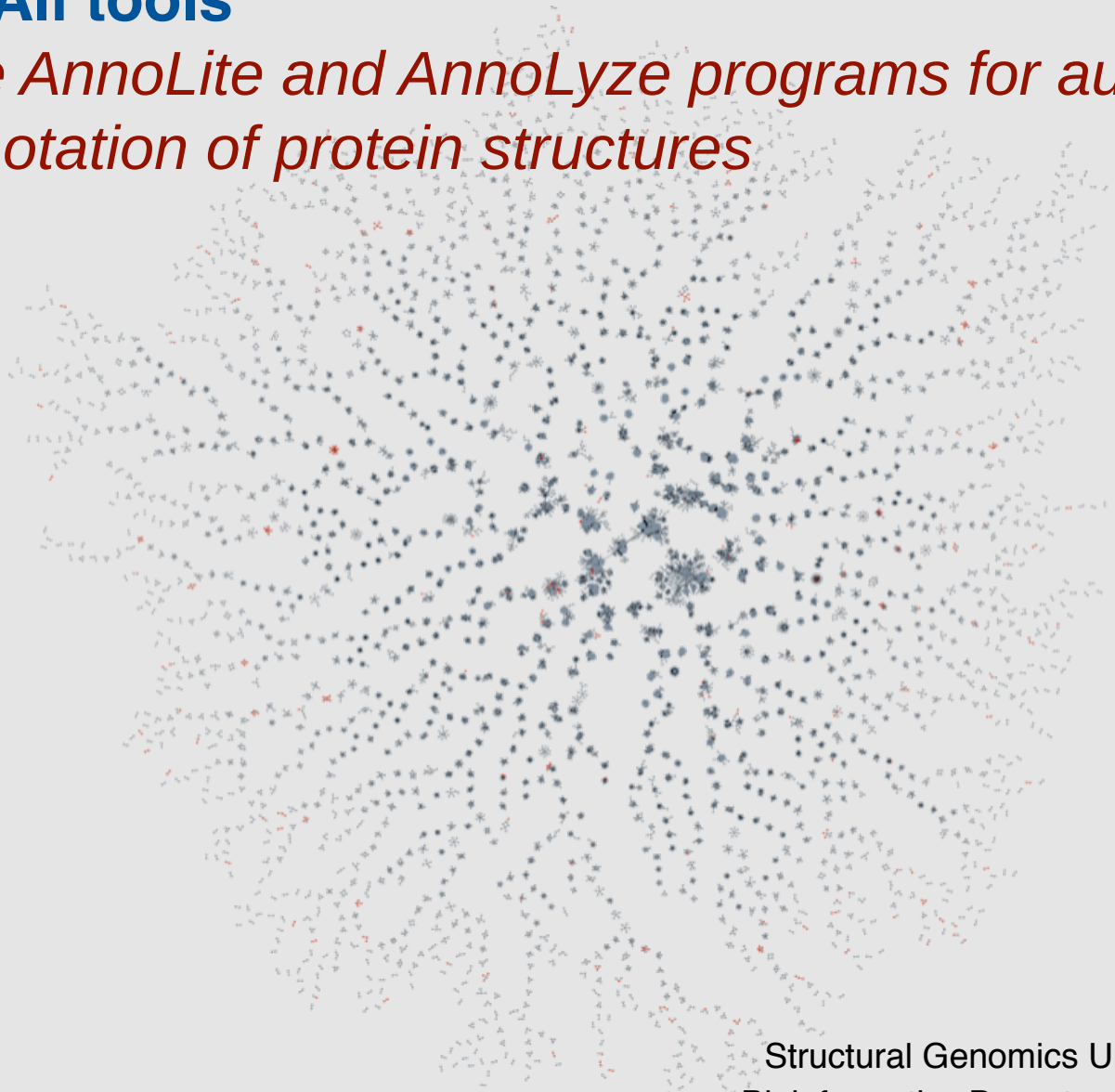
DBAli tools

The AnnoLite and AnnoLyze programs for automatic annotation of protein structures



