Structural Genomics

Iosif Vaisman

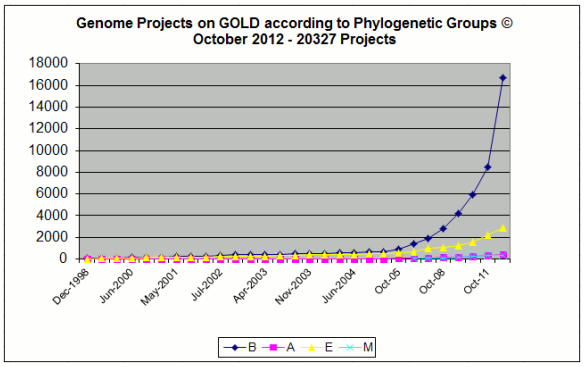
Email: ivaisman@gmu.edu

Genome sequencing projects statistics

Organism	Complete	Draft assembly	In progress	Total
Prokaryotes	1117	966	595	2678
Archaea	100	5	48	153
Bacteria	1017	961	547	2525
Eukaryotes	36	319	294	649
Animals	6	137	106	249
Mammals	3	41	25	69
Birds		3	13	16
Fishes		16	16	32
Insects	2	38	17	57
Roundworms	1	16	11	28
Plants	5	33	80	118
Land plants	3	29	73	105
Fungi	17	107	59	183
Ascomycetes	13	83	38	134
Basidiomycetes	2	16	11	29
Other fungi	2	8	10	20
Protists	8	39	46	93
Apicomplexans	3	11	16	30
Other protists	1	24	28	53
Total:	**1153**	**1285**	**889**	**3327**
Total (2011):	1151	1156	896	3203
Total (2009):	1001	1270	1189	3460

GOLD (Genomes OnLine Database)

ARCHAEA TOTAL: 704	Complete: 447 Permanent Drafts: 71	Draft: 54 In Progress: 95 DNA Received: 63 Awaiting DNA: 0	Targeted: 1
BACTERIA TOTAL: 26956	Complete: 11781 Permanent Draft: 4513	Draft: 2279 In Progress: 11665 DNA Received: 75 Awaiting DNA: 0	Targeted: 424
EUKARYA TOTAL: 6580	Complete: 2050 Permanent Draft: 160	Draft: 607 In Progress: 3509 DNA Received: 1 Awaiting DNA: 0	Targeted: 6



PDB redundancy

Method	Description	# of Clusters	
		2009	2012
blast	100% identity		48584
blast	95% identity	17575	34514
blast	90% identity	16853	33124
blast	70% identity	15114	29705
blast	50% identity	12886	25839
blast	40% identity	11218	22996
blast	30% identity	9294	19676

Sequence-structure correlations

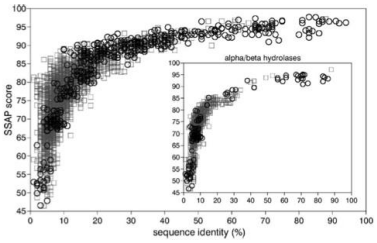
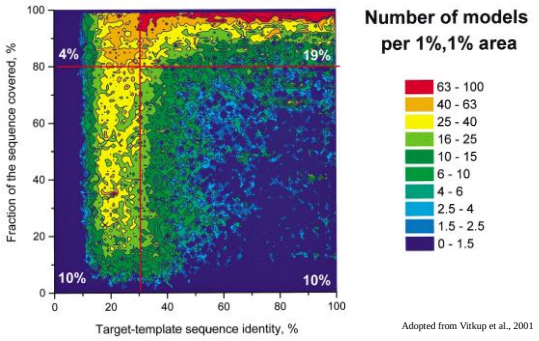


Fig. 1. Correlation between structure similarity (measured by the SSAP structure comparison algorithm, 0–100) and sequence similarity (measured by sequence identity) for all pairs of homologous domain structures in the CATH domain database.

Redfern and Orengo, 2005

Model structure coverage in sequence space



Adopted from Vitkup et al., 2001

Structural Genomics Project

- Organize known protein sequences into families.
- Select family representatives as targets.
- Solve the 3D structure of targets by X-ray crystallography or NMR spectroscopy.
- Build models for other proteins by homology to solved 3D structures.

History of Structural Genomics

1995	**SG** project proposed in Japan
1997 Apr.	**SG** pilot project starts at RIKEN Inst.
1997	**SG** studies initiated through DOE, NIGMS in US
1998/99	Initial **SG** projects start in Canada, Germany, US
1999 June	Call for **SG** pilot projects issued by NIGMS/NIH
2000 Jan.	OECD Committee for Scientific and Technological Policy (CSTP) proposes to initiate **SG** studies
2000 Apr.	1st International **SG** Meeting, Hinxton, UK
2000 June	OECD/Global Science Forum (GSF) and **SG** Workshop, Florence, Italy
2000 Sep.	**SG**: From Gene to Structure to Function, Cambridge, UK
2000 Sep.	NIGMS Protein Structure Initiative starts in US with 7 Centers
2000 Nov.	International Conference on **SG** (ICSG 2000), Yokohama, Japan / International SG Task Force Meeting / OECD/GSF Meeting
2001 Jan.	OECD/CSTP/GSF, Paris, France – Further Study on **SG**
2001 Apr.	2nd International **SG** Meeting, Airlie House, US – Start of ISGO
2001 Sep.	NIGMS Protein Structure Initiative adds 2 new centers
2002 Mar.	European Commission announces funding of Structural Proteomics in Europe (SPINE)
2002 Apr.	National project on Protein Structural and Functional Analyses starts in Japan
2002 Oct.	ISGO International Conference on **SG** (ICSG 2002), Berlin, Germany

Heinemann, 2002

NIGMS Protein Structure Initiative

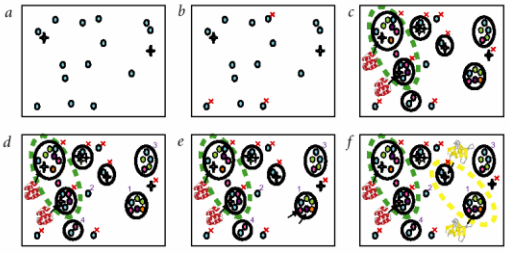
	2001	2002	2003	2004	2013
Selected	11214	21872	42726	74637	323152
Cloned	5465	11277	23237	45353	223094
Expressed	2860	6115	13602	25536	122703
Purified	1505	2823	5291	8398	58315
Crystallized	336	1161	1876	3199	15853
Diffraction	96	438	767	1651	11561
Crystal structure	87	314	545	1260	5919
PDB	76	247	569	1488	9554

Goals of structural genomics

•Provision of enough structural templates to facilitate homology modeling of most proteins
•Structures of all proteins in a complete proteome
•Structural elucidation of a complete biological pathway
•Structural elucidation of a complete disease

Phil Bourne, 2005

Target selection



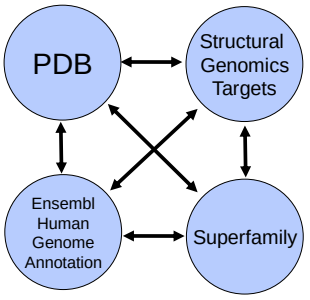
a) realm of interest
b) family exclusion - impossible
c) family exclusion - known
d) prioritization
e) selection
f) analysis and interpretation

S.Brenner, 2000

Target categories

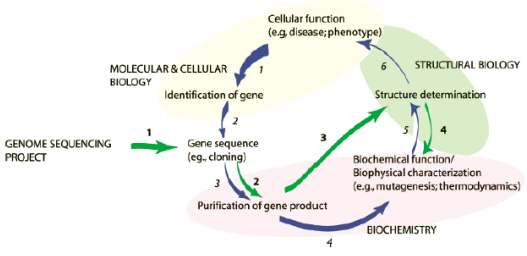
- biomedical
- community nominated
- legacy
- membrane protein
- metagenomic
- structural coverage
- technology development

Coverage of the Human Genome By Structure



Xie and Bourne, 2005

Structural genomics shortcuts



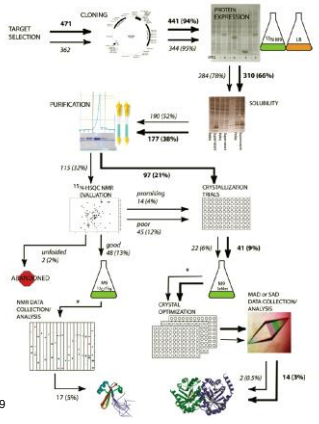
Yee et al., *Acc. Chem. Res.* **2003,** 36, 183-189

Targets by genome

Organism	Number of targets	% Of all targets
Caenorhabditis elegans	4674	17.4
Arabidopsis thaliana	3900	14.5
Homo sapiens	3257	12.1
Pyrococcus furiosus	2179	8.1
Thermotoga maritima	1860	6.9
Mycobacterium tuberculosis	1476	5.5
Escherichia coli	1272	4.7
Saccharomyces cerevisiae	1254	4.7
Bacillus subtilis	1220	4.5
Bacillus stearothermophilus	764	2.8

Adopted from O'Toole et al., 2004

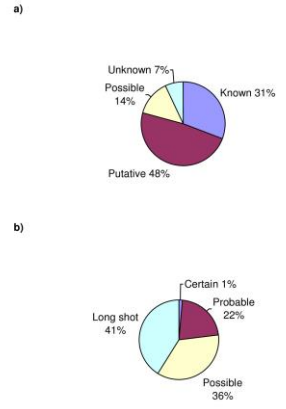
M. thermoautotrophicum structural genomics project



Yee et al., *Acc. Chem. Res.* **2003,** 36, 183-189

Current results

Group or SG center	Targets and nonidentical chains	New Pfam families (total family size)	Novel structures (30% ID)	New SCOP folds	New SCOP fold or superfamily
SG centers					
Berkeley Structural Genomics Center (BSGC)	57 (57 chains)	22 (5757)	41	4	6
Center for Eukaryotic Structural Genomics (CESG)	48 (48 chains)	7 (387)	28	0	0
Joint Center for Structural Genomics (JCSG)	186 (187 chains)	32 (4875)	92	3	4
Midwest Center for Structural Genomics (MCSG)	224 (229 chains)	55 (5512)	163	18	25
Northeast Structural Genomics Consortium (NESGC)	159 (159 chains)	52 (4811)	108	15	26
New York Structural Genomics Research Consortium (NYSGRC)	166 (171 chains)	27 (3982)	90	6	9
Southeast Collaboratory for Structural Genomics (SECSG)	67 (67 chains)	6 (1079)	25	0	1
Structural Genomics of Pathogenic Protozoa Consortium (SGPP)	26 (26 chains)	1 (19)	8	2	2
TB Structural Genomics Consortium (TB)	99 (99 chains)	9 (3938)	42	0	1
PSI centers (total of 9 centers above)	1032 (1043 chains)	211 (30,360)	597	48	74
Japanese center (RIKEN)	686 (718 chains)	50 (6860)	289	10	20
Other international SG (total, excluding all centers above)	169 (183 chains)	33 (5877)	69	6	9
Non-SG groups (since 2000)					
Non-SG structural biology (total)	17,096 (23,747 chains)	928 (249,171)	2,521	269	478
Steitz group	46 (559 chains)	23 (4190)	31	7	12
Huber group	185 (273 chains)	8 (679)	38	5	10
Iwata group	14 (54 chains)	14 (7960)	20	2	3

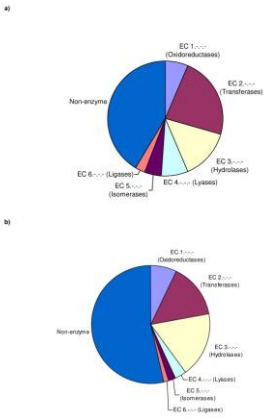


Functional annotation coverage of MCSG structures.

a) Pie chart showing the proportion of MCSG targets manually assigned as having a known, putative, possible, or an unknown function.

b) Pie chart showing the likelihood of the best scoring ProFunc template-search prediction being correct for the targets that are of unknown function following annotation by sequence methods.

D.Lee et al., 2011

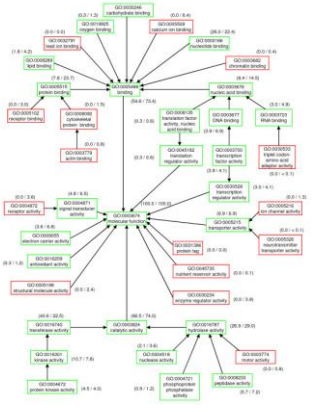


EC classes of MCSG structures compared to the PDB as a whole.

a) Pie chart showing the distribution of EC classes for the MCSG structures that have a known function.

b) Pie chart showing the distribution of EC classes for all PDB entries taken from the Enzyme Structures Database at the EBI

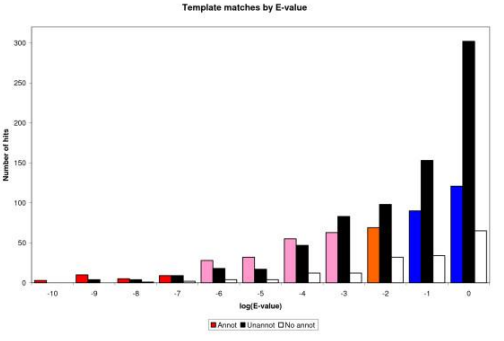
D.Lee et al., 2011



Distribution of molecular function GO terms associated with MCSG structures.

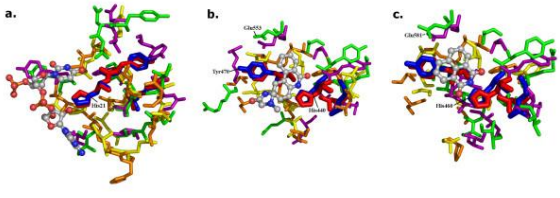
The molecular function ontology terms of the generic GO slim give a broad overview of all protein molecular function categories. Arrows indicate 'is a' relationships. MCSG structures with a known function cover terms in boxes with a green border while terms in boxes with a red border are not covered by MCSG structures. Numbers in parentheses outside of each box show the proportion (%) of MCSG targets with a solved structure and known function that are associated with each term compared to the proportion of all sequence unique PDB structures of known function (MCSG%/PDB%).

D.Lee et al., 2011



Identification of functionally annotated residues as a function of the reverse template E-value.

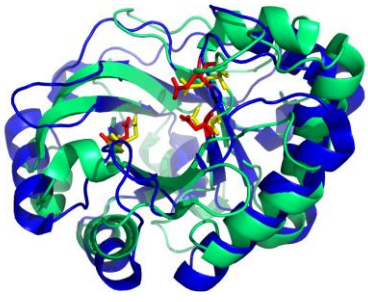
D.Lee et al., 2011



Prediction of function from structure using ProFunc.

Three reverse template matches for PDB entry 2aua, a protein of unknown function from Bacillus cereus. The matches are to the catalytic domains of three toxins: a) diphtheria toxin from *Corynebacterium diphtheriae* (PDB code 1f0l), b) exotoxin A from Pseudomonas aeruginosa (PDB code 1 × k9) and c) cholix toxin from Vibrio cholera (PDB entry 3ess). In each case, the template residues from the 2aua query structure are shown in thick, red sticks while the corresponding residues in the target structure are shown as thick, blue sticks. Neighbouring identical residues, in equivalent 3D positions, are shown in purple for 2aua and green for the target, while similar residues are shown in orange for 2aua and yellow for the target. The inhibitor molecules bound in the target structures are shown in ball-and-stick representation and are: a) adenylyl-3'-5'-phospho-uridine-3'-monophosphate, b) N-(6-oxo-5,6-dihydro-phenanthridin-2-yl)-N,N-dimethylacetamide and c) 1,8-naphthalimide. Catalytic residues are labelled using the residue numbering of the corresponding PDB entries.

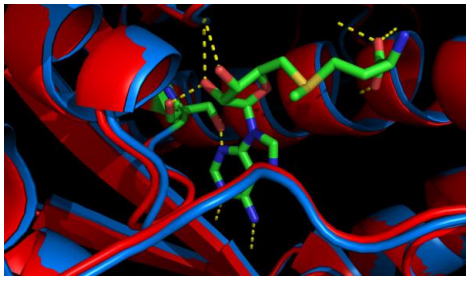
D.Lee et al., 2011



Refining function prediction using ProFunc.

Structural superposition of an uncharacterised protein with a possible functional annotation following sequence analysis (PDB entry 1sfs, in blue) and its top reverse template match, a bacterial muramidase (PDB entry 1jfx, in green). The folds of the two proteins are similar. The residues depicted by yellow sticks are the known catalytic residues in 1jfx (Asp9, Asp98 and Glu100), while the red sticks show the equivalenced residues in 1sfs (Asp9, Asn102 and Glu104).

D.Lee et al., 2011



Explaining the effect of an nsSNP using a homology model based on a MCSG structure.

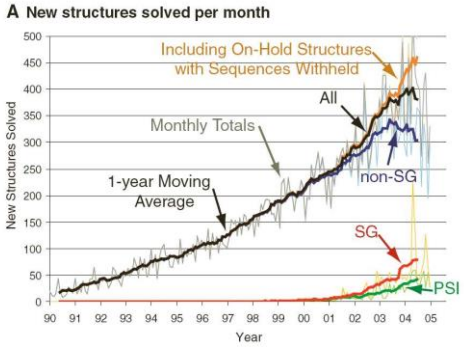
The interaction between S-adenosyl methionine (SAM) and mitochondrial tRNA-specific 2-thiouridylase 1. The Ala10Ser variant probably introduces a hydrogen bond between SAM and the enzyme that increases binding affinity and thus slows down SAM release hence reducing activity. The wild type model is shown in red and the Ala10Ser variant is shown in blue. The variant residue and SAM are coloured according to their atom types and potential hydrogen bonds are shown in yellow.

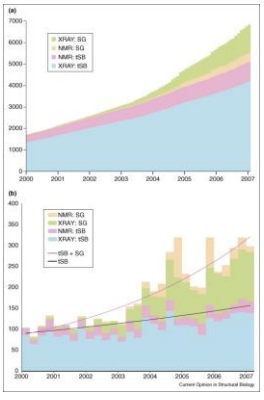
D.Lee et al., 2011

Structural genomics target database

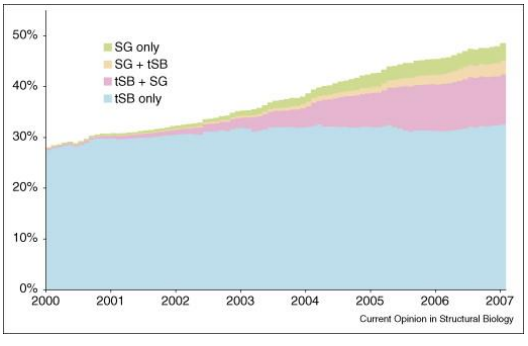
Replaced by
TargetTrack | Structural Biology Target Registration Database
http://sbkb.org/tt/

Current results





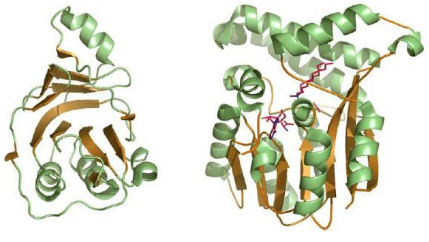
Structural coverage of the Swiss-Prot database

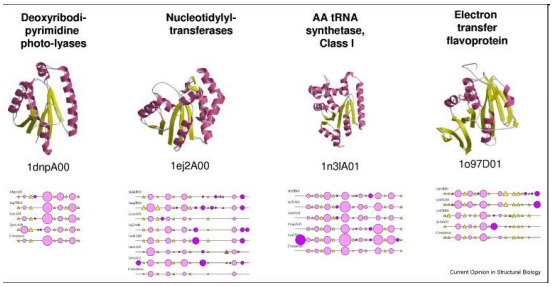


Grabowski et al., 2007

Practical applications of structural genomics

Fig. 3 Mis-annotation of the Rv3853 gene from *M. tuberculosis*. Originally annotated as the terminal SAM-dependent methyltransferase of menaquinone biosynthesis (MenG), Rv3853 has a monomer fold (left) that is completely different from that of a typical SAM-dependent methyltransferase such as CmaA1 (right)





Redfern et al., 2008