Structural Genomics

Iosif Vaisman

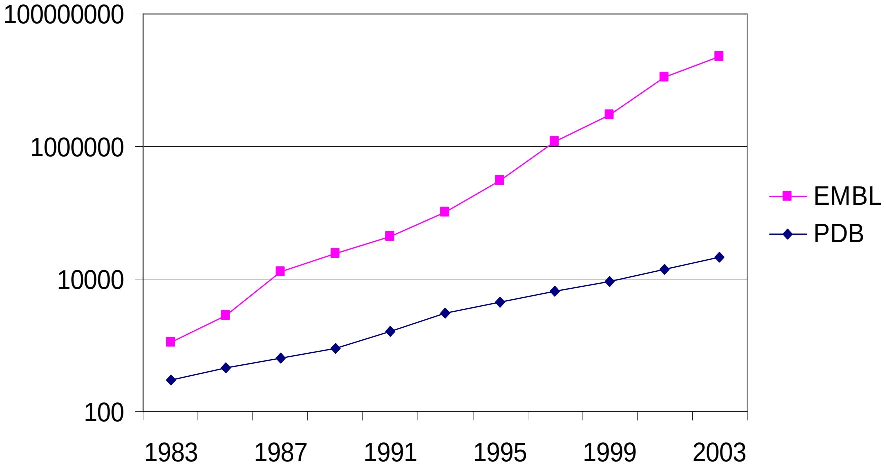
Email: ivaisman@gmu.edu

Protein Folding Problem

Given:	sequence
Find:	structure

The problem is NP-complete

Dynamics of Database Growth



Protein Structure Determination

X-ray crystallography

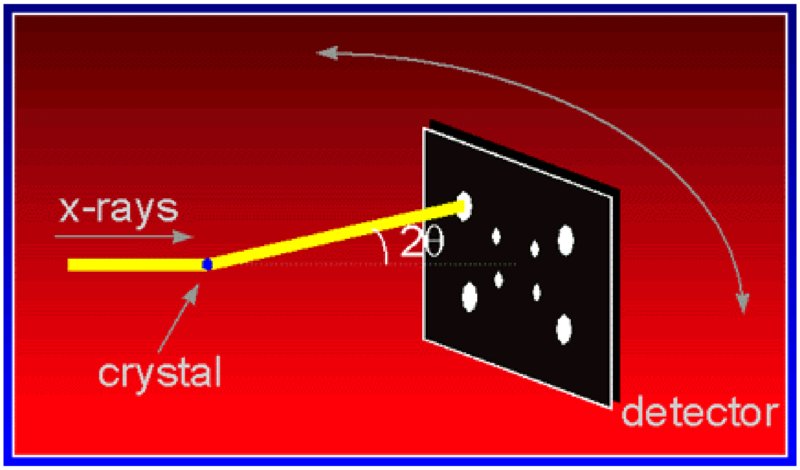
NMR spectroscopy

Neutron diffraction

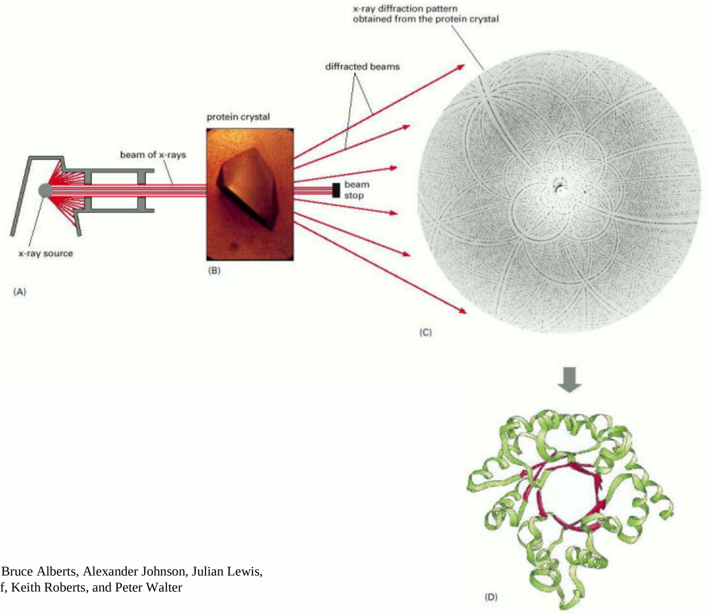
Electron microscopy

Atomic force microscopy

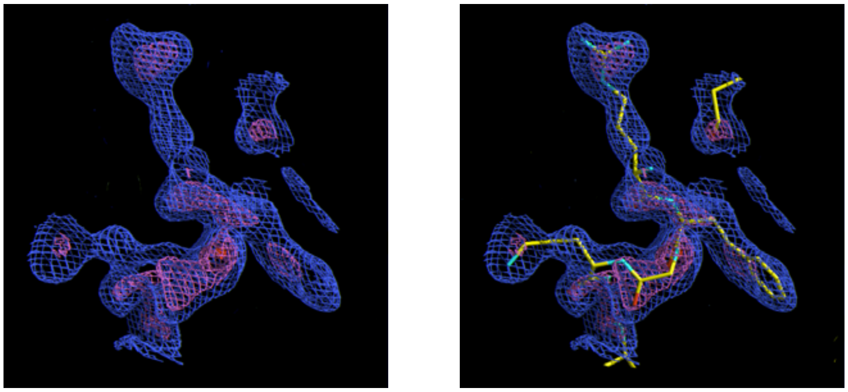
X-ray crystallography



X-ray crystallography

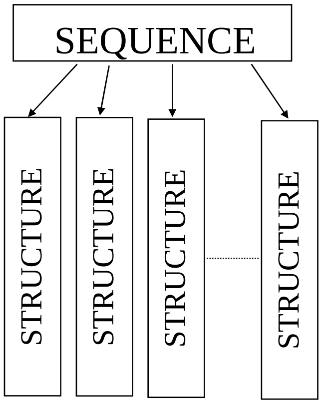


X-ray crystallography



Electron density map created from multi-wavelength data (Arg)