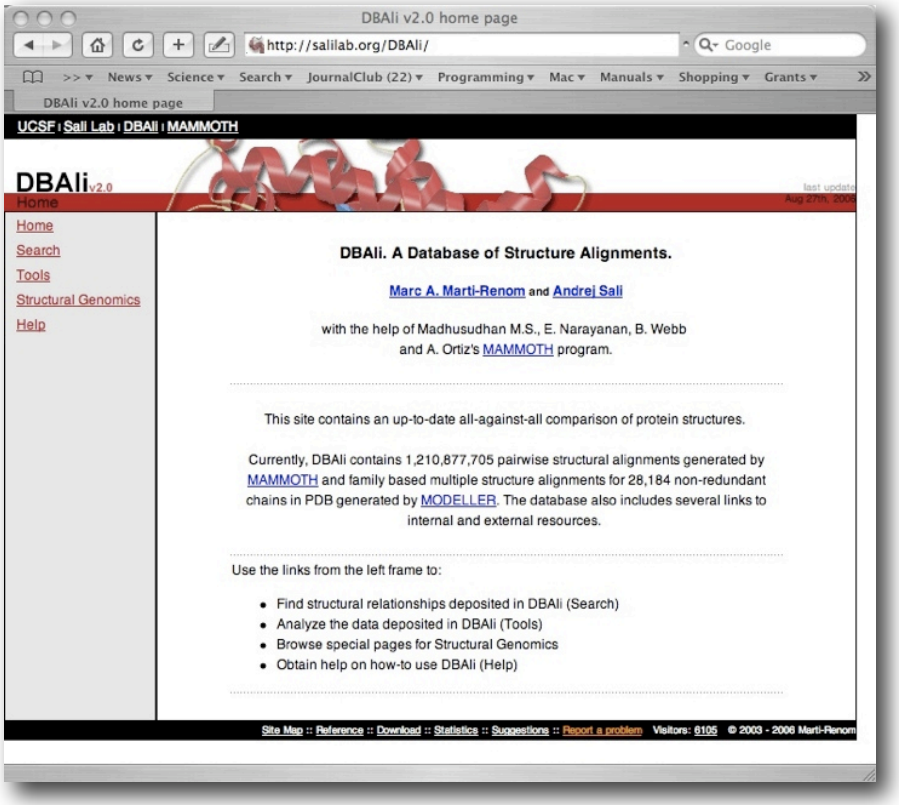
Marc A. Marti-Renom

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

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

DBAli$_{v2.0}$ database

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



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

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

Pairwise structure alignments	
Last update:	November 6th, 2006
Number of chains:	84,180
Number of structure-structure comparisons:*	1,316,585,828
Multiple structure alignments	
Last update:	November 6th, 2006
Number of representative chains:	30,150
Number of families:	11,405

Uses MAMMOTH for similarity detection

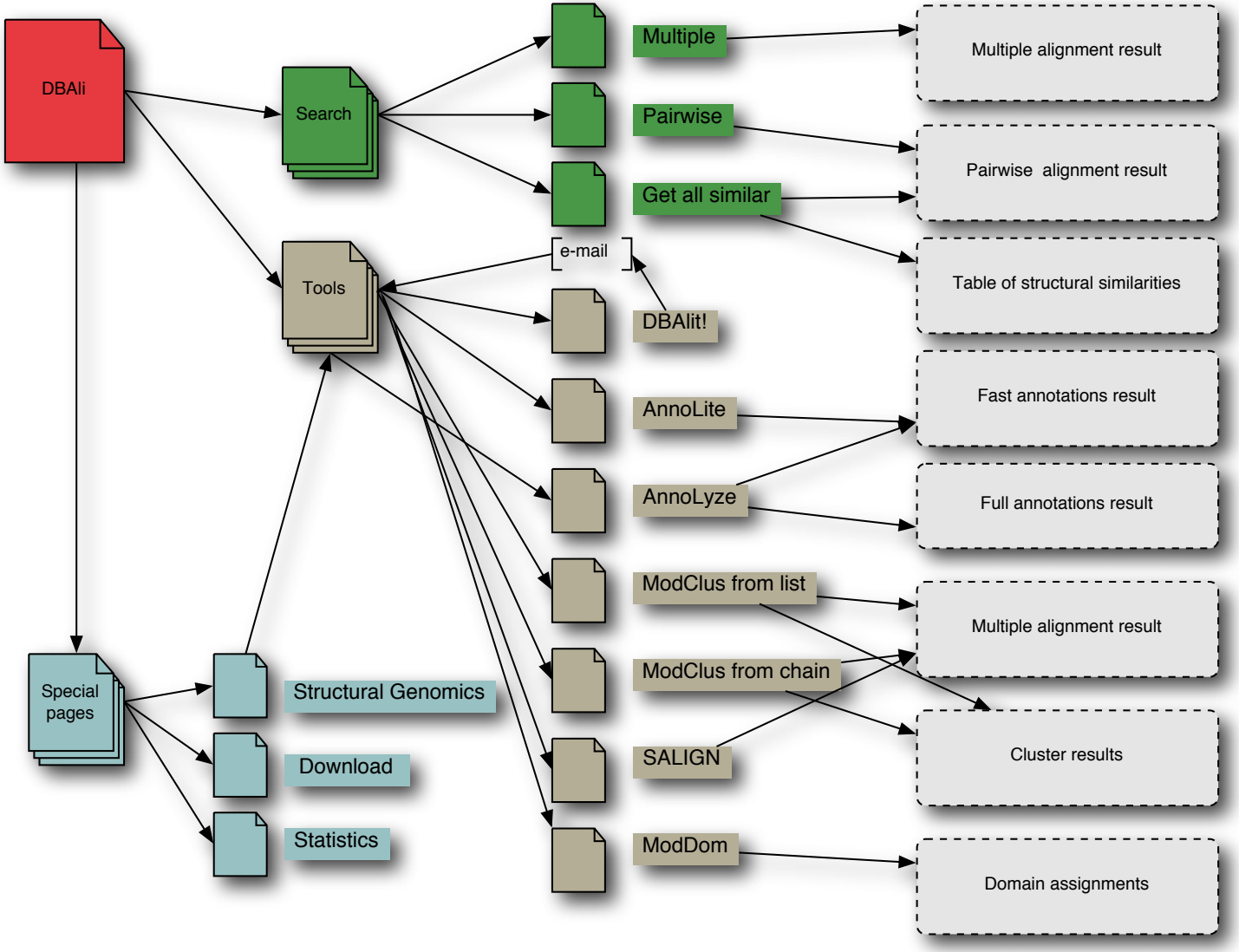
- ✓ **VERY FAST!!!**
- ✓ **Good scoring system with significance**

Ortiz AR, (2002) Protein Sci. 11 pp2606

Marti-Renom et al. 2001. Bioinformatics. 17, 746
Marti-Renom et al. 2006 Submitted.

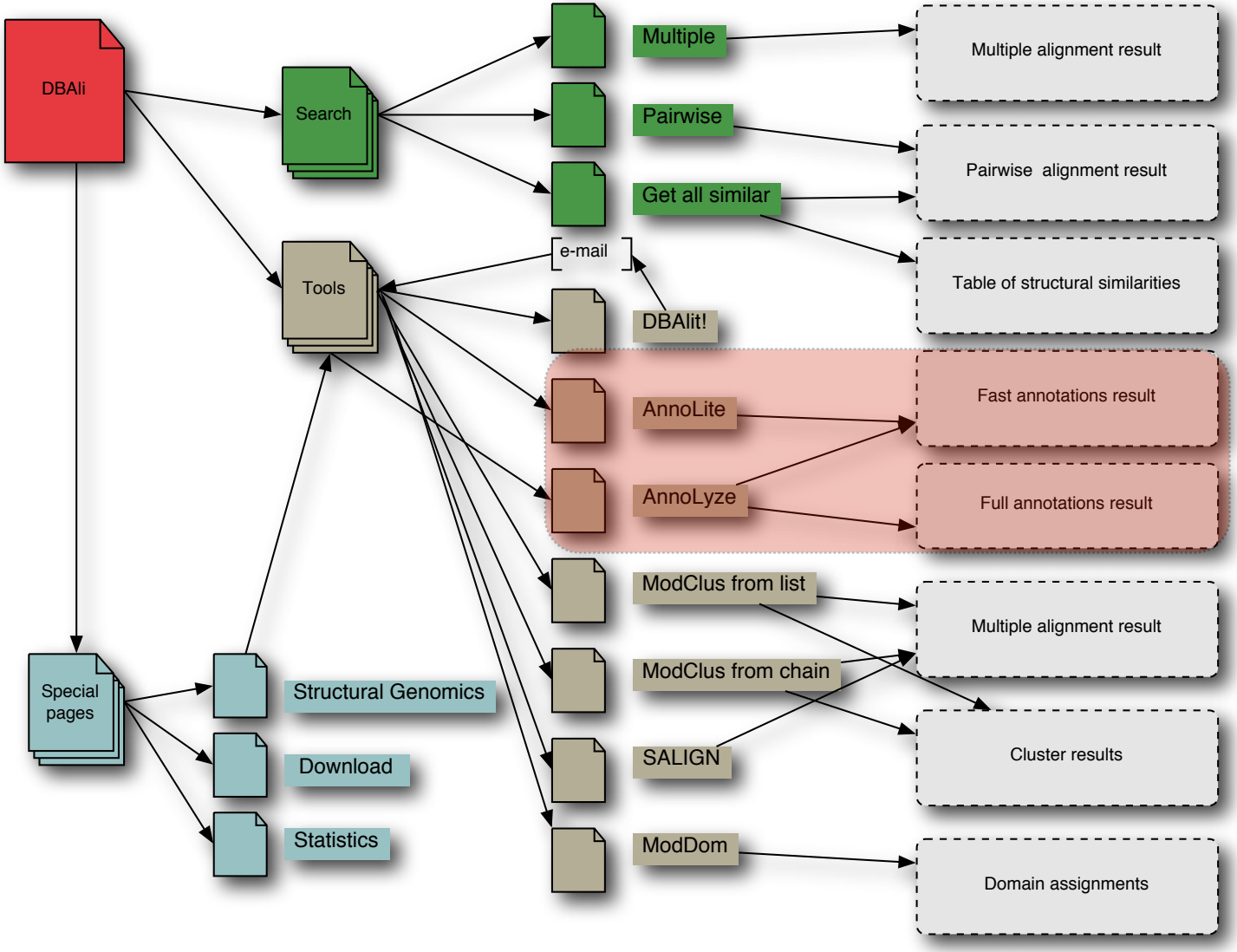
DBAli$_{v2.0}$ database

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



DBAli$_{v2.0}$ database

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



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

AnnoLite

AnnoLite results for chain 1gpl: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

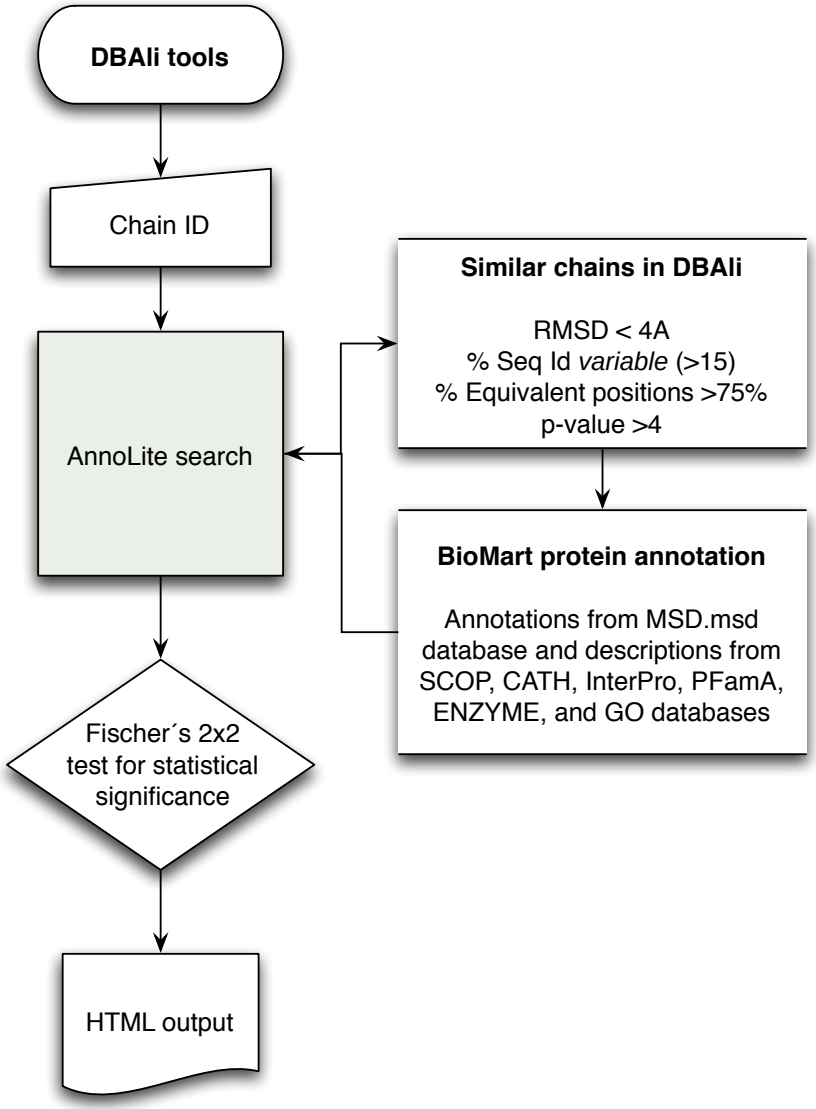
- Information annotated in the MSD database.
- High, medium and low confidence annotations not annotated in the MSD database.
- High, medium and low confidence annotations already annotated in the MSD database.

Benchmark set

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

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

Method



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

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

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

Scoring function

Fisher's 2x2 contingency test

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

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

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

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

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

Sensitivity .vs. Precision

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

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

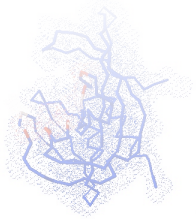
AnnoLyze

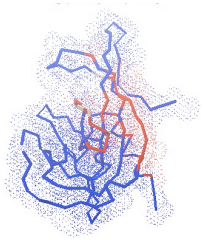
Inherited ligands: 4

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

Inherited partners:1

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





Benchmark

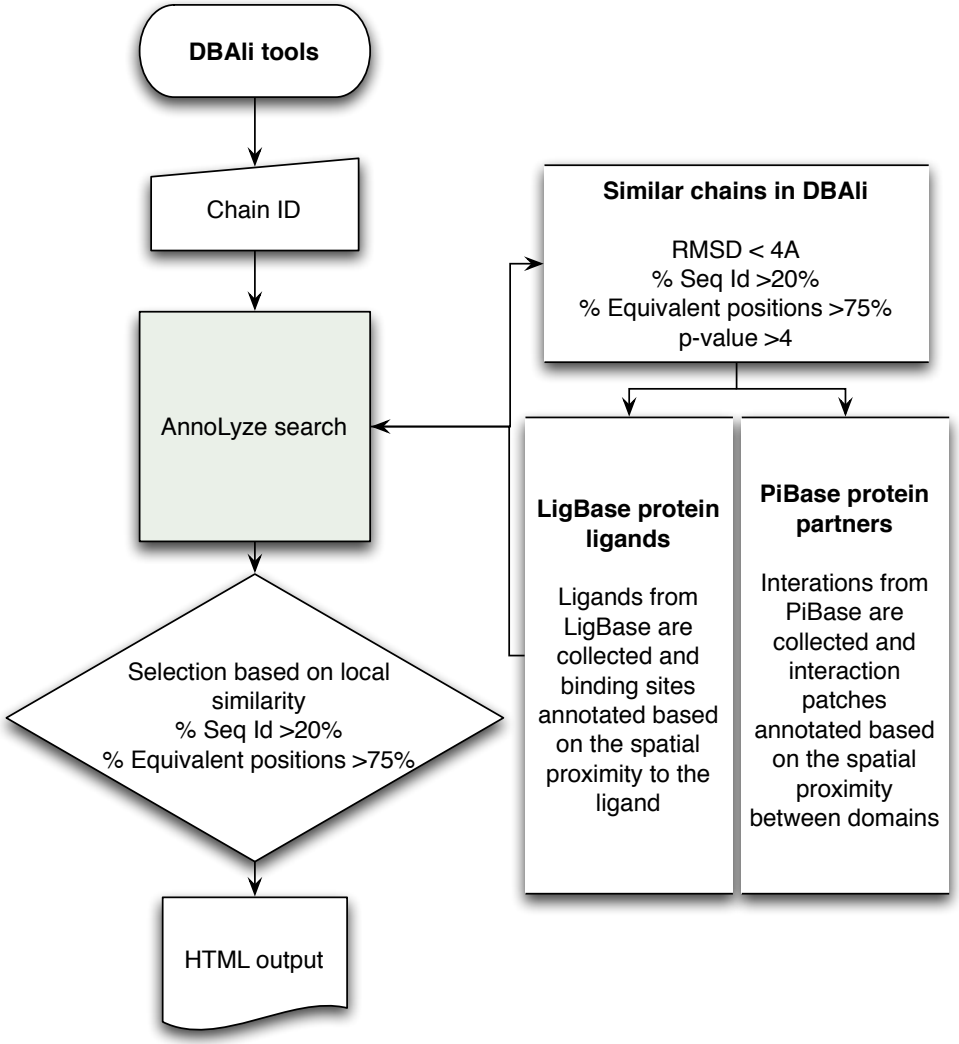
	Number of chains
Initial set*	78,167
LigBase**	30,126
Non-redundant set***	4,948 (8,846 ligands)

