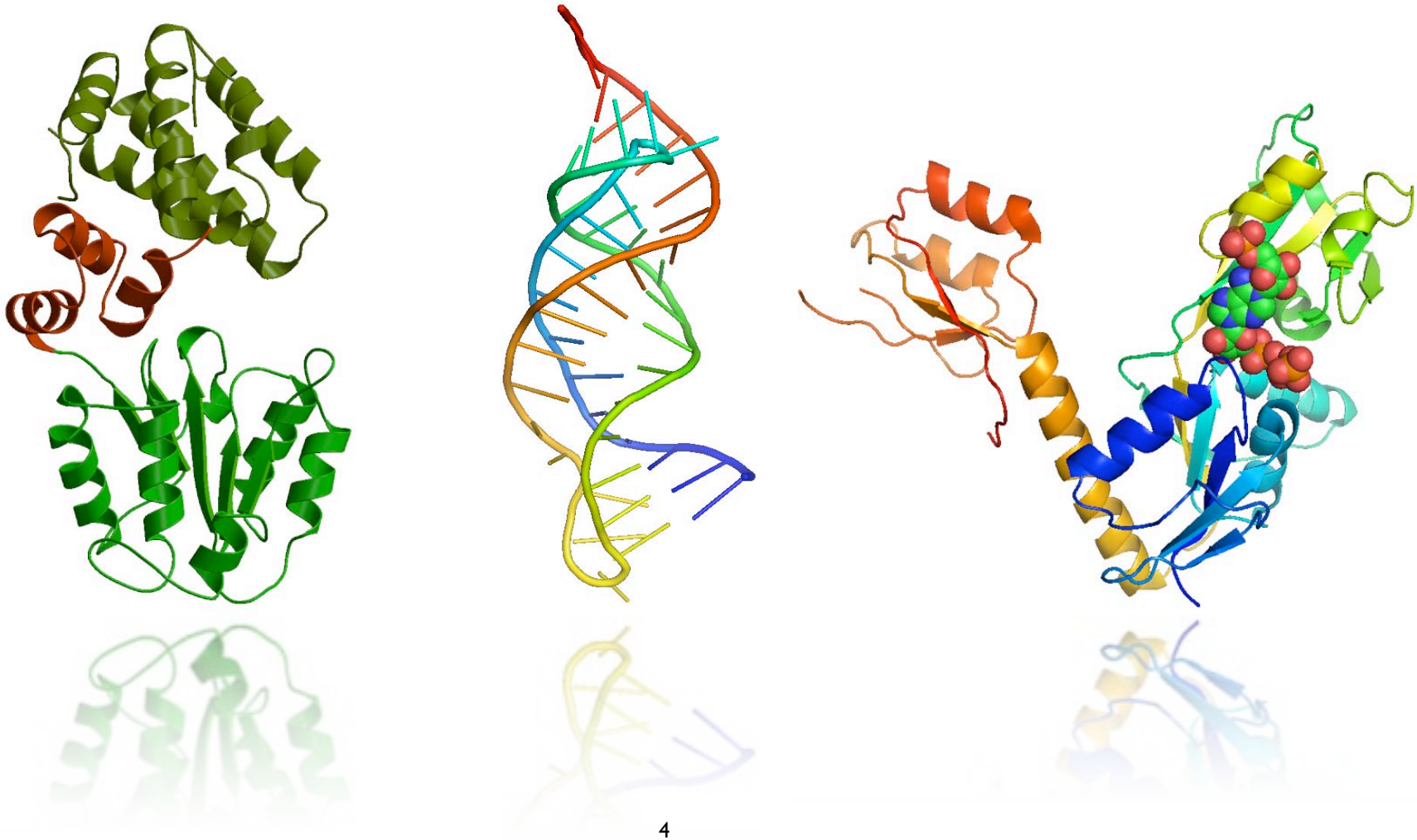
http://www.dbali.org
Structural Biology Analysis Suite

Bioinformatics Department http://bioinfo.cipf.es



Structural Genomics Unit

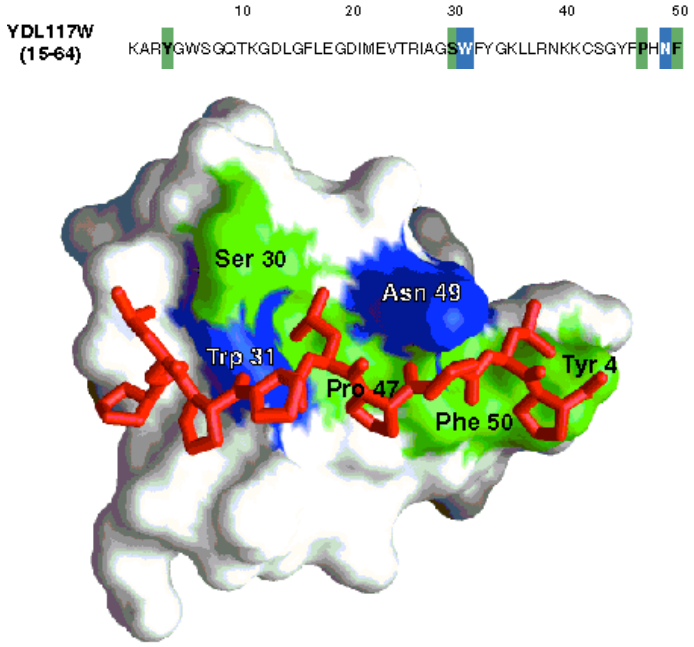
Bioinformatics Department, CIPF



Comparative Modeling
AnnoLyze - AnnoLite
Tropical Disease Initiative

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.

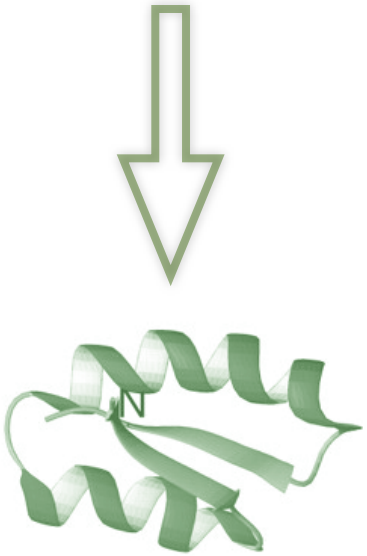


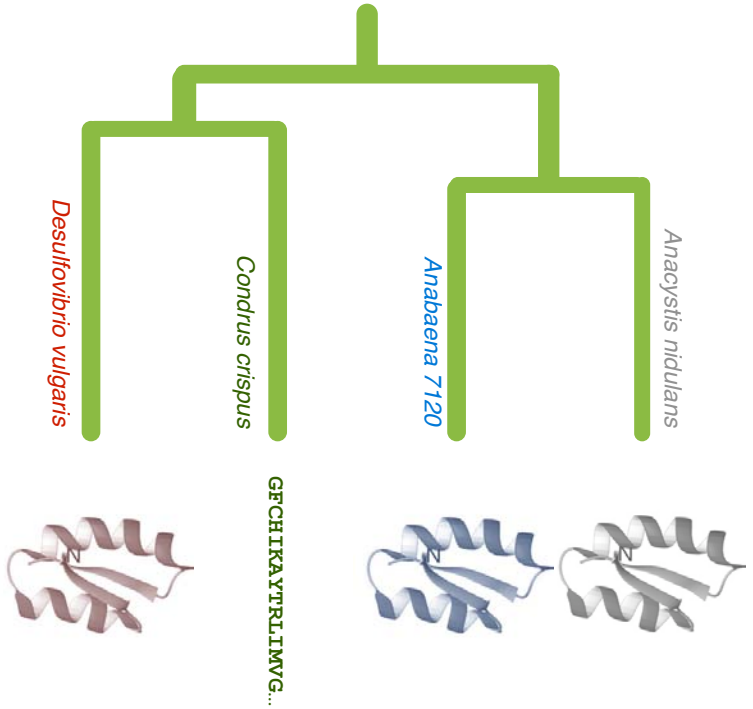
In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence**.

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.

protein prediction vs protein determination

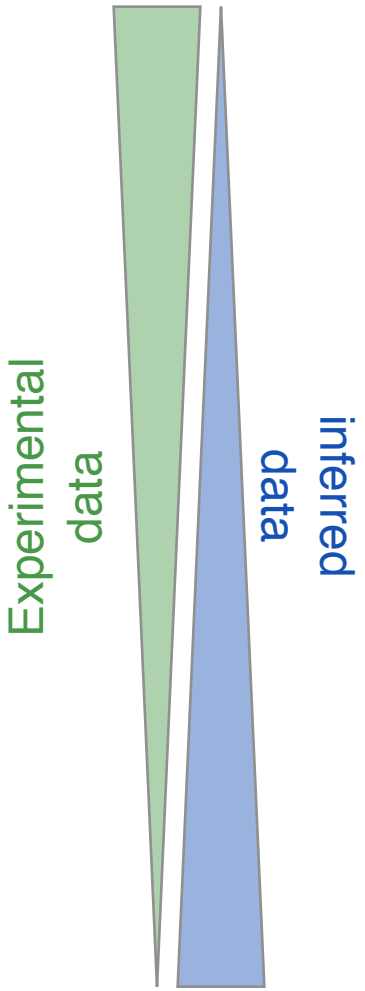
X-Ray

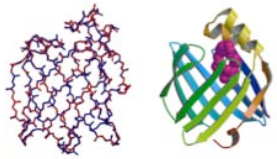
NMR

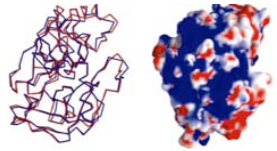
Comparative Modeling

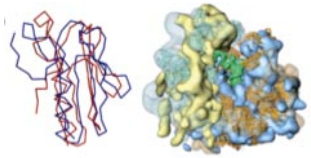
Threading

Ab-initio

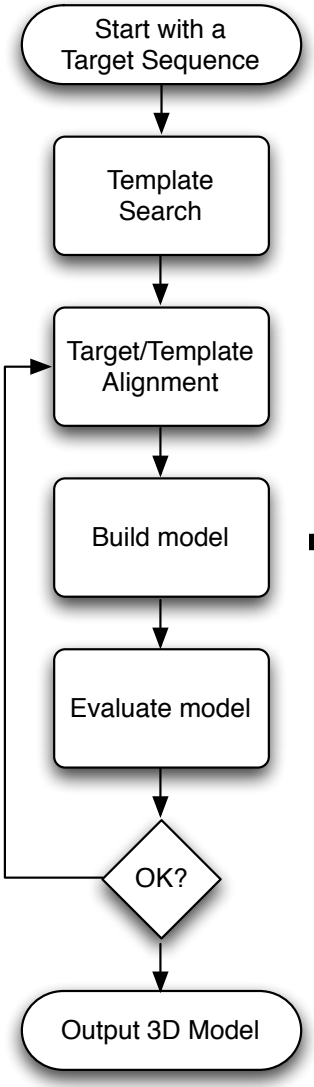








Comparative modeling by satisfaction of spatial restraints
MODELLER

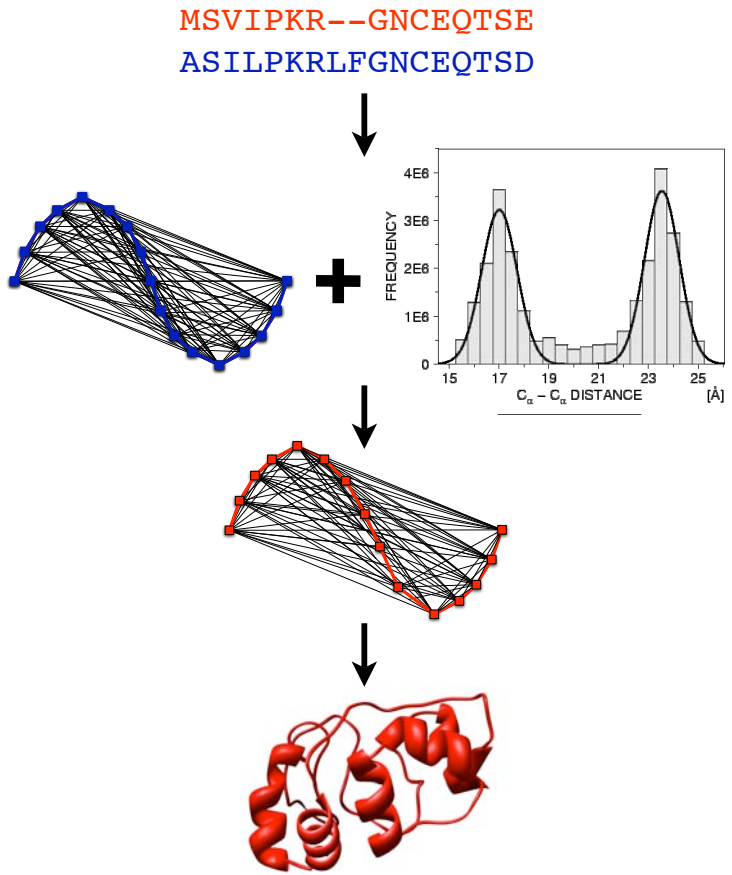


Given an alignment...

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

apply these features as restraints on your target sequence

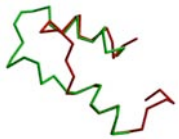
optimize to find the best solution for the restraints to produce your 3D model



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.

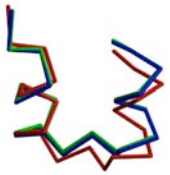
Comparative modeling by satisfaction of spatial restraints Types of errors and their impact

Wrong fold



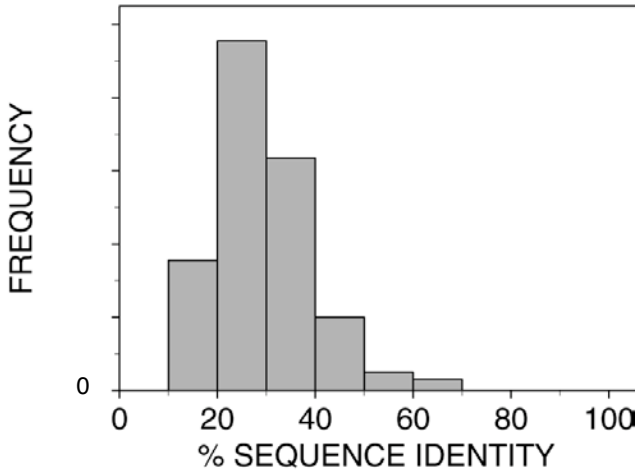
Miss alignments

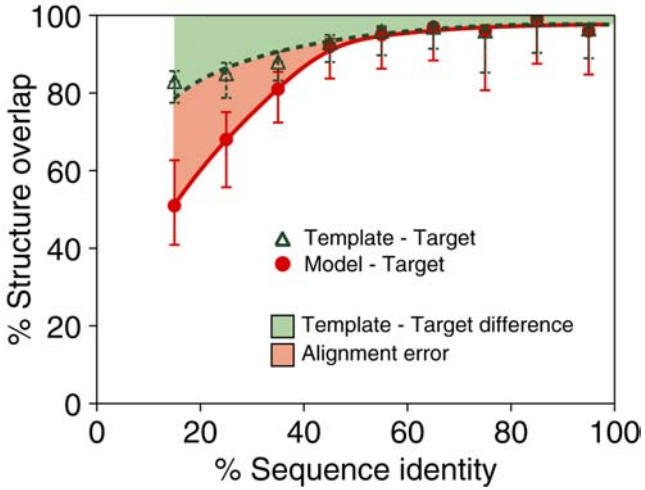
Loop regions



Rigid body distortions

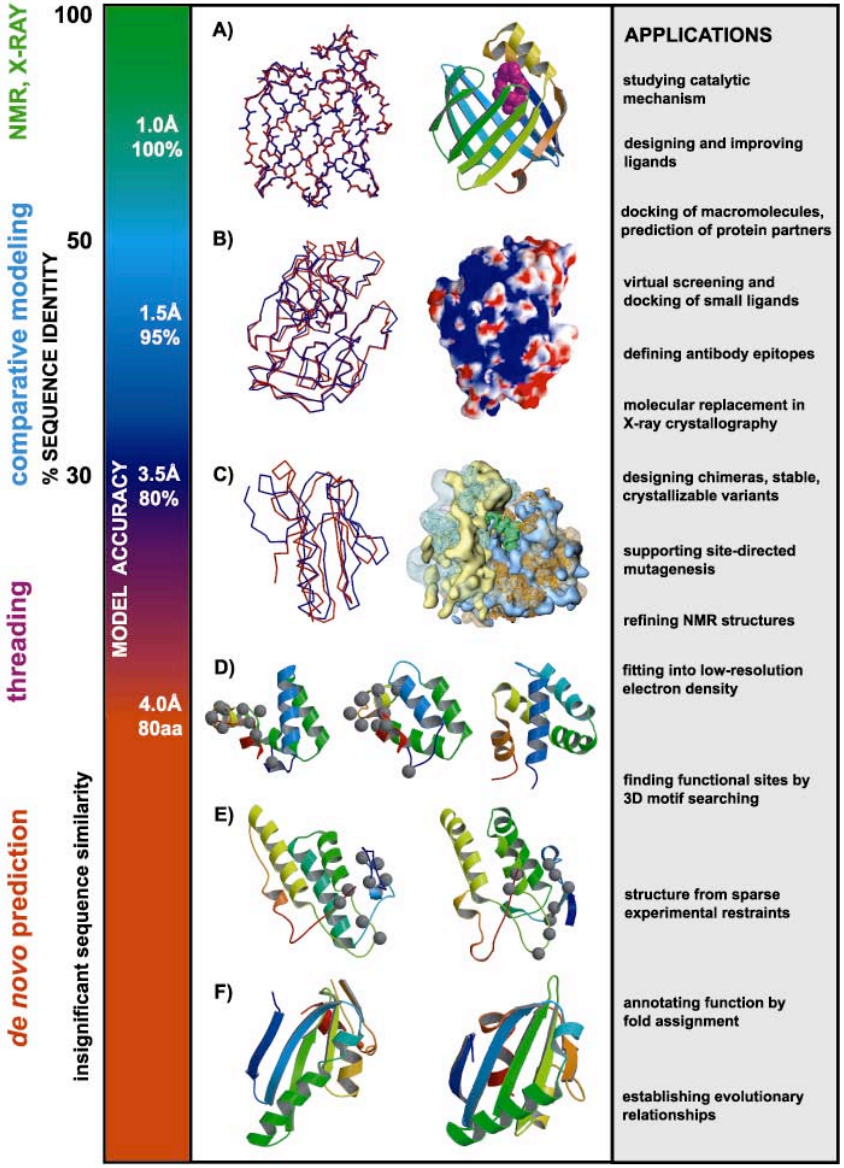
Side-chain packing





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

Utility of protein structure models, despite errors

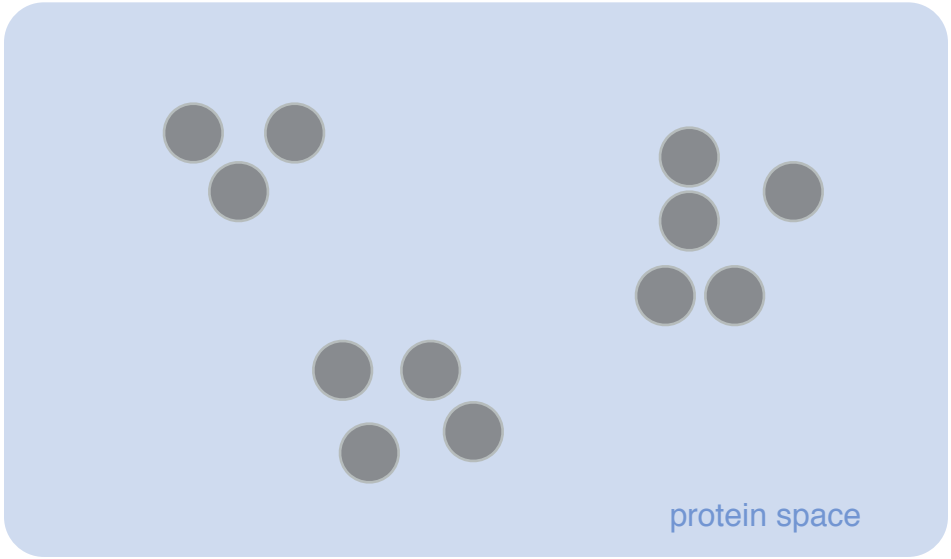


D. Baker & A. Sali. Science 294, 93, 2001.

Structural Genomics

Characterize most protein sequences based on related known structures

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



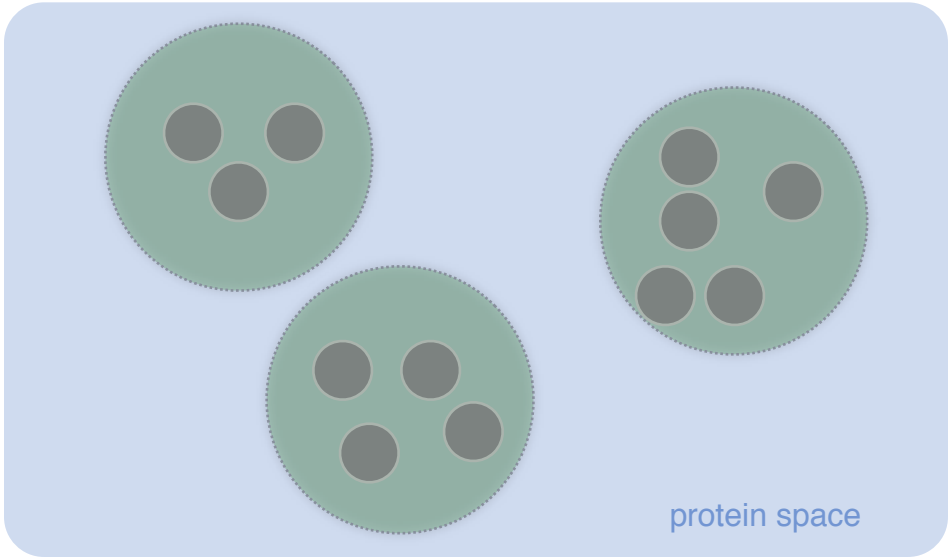
There are **~16,000** families (90%)
@ 30% sequence identity cutoff

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

Structural Genomics

Characterize most protein sequences based on related known structures

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



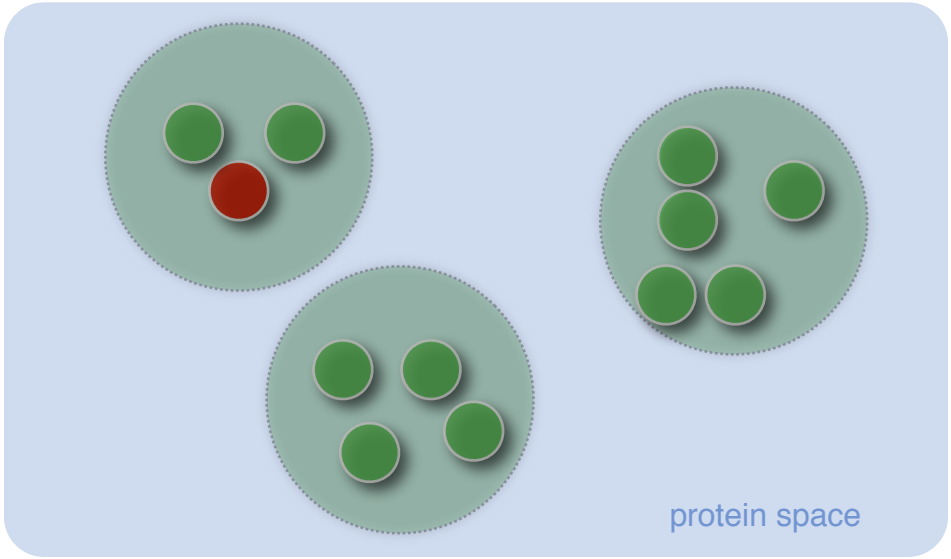
There are **~16,000** families (90%)
@ 30% sequence identity cutoff

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

Structural Genomics

Characterize most protein sequences based on related known structures

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



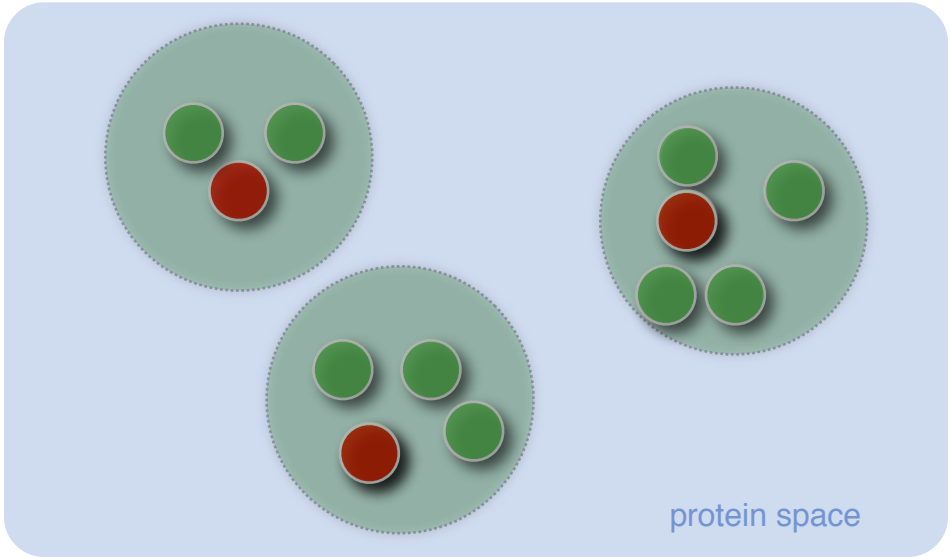
There are **~16,000** families (90%)
@ 30% sequence identity cutoff

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

Structural Genomics

Characterize most protein sequences based on related known structures

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

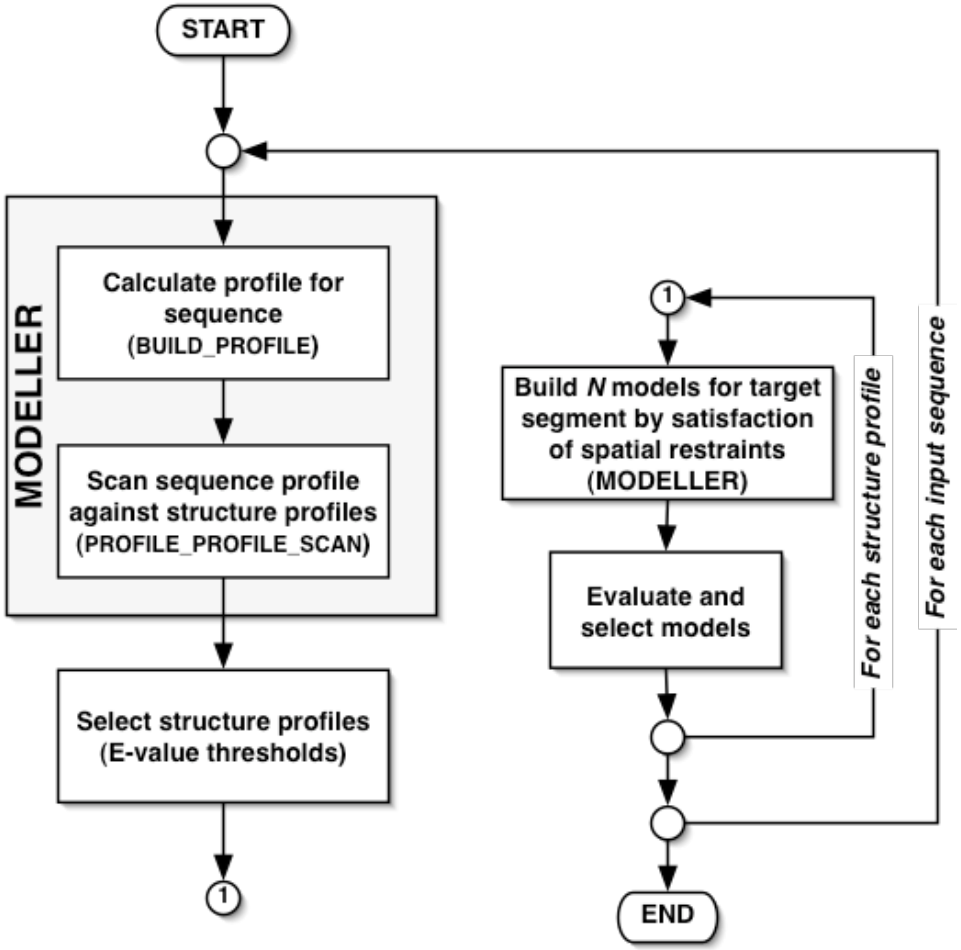


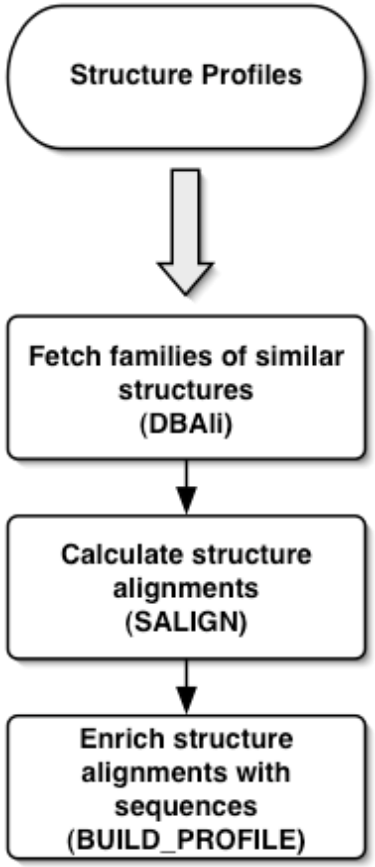
There are **~16,000** families (90%)
@ 30% sequence identity cutoff

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

MODPIPE2.0

Large-Scale Protein Structure Modeling





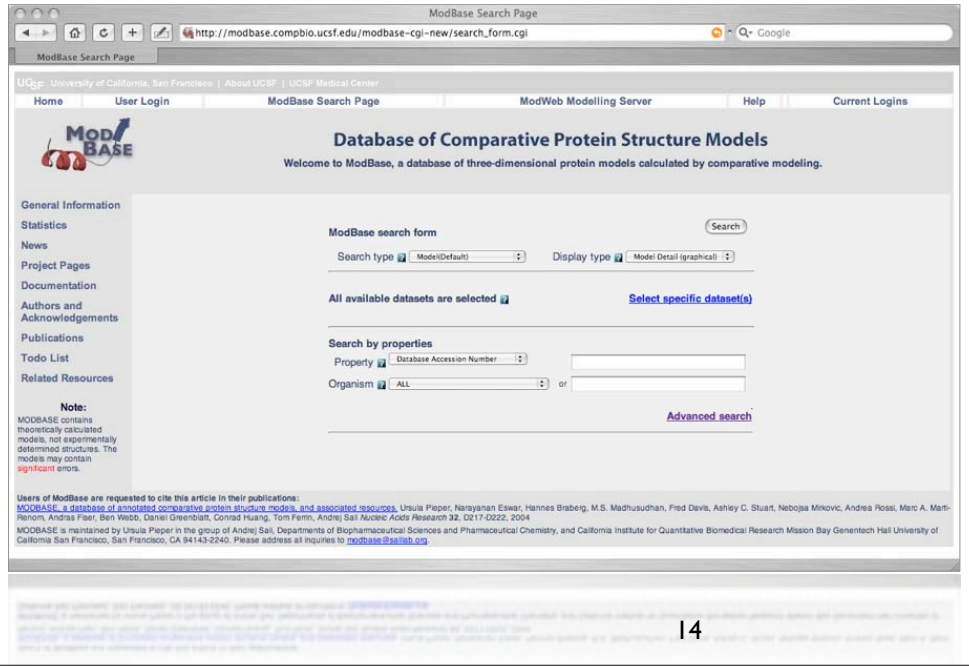
Eswar et.al., (2003) Nucl.Acids.Res. 31(13)

ModBase Statistics

Large-scale modeling of the TrEMBL-SWISSPROT databases

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

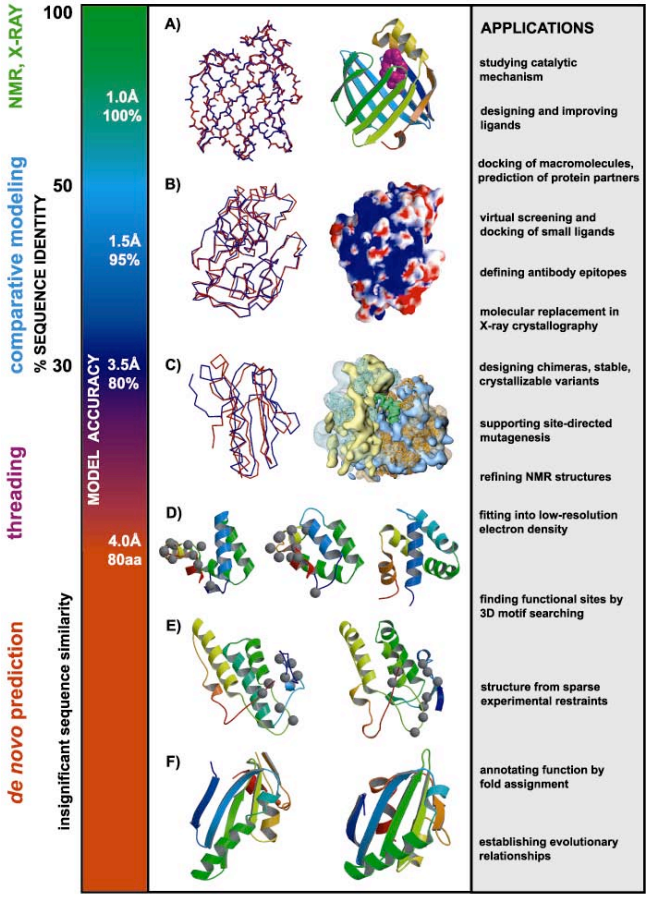
Sequences (total)	2,186,210
Sequences (modeled)	1,340,687
Models	4,284,570

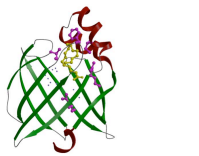


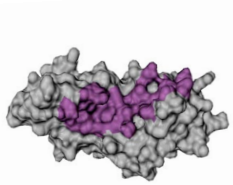
UCSF
University of California
San Francisco

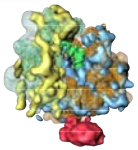
Pieper et al. NAR 34, D291 (2006)

Can we use models to infer function?

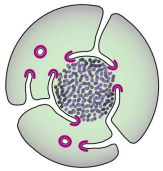








T. cruzi



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

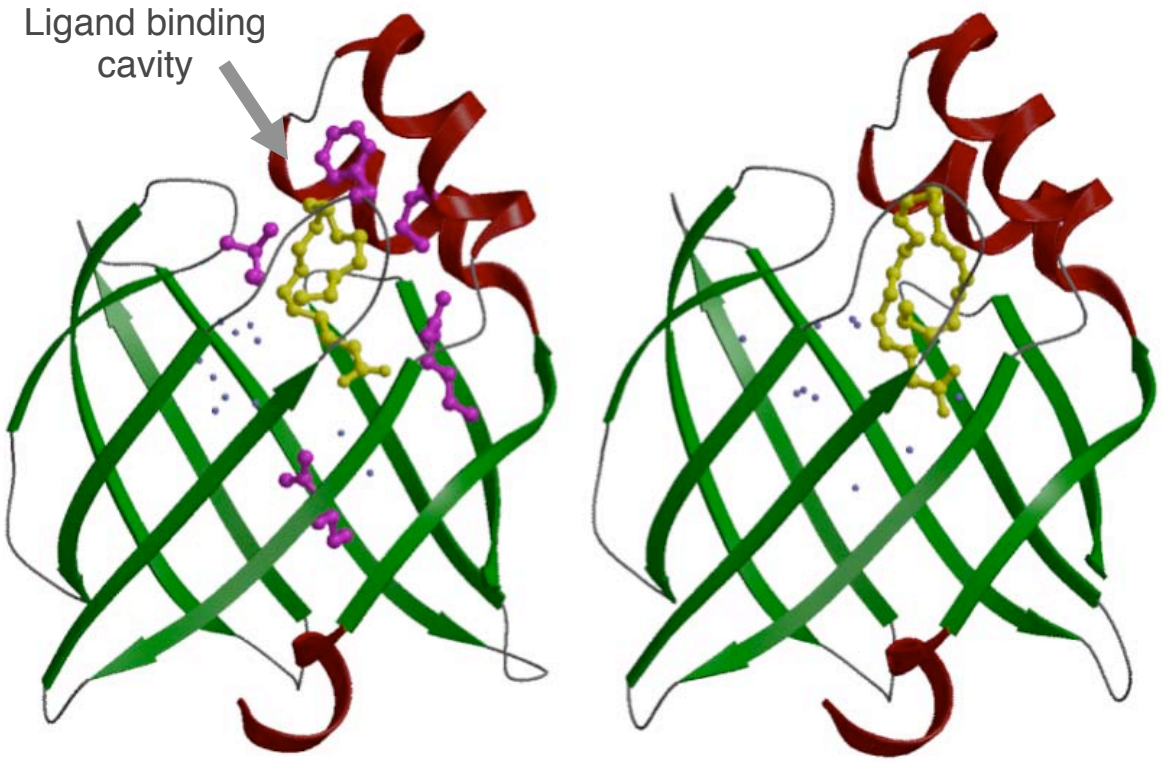
Predicting features of a model that are not present in the template

BLBP/oleic acid

Cavity is not filled

BLBP/docosahexaenoic acid

Cavity is filled



1. BLBP binds fatty acids.

2. Build a 3D model.

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

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

Structural analysis of missense mutations in human BRCA1 BRCT domains

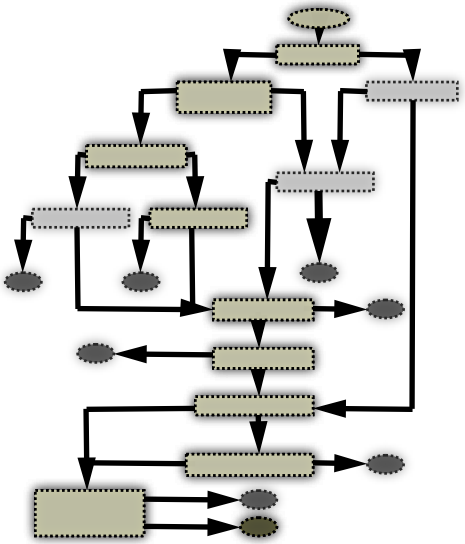
Nebojsa Mirkovic, Marc A. Marti-Renom, Barbara L. Weber, Andrej Sali and Alvaro N.A. Monteiro

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

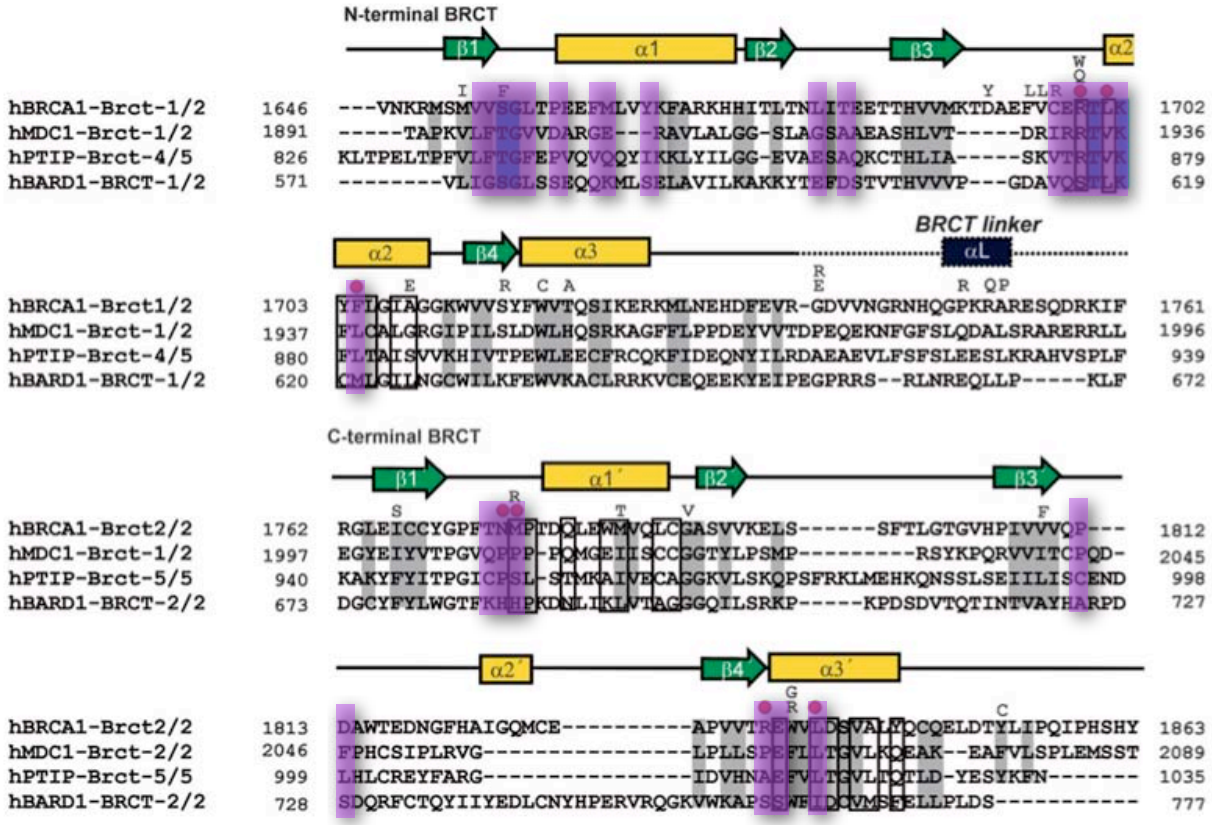
Cannot measure the functional impact of every possible SNP at all positions in each protein! Thus, prediction based on general principles of protein structure is needed.

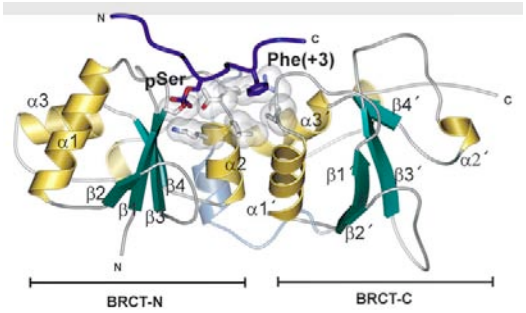
Missense mutations in BRCT domains by function

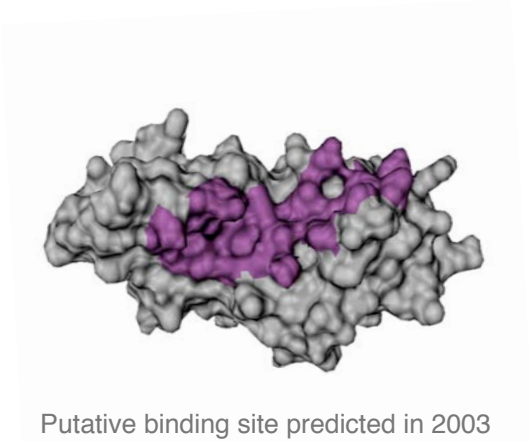
	cancer associate	not cancer associated	?
no transcription activation	C1697R R1699W A1708E S1715R P1749R M1775R		M1652K L1657P E1660G H1686Q R1699Q K1702E Y1703HF 1704S L1705PS 1715NS1 722FF17 34LG173 8EG1743 RA1752 PF1761I F1761S M1775E M1775K L1780P I1807S V1833E A1843T
transcription activation		M1652I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



Putative binding site on BRCA1





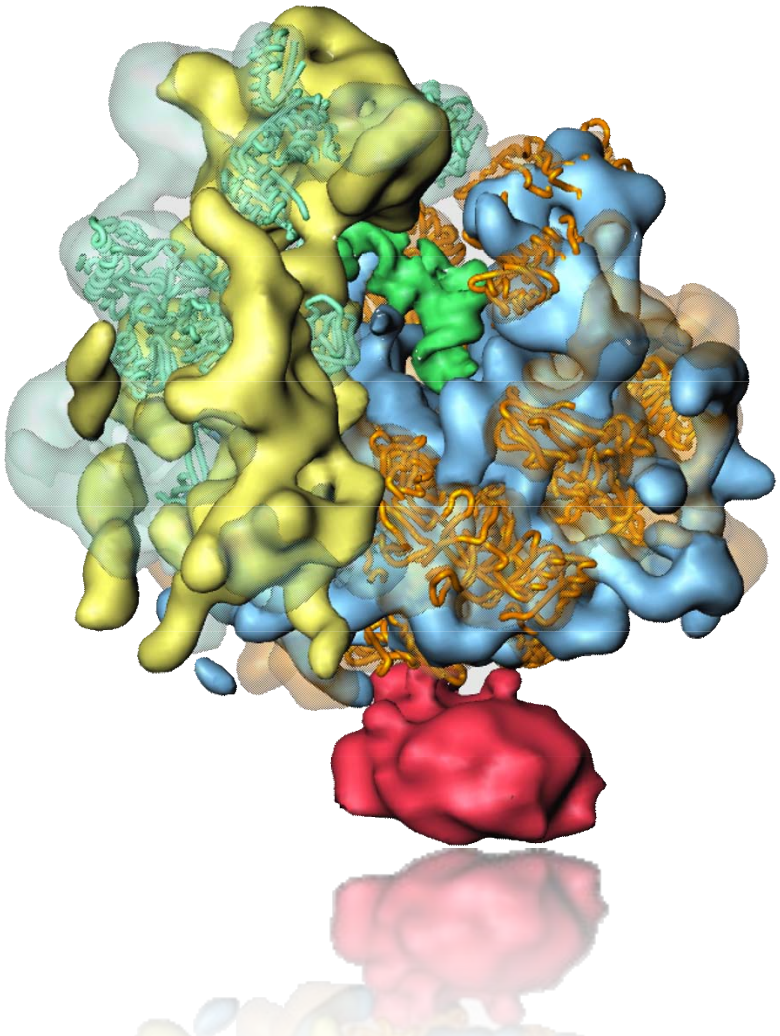


Putative binding site predicted in 2003 and accepted for publication on March 2004.

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

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

S. cerevisiae ribosome



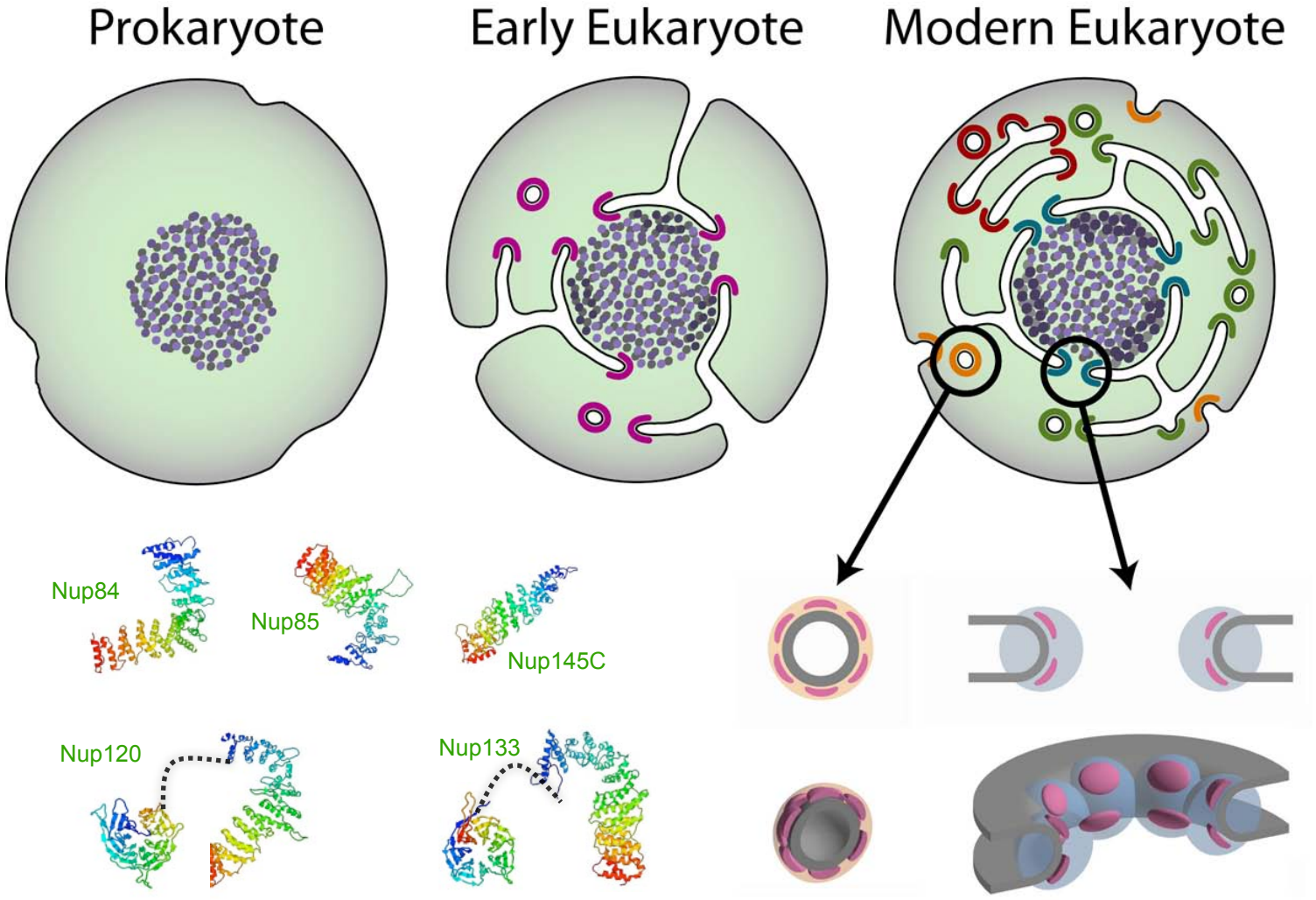
Fitting of comparative models into 15Å cryo-electron density map.

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

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

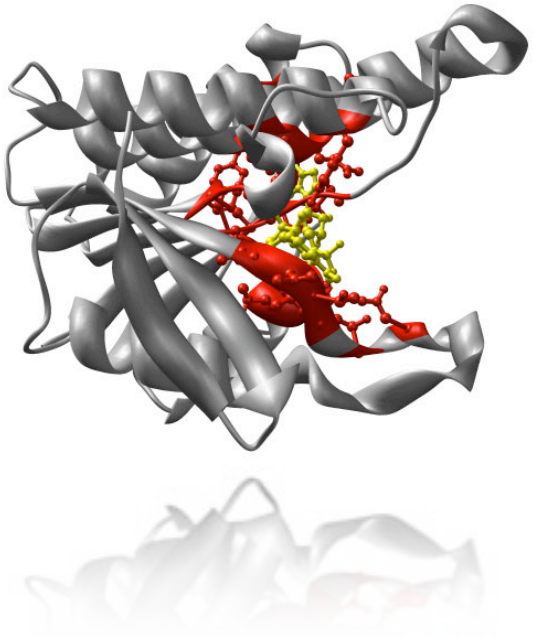
C. Spahn, R. Beckmann, N. Eswar, P. Penczek, A. Sali, G. Blobel, J. Frank. Cell 107, 361-372, 2001.

The Nucleopore complex
Cell evolution (?)



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

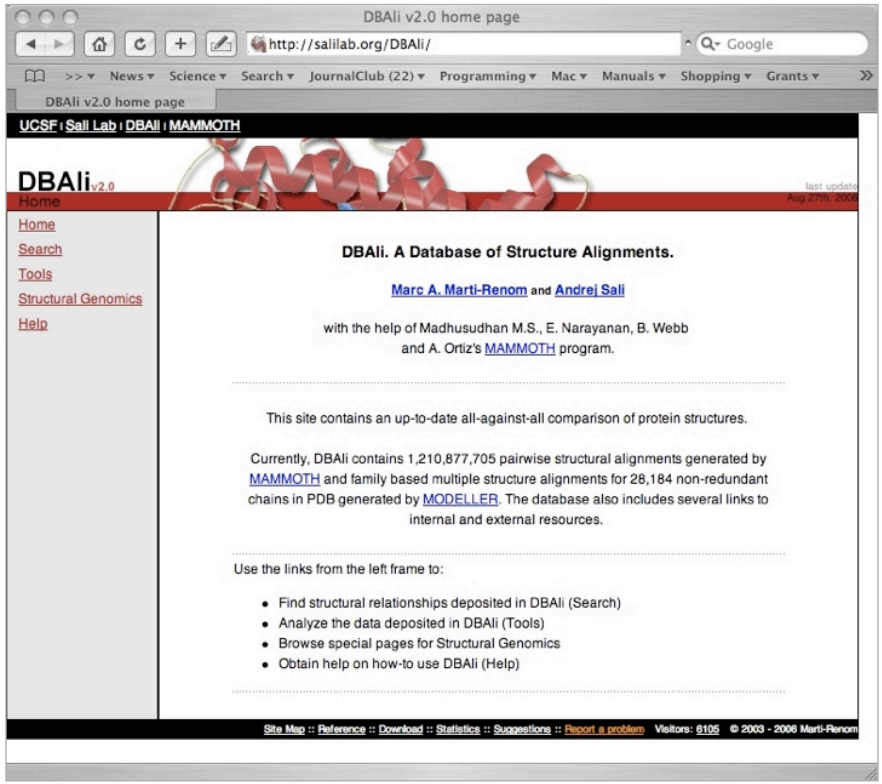
Protein function from structure
Comparative annotation. AnnoLite and AnnoLyze.



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

DBAli$_{v2.0}$ database

http://www.dbali.org



- ✓ 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:	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

Uses MAMMOTH for similarity detection

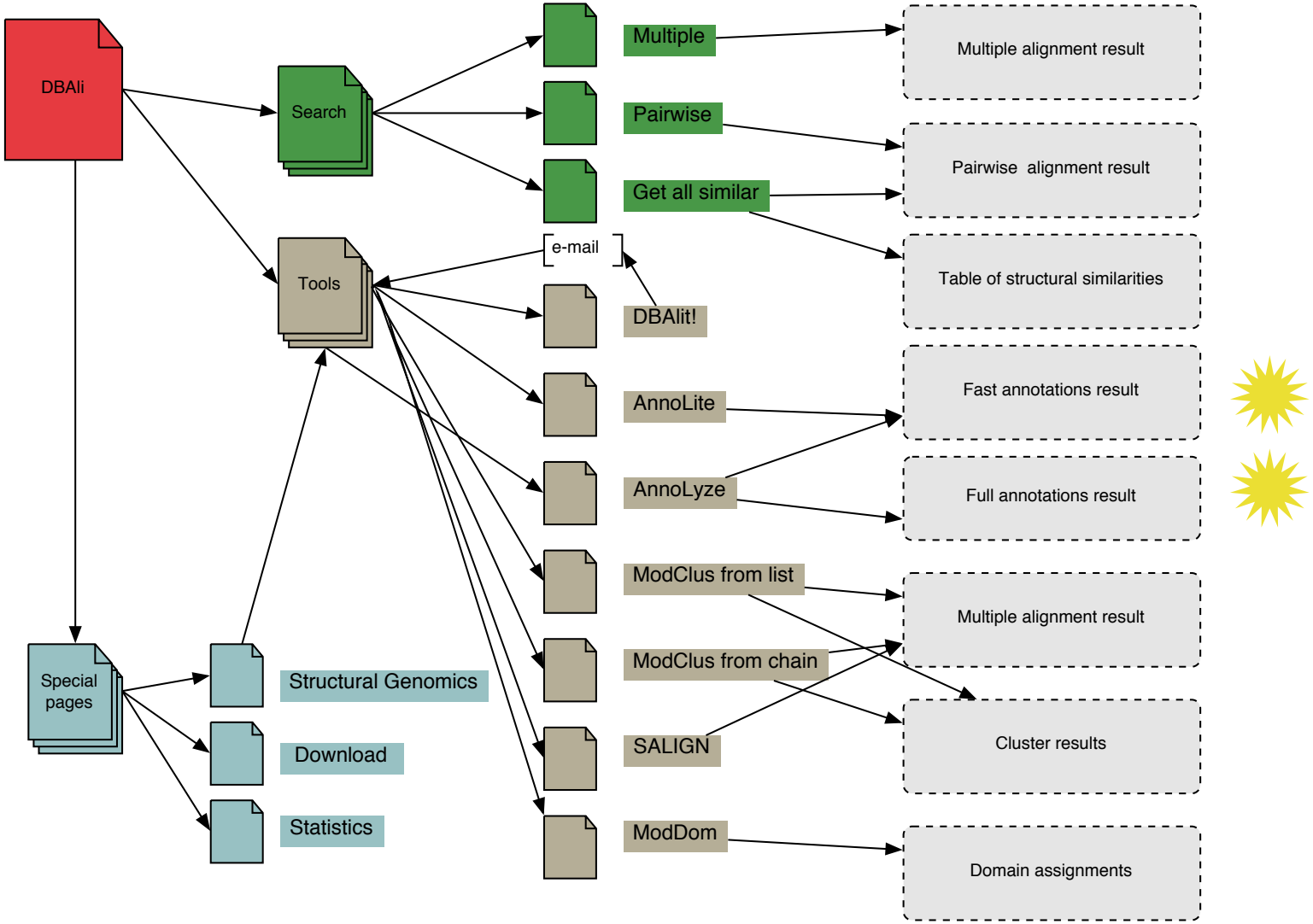
- ✓ **VERY FAST!!!**
- ✓ **Good scoring system with significance**

Ortiz AR, (2002) Protein Sci. 11 pp2606

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

DBAli$_{v2.0}$ database

http://www.dbali.org



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

AnnoLite

AnnoLite results for chain 1cgiA based on 44 structural similar chains.

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

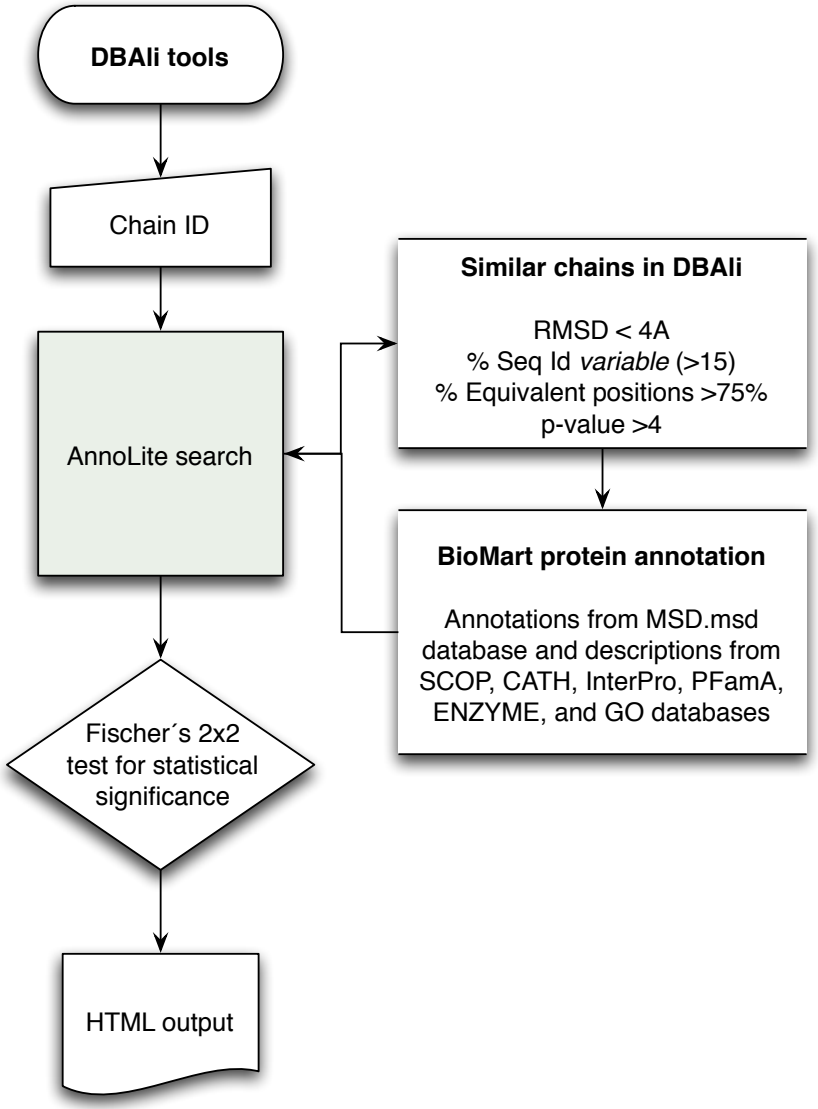
- Information annotated in the MSD database.
- High, medium and low confidence annotations not annotated in the MSD database.
- High, medium and low confidence annotations already annotated in the MSD database.

Benchmark set

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

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

Method



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

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

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

Scoring function

Fisher's 2x2 contingency test

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

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

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

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

$p = 1.78e^{-19}$

Sensitivity .vs. Precision

	Optimal cut-off	Sensitivity (%) Recall or TPR	Precision (%)
SCOP fold	1E-06	92.7	88.4
CATH fold	1E-03	95.7	90.1
InterPro	1E-03	88.4	78.2
PFam family	1E-04	90.5	82.8
EC number	1E-04	93.3	79.7
GO Molecular Function	1E-01	84.3	80.9
GO Biological Process	1E-03	85.5	74.8
GO Cellular Component	1E-02	77.6	58.6

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

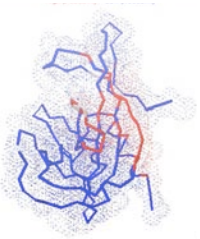
AnnoLyze

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
8OG	20.00	0.111	19 20 21 48 49 51 96 98 136
ACY	15.87	0.163	23 29 31 37 44 45 81 83 85 94 96 98 103 121 135

Inherited partners:1

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



Benchmark

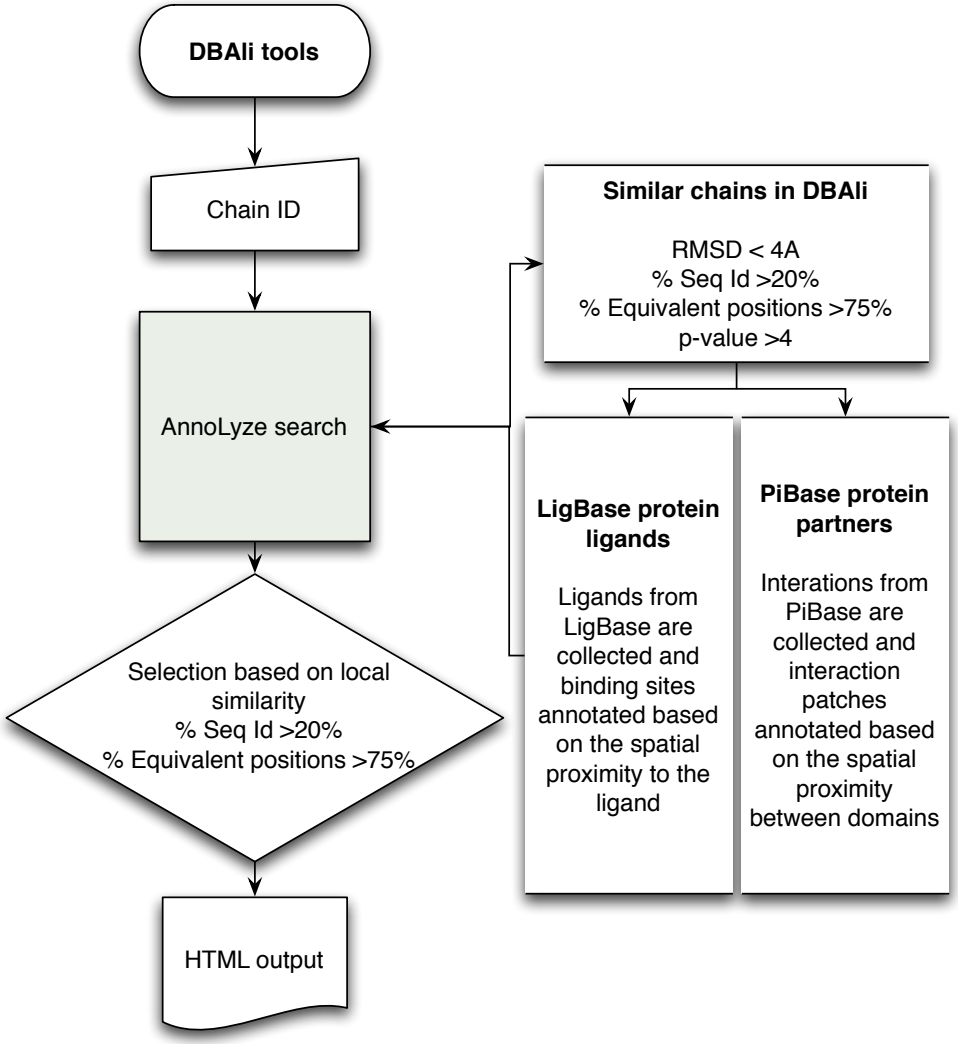
	Number of chains
Initial set*	78,167
LigBase**	30,126
Non-redundant set***	4,948 (8,846 ligands)

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

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

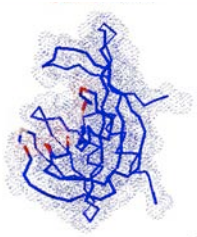
*all PDB chains larger than 30 aminoacids in length (8th of August, 2006)
**annotated with at least one partner in the πBase database
***not two chains can be structurally aligned within 3A, superimposing more than 75% of their Cα atoms, result in a sequence alignment with more than 30% identity, and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

Method



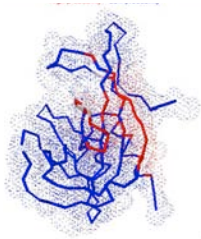
Inherited ligands: 4

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



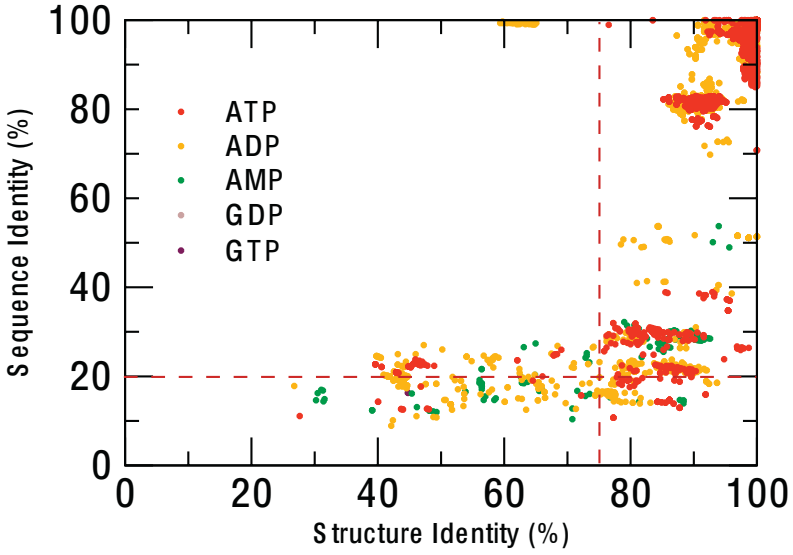
Inherited partners:1

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

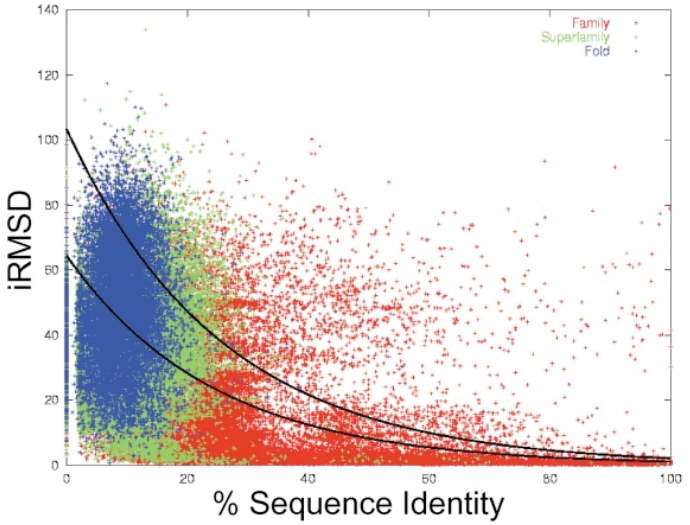


Scoring function

Ligands



Partners



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

Sensitivity .vs. Precision

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

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

Example (2azwA)

Structural Genomics Unknown Function

Molecule: MutT/nudix family protein

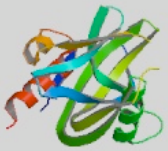
PDB ID: 2azw:A

Header:
STRUCTURAL GENOMICS, UNKNOWN FUNCTION

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

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

Resolution: 1.90Å



Links: none

SCOP: none

CATH: none

Sequence:
Md5: 09b13d23ceae0fdfcaddec636e2ddfa6KTPTAAAS
Length: 146

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

Ligands: none

Interacting partners: none

Similar structures: 20

Similar sequences: 890

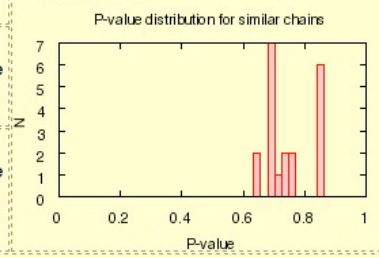
Most similar structure in DBAli:

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

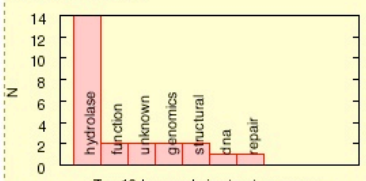
Most similar sequence in DBAli:

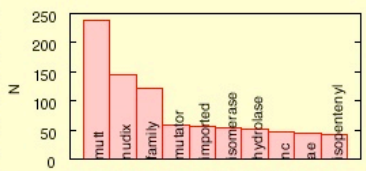
Code	SeqId(%)	EqPos	RMSD	P-Value	See
1vcd:B	24.59	122	3.47	17.00	ali

P-value distribution:

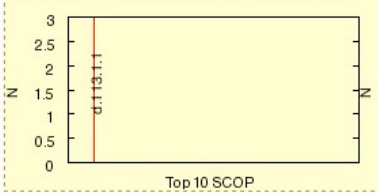


Keyword distribution:





SCOP and CATH distribution for similar structures:

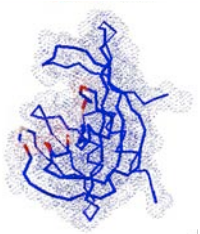


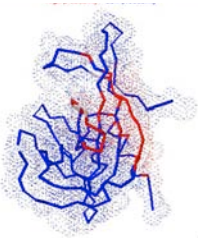
Inherited ligands: 4

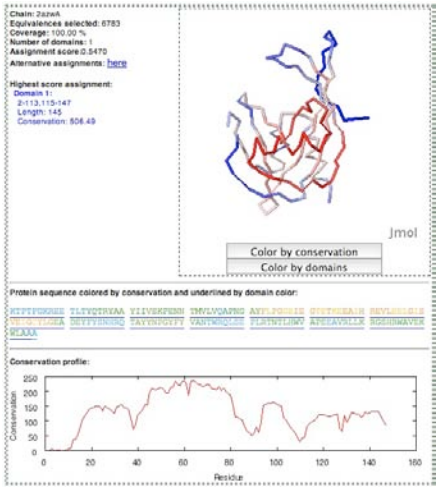
Ligand	Av. binding site seq. id.	Av. residue conservation	Residues in predicted binding site (size proportional to the local conservation)
MO2	59.03	0.185	48 49 52 62 63 66 67 113 116
CRY	20.00	0.111	23 29 31 37 44 48 49 83 85 94 96 103 121
8OG	20.00	0.111	19 20 21 48 49 51 96 98 136
ACY	15.87	0.163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

Inherited partners:1

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



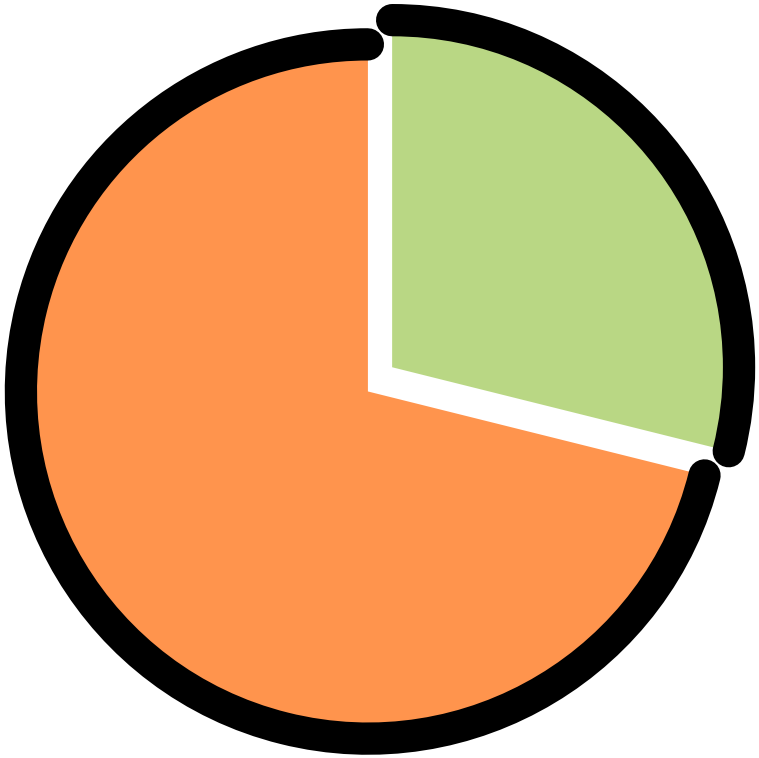




	Conf.	P-value	Link	Description
CATH:	●	1.1e-20	3.90.79.10	Nucleoside Triphosphate Pyrophosphohydrolase
SCOP:	●	4.2e-29	d.113.1.1	MutT-like
PFAM:	●	2.0e-74	PF00293	NUDIX domain
InterPro:	●	1.9e-65	IPR000086	NUDIX hydrolase
	●	2.7e-20	IPR003561	Mutator MutT
	●	2.9e-14	IPR002667	Isopentenyl-diphosphate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

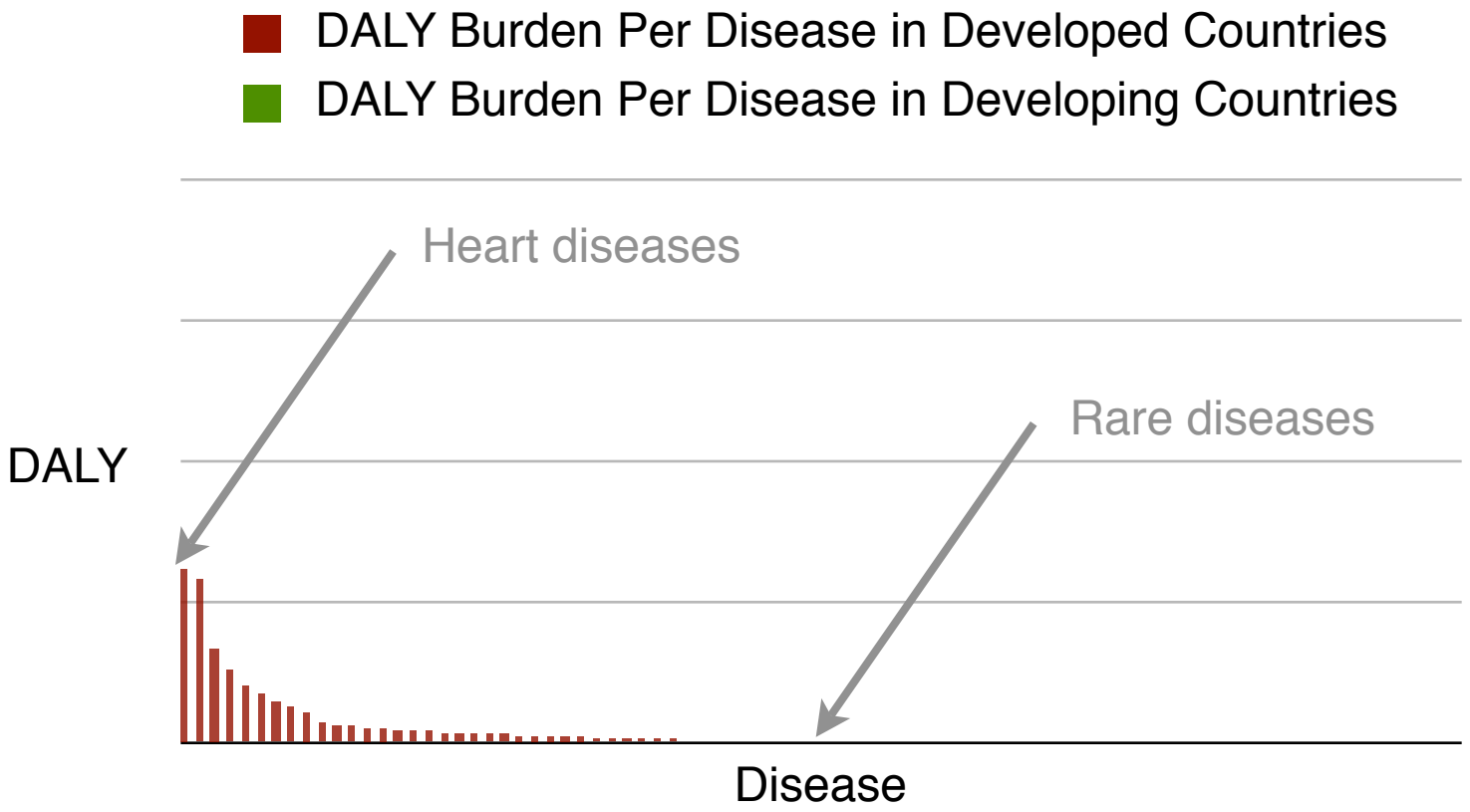
Tropical Disease Initiative (TDI)

Predicting binding sites in protein structure models.



http://www.tropicaldisease.org

Need is High in the Tail



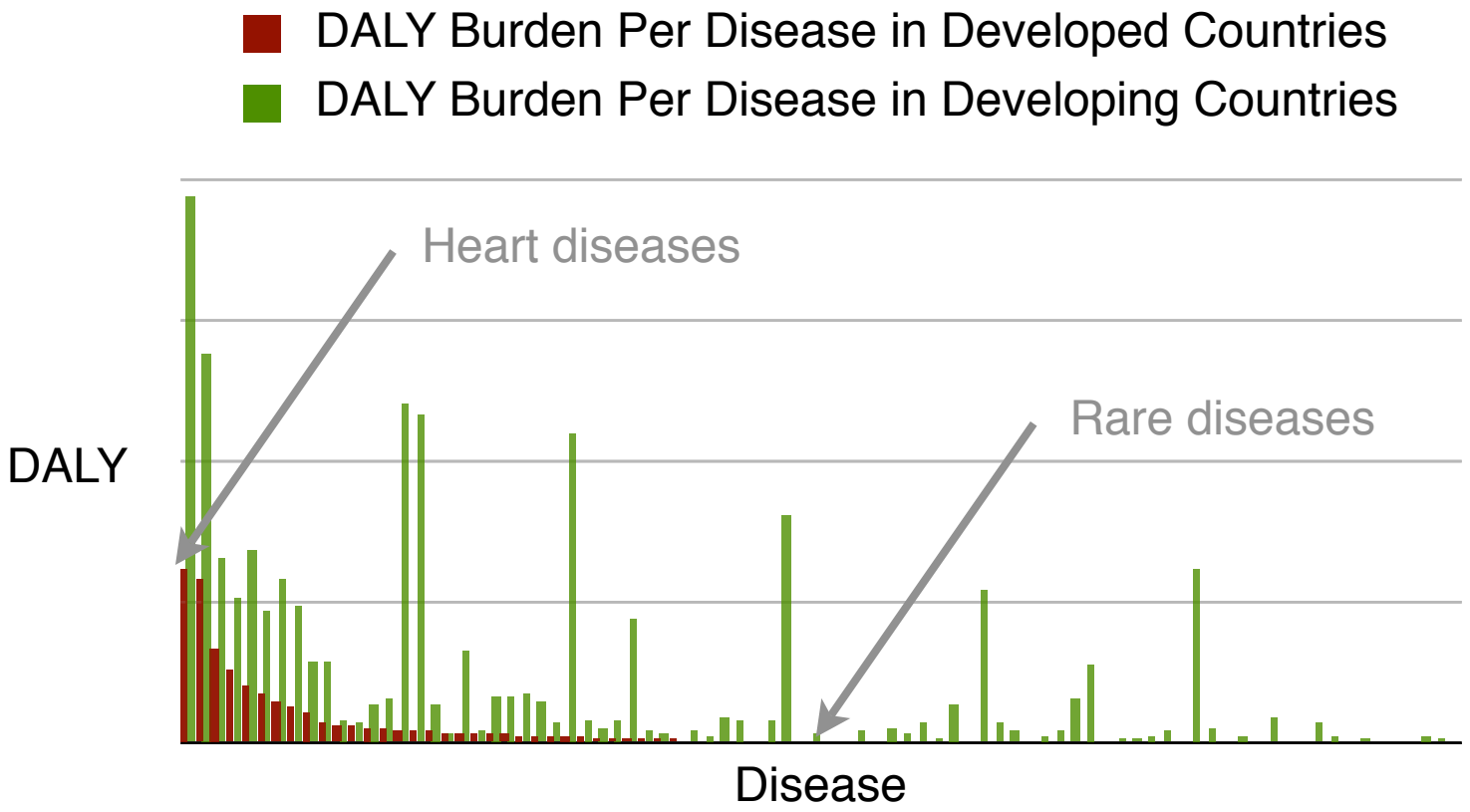
Disease data taken from WHO, *World Health Report 2004*

DALY - Disability adjusted life years

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

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

Need is High in the Tail



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

DALY - Disability adjusted life years

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

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

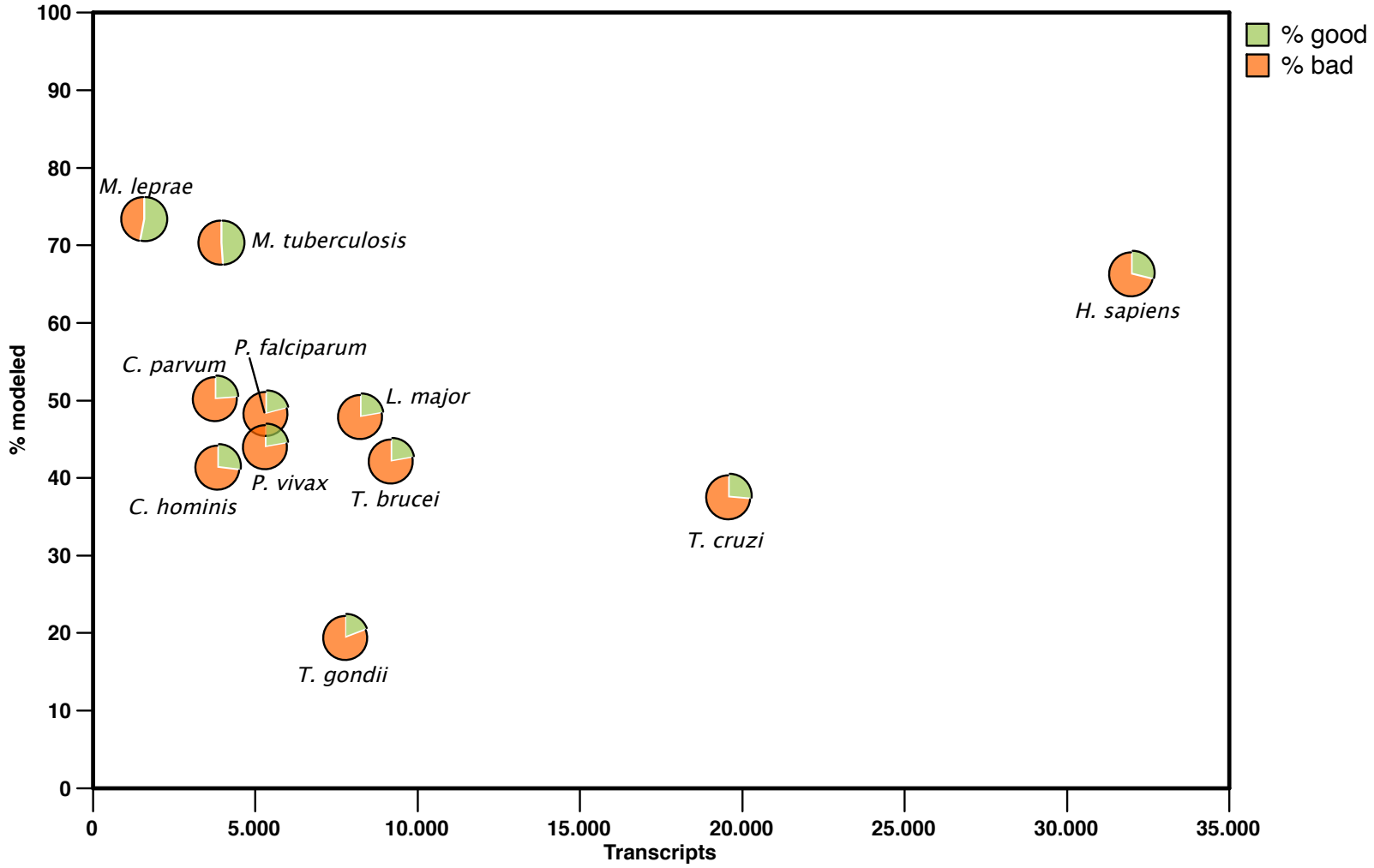
"Unprofitable" Diseases and Global DALY (in 1000's)

Malaria*	**46,486**	Trichuriasis	1,006
Tetanus	7,074	Japanese encephalitis	709
Lymphatic filariasis*	**5,777**	**Chagas Disease***	**667**
Syphilis	4,200	**Dengue***	**616**
Trachoma	2,329	**Onchocerciasis***	**484**
Leishmaniasis*	**2,090**	**Leprosy***	**199**
Ascariasis	1,817	Diphtheria	185
Schistosomiasis*	**1,702**	Poliomyelitise	151
Trypanosomiasis*	**1,525**	Hookworm disease	59

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

Modeling Genomes

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

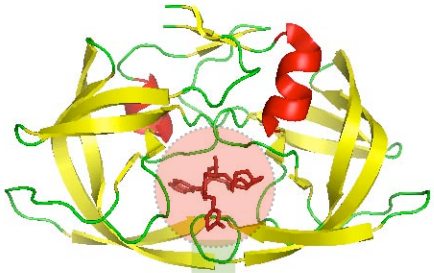


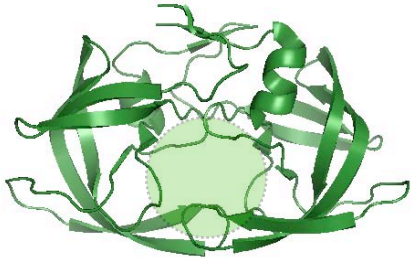
A good model has MPQS of 1.1 or higher

Comparative docking

1. Expansion

co-crystalized protein/ligand

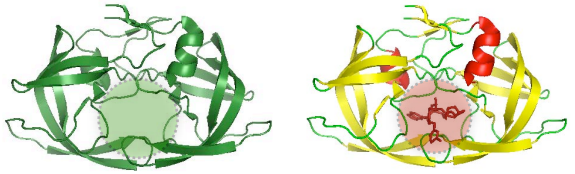




crystalized protein

2. Inheritance

model



template

Summary table

models with inherited ligands

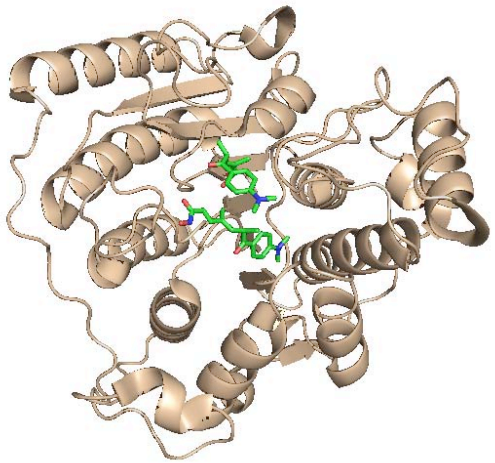
from 16,284 good models, 295 inherited a ligand/substance with at least one compound already approved by FDA and ready to be used from ZINC

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

Example of inheritance (expansion)

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

Template 1t64A a human HDAC8 protein.



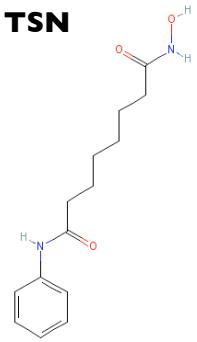
	Origen	Formula	Name	Cov.	Seq, Id. (%)
ZN	X-ray	Zn^{2+}	Zinc ion	--	--
NA	X-ray	Na^{+}	Sodium ion	--	--
CA	X-ray	Ca^{2+}	Calcium ion	--	--
TSN	X-ray	$C_{17}H_{22}N_2O_3$	Trichostatin A	--	--
SHH	Expanded	$C_{14}H_{20}N_2O_3$	Octadenioic acid hudroxyamide phenylamide	100.00	83.8

Example of inheritance (inheritance)

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

	Formula	Name	Cov.	Seq, Id. (%)	Residues
TSN	C_{17} H_{22} N_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



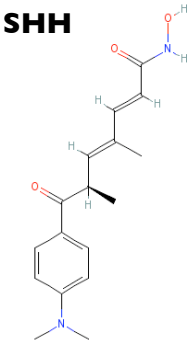
suberoylanilide hydroxamic acid

Pharmacological Action:

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

Inhibits histone deacetylase I and 3

SHH



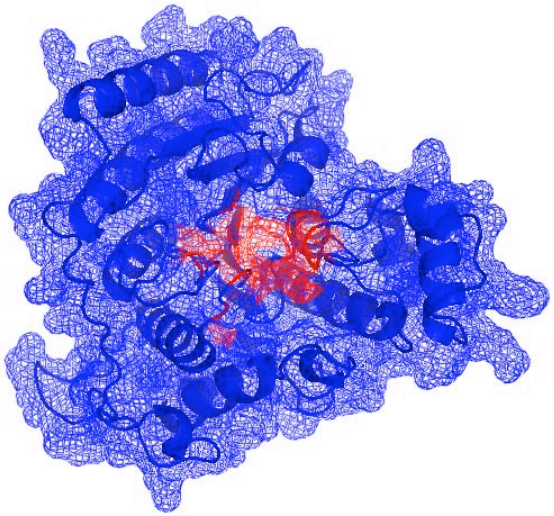
trichostatin A

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*

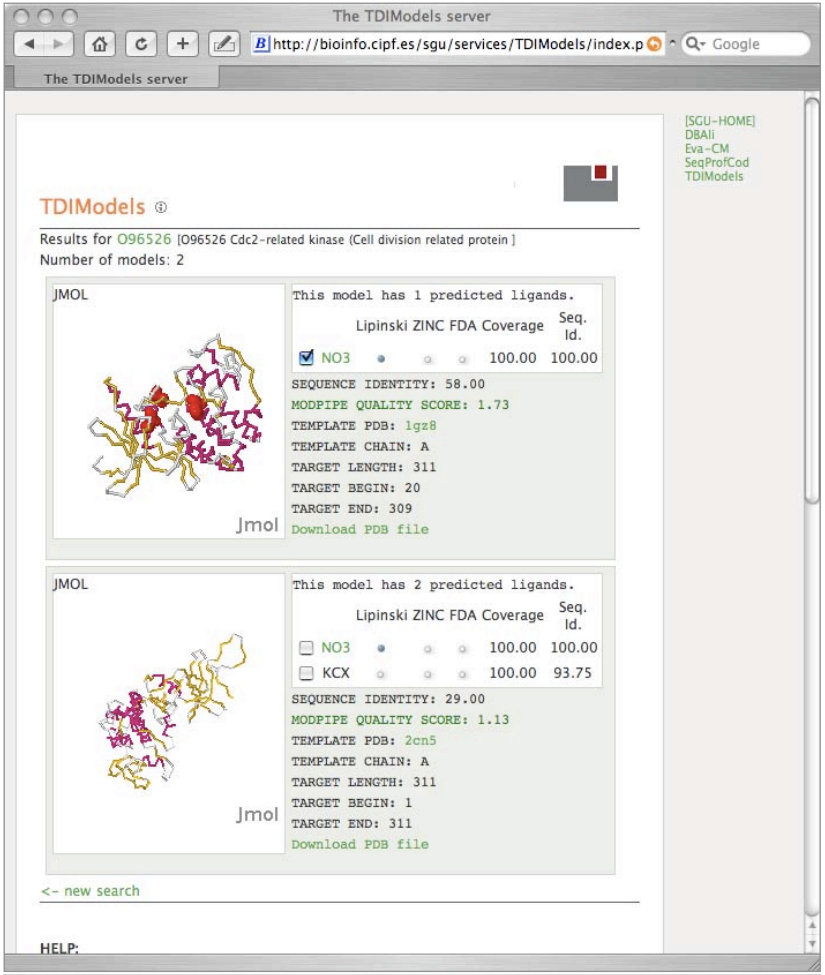
Departments of *Parasite Biochemistry and Cell Biology, ‡Medicinal Chemistry, and §Natural Products Drug Discovery, Merck Research Laboratories, P.O. Box 2000, Rahway, NJ 07065

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

Models database

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



Acknowledgments

Structural Genomics Unit (CIPF)
Marc A. Marti-Renom
Emidio Capriotti
Peio Ziarsolo Areitioaurtena

Comparative Genomics Unit (CIPF)
Hernán Dopazo
Leo Arbiza
François Serra

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

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

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

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

MAMMOTH
Angel R. Ortiz

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

FUNCTIONAL ANNOTATION
Fatima Al-Shahrour
Joaquin Dopazo

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

FUNCTIONAL ANNOTATION
Andrea Rossi
Fred Davis

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