Threading



SEQUENCE

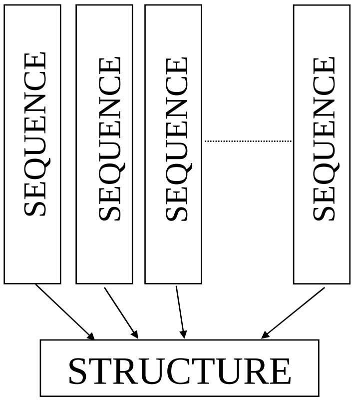
STRUCTURE

STRUCTURE

STRUCTURE

STRUCTURE

Sequence-structure
compatibility
(fold recognition)



SEQUENCE

SEQUENCE

SEQUENCE

SEQUENCE

STRUCTURE

Structure-sequence
compatibility
(inverse folding)

Sequence-structure correlations

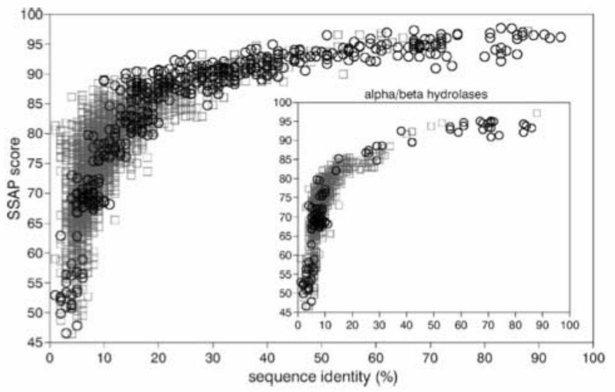
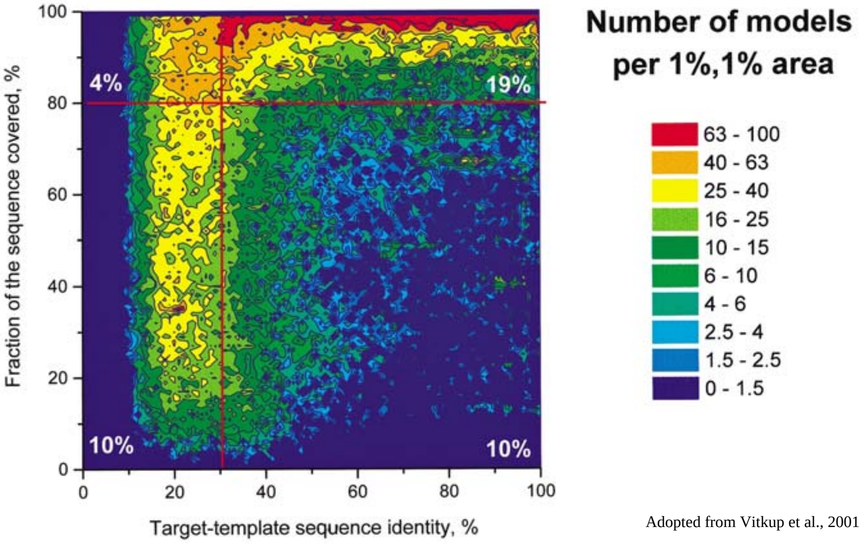


Fig. 1. Correlation between structure similarity (measured by the SSAP structure comparison algorithm, 0–100) and sequence similarity (measured by sequence identity) for all pairs of homologous domain structures in the CATH domain database.