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

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

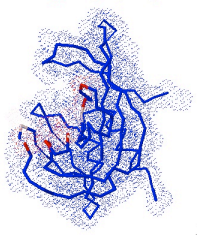
*all PDB chains larger than 30 aminoacids in length (8th of August, 2006)
**annotated with at least one partner in the πBase database
***not two chains can be structurally aligned within 3A, superimposing more than 75% of their Cα atoms, result in a sequence alignment with more than 30% identity, and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

Method



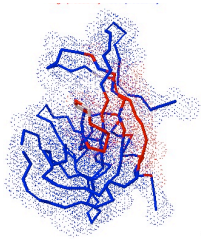
Inherited ligands: 4

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



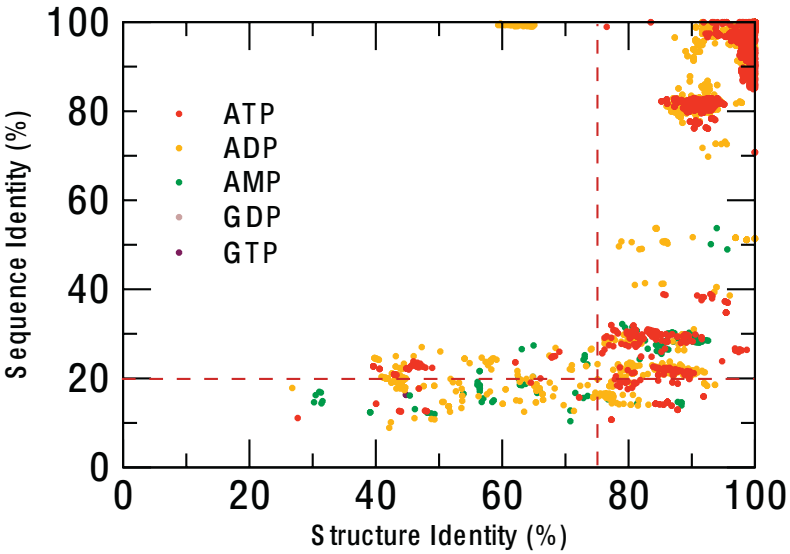
Inherited partners:1

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

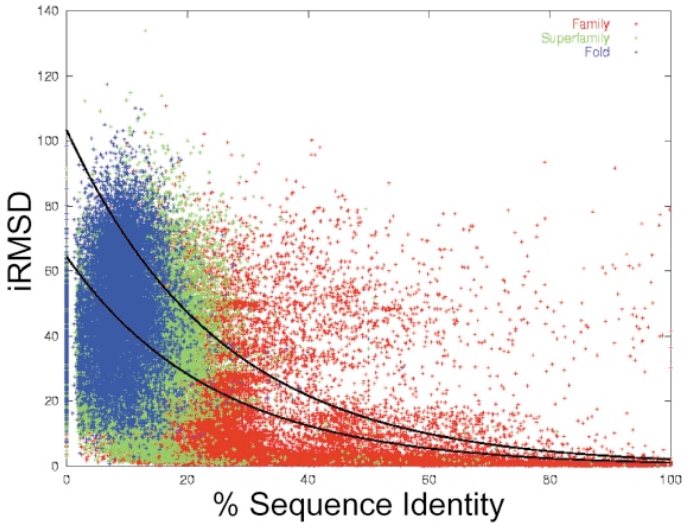


Scoring function

Ligands



Partners



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

Sensitivity .vs. Precision

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

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

Example (2azwA)

Structural Genomics Unknown Function

Molecule: MutT/nudix family protein

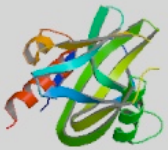
PDB ID: 2azw:A

Header:
STRUCTURAL GENOMICS, UNKNOWN FUNCTION

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

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

Resolution: 1.90Å



Links: none

SCOP: none

CATH: none

Sequence:
Md5: 09b13d23ceae0fdfcaddec636e2ddfa6KTPTAAAS
Length: 146

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

Ligands: none

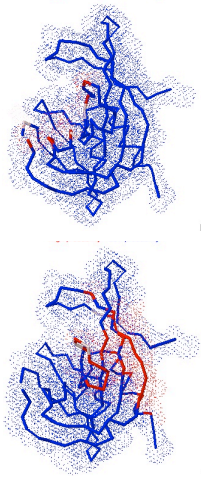
Interacting partners: none

Inherited ligands: 4

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

Inherited partners: 1

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



Similar structures: 20
Similar sequences: 890

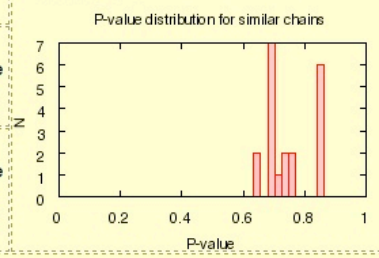
Most similar structure in DBAli:

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

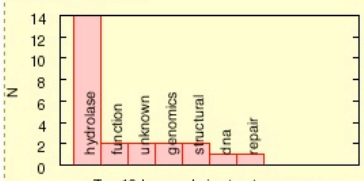
Most similar sequence in DBAli:

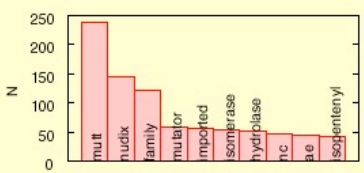
Code	SeqId(%)	EqPos	RMSD	P-Value	See
1vcd:B	24.59	122	3.47	17.00	ali

P-value distribution:

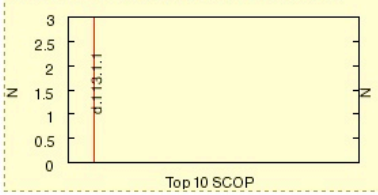


Keyword distribution:



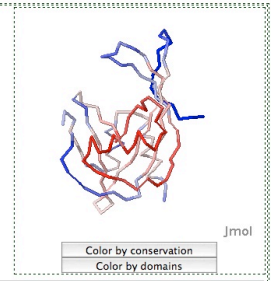


SCOP and CATH distribution for similar structures:



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

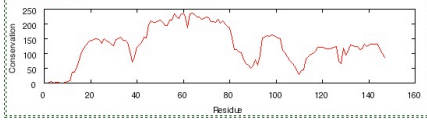
Highest score assignment:
Domain 1:
2-113,115-147
Length: 145
Conservation: 506.49



Protein sequence colored by conservation and underlined by domain color:

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

Conservation profile:



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

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 Krilli (St.George H)

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

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)