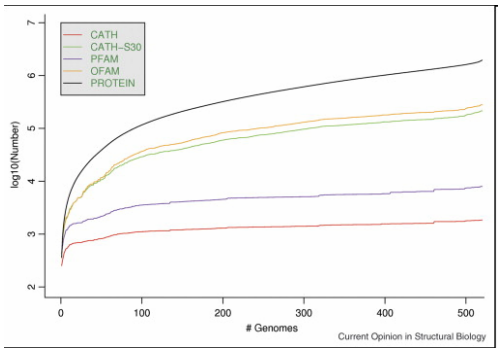
Introduction to Bioinformatics

Iosif Vaisman

Email: ivaisman@gmu.edu

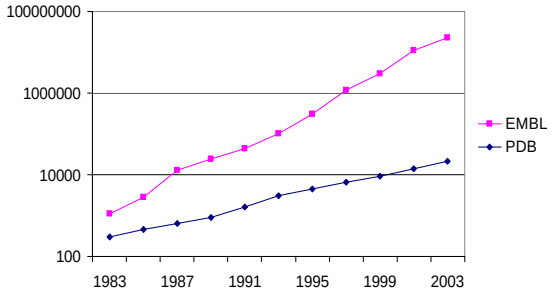
Genome sequencing projects statistics

Organism	Complete	Draft assembly	In progress	total
Prokaryotes	979	1081	1018	3078
Archaea	67	13	38	118
Bacteria	912	1068	980	2960
Eukaryotes	22	189	171	382
Animals	4	75	59	138
Mammals	2	28	17	47
Birds		2	2	4
Fishes		3	6	9
Insects	1	21	7	29
Flatworms		2	2	4
Roundworms	1	9	11	21
Amphibians			1	1
Reptiles			1	1
Other animals		12	14	26
Plants	2	12	45	59
Land plants	2	9	39	50
Green Algae		3	6	9
Fungi	10	76	39	125
Ascomycetes	8	60	27	95
Basidiomycetes	1	10	8	19
Other fungi	1	6	4	11
Protists	6	24	24	54
Apicomplexans	1	11	7	19
Kinetoplasts	1	2	5	8
Other protists	4	10	12	26
total:	1001	1270	1189	3460

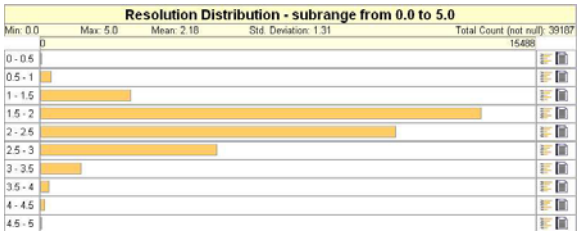


Redfern et al., 2008

Dynamics of Database Growth



PDB resolutions



PDB redundancy

Method	Description	# of Clusters
blast	95% identity (one chain)	17575
blast	90% identity (one chain)	16853
blast	70% identity (one chain)	15114
blast	50% identity (one chain)	12886
blast	40% identity (one chain)	11218
blast	30% identity (one chain)	9294

Sequence-structure correlations

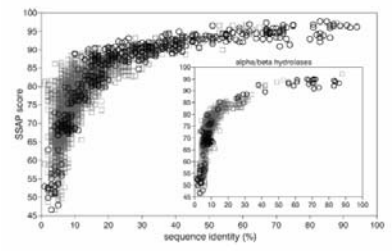
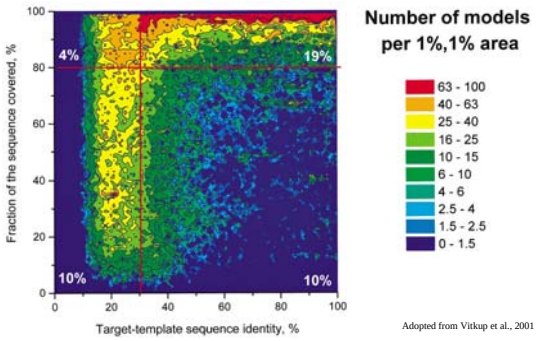


Fig. 1. Correlation between structure similarity (measured by the SSAP structure comparison algorithm, 0–100) and sequence similarity (measured by sequence identity) for all pairs of homologous domain structures in the CATH domain database.

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

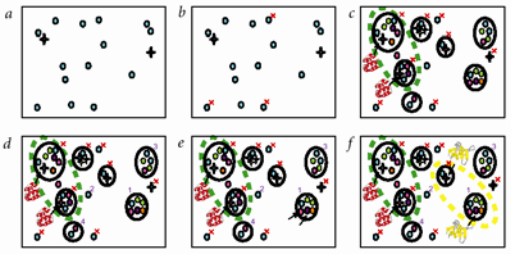
Heinemann, 2002

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

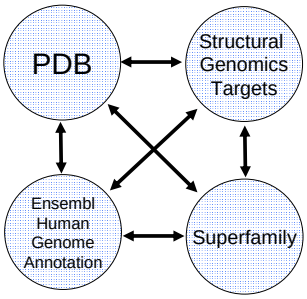
Target selection



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

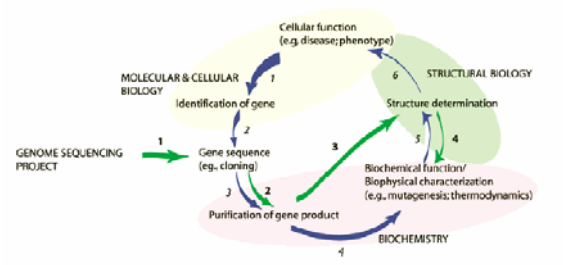
S.Brenner, 2000

Coverage of the Human Genome By Structure



Xie and Bourne, 2005

Structural genomics shortcuts



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

NIGMS Protein Structure Initiative

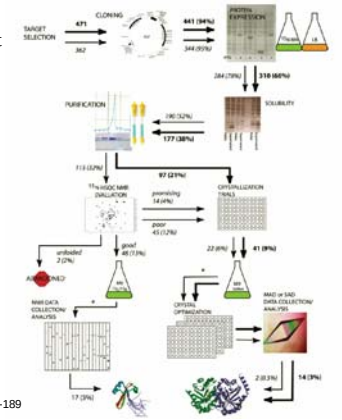
	12/4/2001	12/4/2002	12/4/2003	12/1/2004
Selected	11214	21872	42726	74637
Cloned	5465	11277	23237	45353
Expressed	2860	6115	13602	25536
Purified	1505	2823	5291	8398
Crystallized	336	1161	1876	3199
Diffraction	96	438	767	1651
Crystal structure	87	314	545	1260
PDB	76	247	569	1488

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



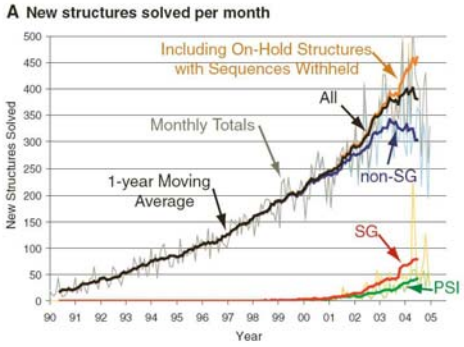
Yee et al., *Acc. Chem. Res.* **2003,** 36, 183-189

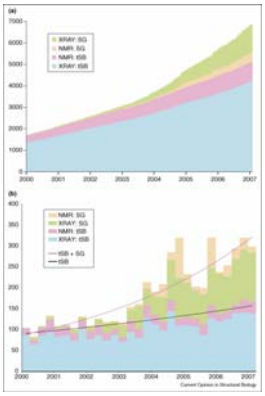
Current results

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 (3079)	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
Stetz 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

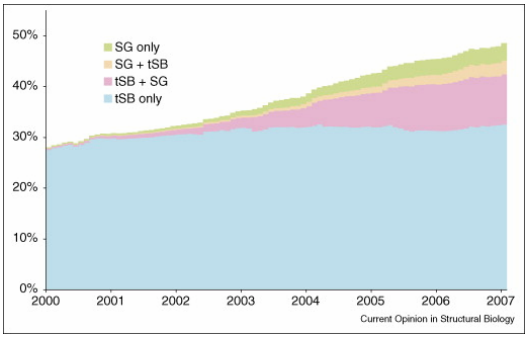
Structural genomics target database

Current results





Structural coverage of the Swiss-Prot database



Grabowski et al., 2007

Practical applications of structural genomics

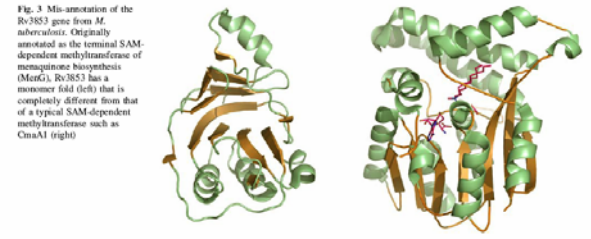
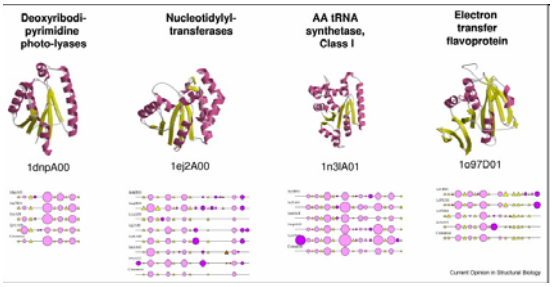


Fig. 3 Mis-annotation of the Rv3853 gene from *M. tuberculosis*. Originally annotated as the terminal SAM-dependent methyltransferase of menaquinone biosynthesis (MenG), Rv3853 has a monomer fold (left) that is completely different from that of a typical SAM-dependent methyltransferase such as CmaA1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



Redfern et al., 2008