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

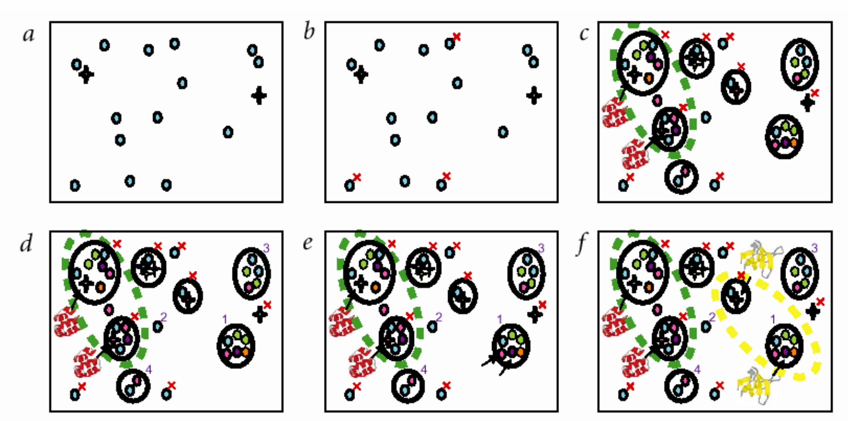
Heinemann, 2002

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

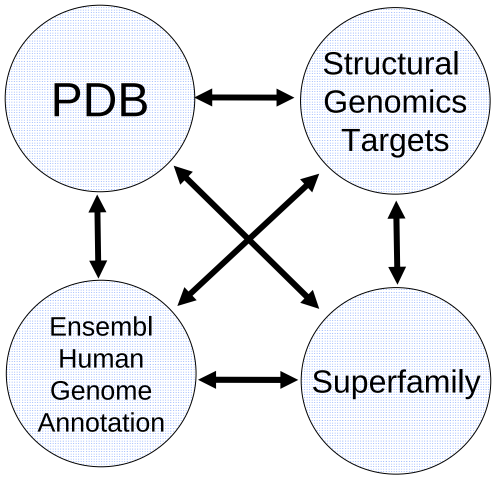
Target selection



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



PDB

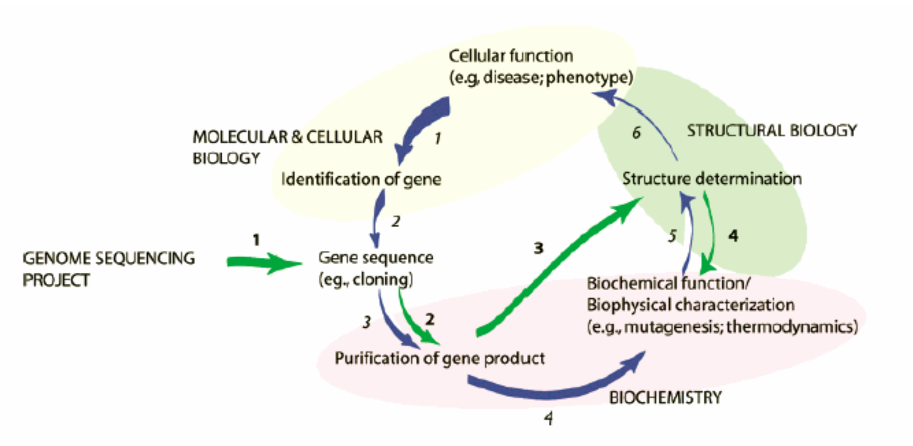
Structural Genomics Targets

Ensembl Human Genome Annotation

Superfamily

Xie and Bourne, 2005

Structural genomics shortcuts

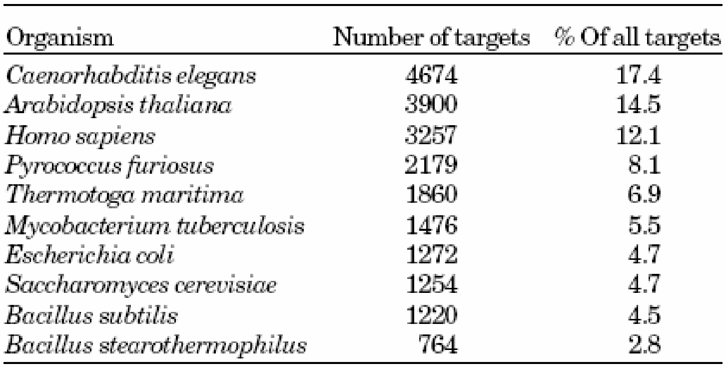


Yee et al., *Acc. Chem. Res.* **2003,** 36, 183-189

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

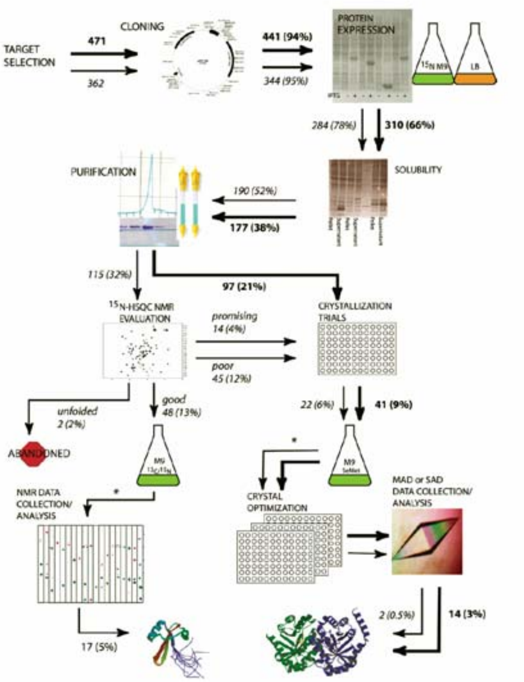
Targets by genome



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

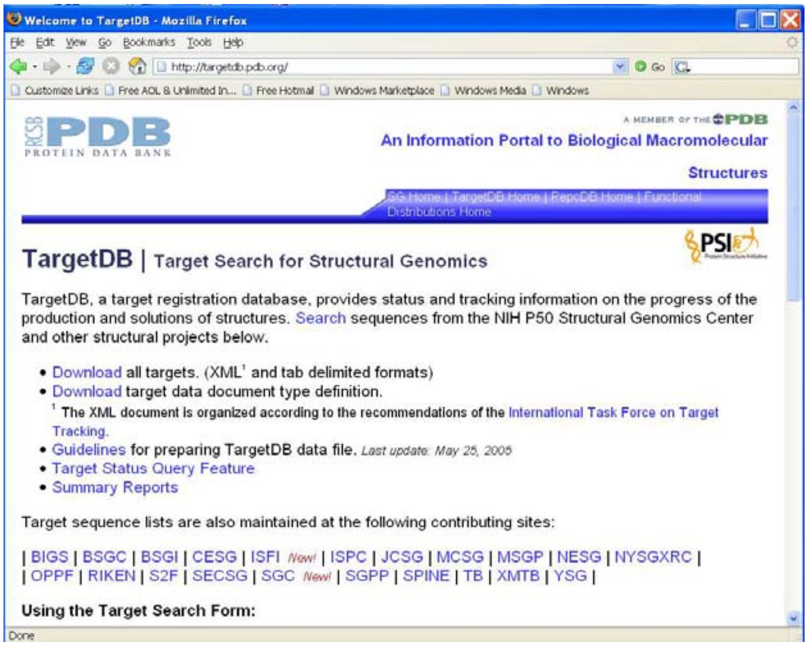
Adopted from O'Toole et al., 2004

M. thermoautotrophicum
structural genomics project



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

Structural genomics target database



An Information Portal to Biological Macromolecular Structures

SG Home | TargetDB Home | PepcDB Home | Functional Distributions Home

TargetDB | **Target Search for Structural Genomics**

TargetDB, a target registration database, provides status and tracking information on the progress of the production and solutions of structures. Search sequences from the NIH P50 Structural Genomics Center and other structural projects below.

- Download all targets. (XML[1] and tab delimited formats)
- Download target data document type definition.
  [1] The XML document is organized according to the recommendations of the International Task Force on Target Tracking.
- Guidelines for preparing TargetDB data file. *Last update: May 25, 2005*
- Target Status Query Feature
- Summary Reports

Target sequence lists are also maintained at the following contributing sites:

| BIGS | BSGC | BSGI | CESG | ISFI *New!* | ISPC | JCSG | MCSG | MSGP | NESG | NYSGXRC |
| OPPF | RIKEN | S2F | SECSG | SGC *New!* | SGPP | SPINE | TB | XMTB | YSG |

Using the Target Search Form: