Structural genomics

Bostjan Kobe
Professor of Structural Biology
SMMS and IMB
Room 76-452, 3365-2132, b.kobe@uq.edu.au

Content:

- Protein function depends on its structure
- What is structural genomics
- Protein structure classification
  - SCOP, CATH, FSSP/DALI
  - Overview of protein folds
- Structural genomics
  - Steps
  - Target selection
  - Expected benefits/limitations
  - Current scope
  - Structure to function
    - Examples

* *Nature Struct Biol, Structural Genomics Supplement, November 2000*

3D structure of proteins

- 3D structure of a protein is determined by its amino acid sequence
- Protein function depends on its structure

Structural genomics

- A systematic program of 3D structure determination aimed at developing a comprehensive view of protein structure universe
  - Experimentally determine representative protein structures
    - X-ray crystallography
    - NMR spectroscopy
  - Computationally predict remaining protein structures
    - Comparative modelling
- Goal: infer functional information

Protein structure classification

- Hierarchical organization
  - SCOP: Structural Classification of Proteins (Murzin et al.)
    - http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.1.html
  - CATH: Class Architecture Topology Homology (Thornton et al.)
    - http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html
  - Class: α, β, α/β, α+β, little secondary structure...
  - Fold
    - ~1000-5000 different folds expected
  - Family: significant sequence similarity (>30%)
    - Superfamily: families with functional similarities
- Automated geometrical comparison
  - FSSP: Families of Structurally Similar Proteins (Sander et al.)
    - http://www2.ebi.ac.uk/dali/fssp/

SCOP: Structural Classification of Proteins

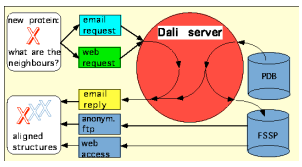
Murzin et al (1995). *J. Mol. Biol.* 247, 536-540.

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	226	392	645
All beta proteins	149	300	594
Alpha and beta proteins (a/b)	134	221	661
Alpha and beta proteins (a+b)	286	424	753
Multi-domain proteins	48	48	64
Membrane and cell surface proteins	49	90	101
Small proteins	79	114	186
Total	971	1589	3004

FSSP: Fold Classification based on Structure -Structure Alignment of Proteins

Holm et al. Protein Science 1, 1691-1698.

- FSSP database based on exhaustive all-against-all 3D structure comparison of protein structures in PDB
- The classification and alignments automatically maintained and continuously updated using the Dali search engine



- **DALI method**
  - **3D structures are represented as Ca-Ca distance matrix. Similarity in terms of equivalent intramolecular distances is optimized.**
  - **Similarity score expressed in terms of statistical significance**
    - **Z = standard deviations above that expected. Z < 2.0 means no significant similarity.**

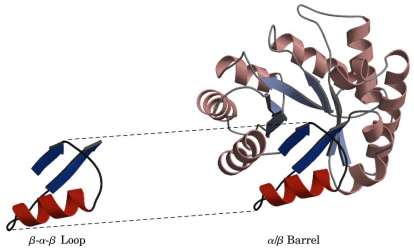
OUTPUT FROM DALI

STRID2 Z RMSD LALI LSEQ2 %IDE PROTEIN

1bk5A 61.5 0.0 422 422 100 karyopherin alpha fragment (importin alpha, srp1p)
1bk5B 58.6 0.4 422 422 100 karyopherin alpha fragment (importin alpha, srp1p)
1bk6A 54.5 0.8 422 422 99 karyopherin alpha fragment (importin alpha, srp1p) biol
1bk6B 54.5 0.8 422 422 99 karyopherin alpha fragment (importin alpha, srp1p) biol
1ialA 47.0 2.0 412 438 48 importin alpha (karyopherin alpha) biological_unit
3bct 34.1 3.8 395 457 17 beta-catenin fragment
1ee4A 33.1 2.3 354 423 24 karyopherin alpha fragment (serine-rich RNA polymerase
1qgrA 19.6 10.4 386 871 14 importin beta subunit (karyopherin beta-1, nuclear fact
1b3uA 15.7 11.1 363 588 14 protein phosphatase pp2a fragment
1qbkB 13.6 9.1 350 879 11 karyopherin beta2 fragment ran fragment
1lrv 11.1 8.2 221 233 11 leucine-rich repeat variant (lrv) biological_unit

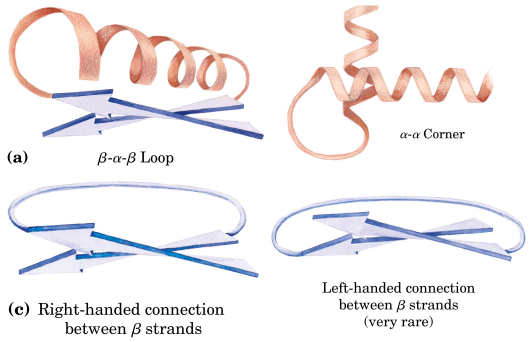
Protein fold

- **A specific combination of smaller supersecondary structure motifs**



β-α-β Loop α/β Barrel

Supersecondary structure motifs



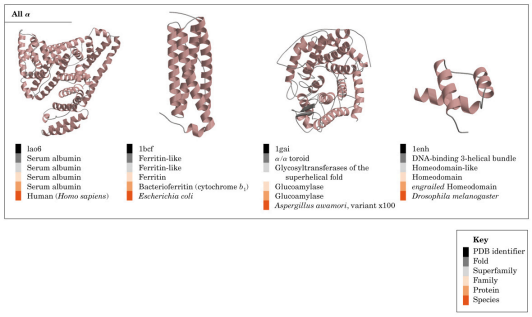
(a) β-α-β Loop

α-α Corner

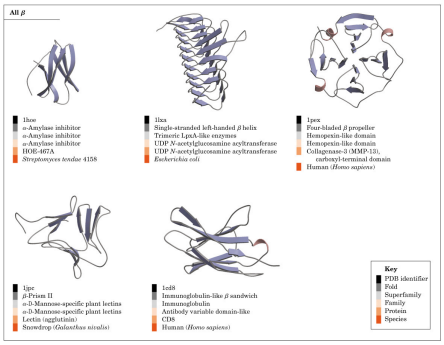
(c) Right-handed connection between β strands

Left-handed connection between β strands (very rare)

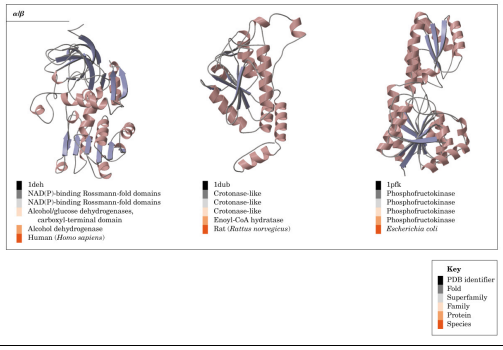
Examples of protein structure (1)



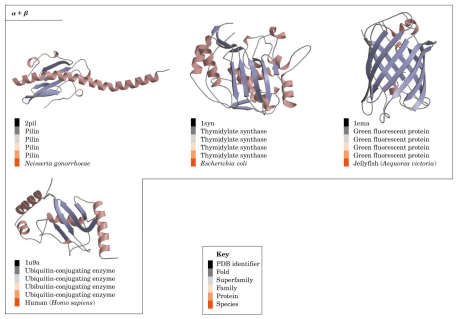
Examples of protein structure (2)



Examples of protein structure (3)

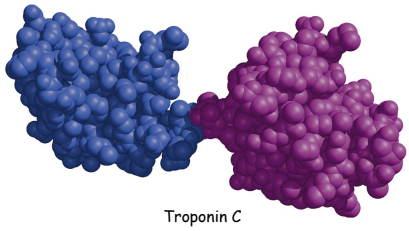


Examples of protein structure (4)



Domains

- Independent globular folding units



Troponin C

Protein structure universe

- 1,000-5,000 distinct protein folds predicted
  - PDB currently contains ~970 distinct folds
- Each new structure enables modelling of 15-40 sequences (>30-35% identity)
  - Yeast genome: portions of 50% sequences can be modelled (18% all residues in yeast proteins)
  - 10,000-20,000 templates needed to model all proteins

Structural genomics: how can it be done?

- High throughput
  - X-ray crystallography
  - NMR spectroscopy
  - Comparative modelling
- Integrative database
  - Structure classification
  - Link data with genome information (phylogenetic occurrence, protein function, gene expression, protein-protein interactions)

Structural genomics: steps

1. PCR amplification of coding sequence
2. Cloning coding sequence into expression vector
   - E.g. His-tag
   - Sequencing cloned gene for verification
3. Protein expression and purification
4. Characterization of expressed protein
5. Defining suitable crystallization/NMR solution conditions
6. X-ray/NMR measurement
7. Structure determination and refinement
8. Comparative structure modelling with the new template
9. Making functional inferences

Automation developed in all steps

Structural genomics: target selection

- Unknown structure
- Tractable
- Prioritization

1. Realm identification
   - E.g. selected organism, cell type, signalling protein...
2. Family exclusion: cluster into families using sequence analysis
   - BLAST, PSI-BLAST, HMMs; COGs, Pfam
   - Difficult or impossible to study
   - Known structure
3. Family prioritization
   - E.g. taxonomically dispersed, large family...
   - Experimental target selection
4. Protein/region selection
   - Desirable characteristics: size, thermostability, # Met

Protein production and purification

Structural Genomics Consortium[1–3], Architecture et Fonction des Macromolécules Biologiques[4], Berkeley Structural Genomics Center[5], China Structural Genomics Consortium[6,7], Integrated Center for Structure and Function Innovation[8], Israel Structural Proteomics Center[9], Joint Center for Structural Genomics[10,11], Midwest Center for Structural Genomics[12], New York Structural GenomiX Research Center for Structural Genomics[13–17], Northeast Structural Genomics Consortium[18,19], Oxford Protein Production Facility[20], Protein Sample Production Facility, Max Delbrück Center for Molecular Medicine[21], RIKEN Structural Genomics/ Proteomics Initiative[22] & SPINE2-Complexes[23,25]

In selecting a method to produce a recombinant protein, a researcher is faced with a bewildering array of choices as to where to start. To facilitate decision-making, we describe a consensus 'what to try first' strategy based on our collective analysis of the expression and purification of over 10,000 different proteins. This review presents methods that could be applied at the outset of any project, a prioritized list of alternate strategies and a list of pitfalls that trip many new investigators.

NATURE METHODS | VOL.5 NO.2 | FEBRUARY 2008 | 135

Structural genomics: expected benefits

- Infer function
  - Generate hypotheses
  - Test experimentally
    - Site-directed mutagenesis
    - Ligand binding studies
    - Enzyme assays
    - Protein-protein interaction studies
- Medically relevant proteins: disease-oriented research
  - Templates for drug design
  - Protein pharmaceuticals
- Source of reagents
- Method development

Structural genomics: limitations

- Some proteins will not express, crystallize...
  - Post-translational modifications, cofactors
  - → Choose another member of the family
- Membrane proteins
  - Technical challenge
- Proteins from macromolecular complexes
  - Unstable in isolation
- Low complexity regions
  - Unstructured
- Regulation, protein-protein interactions, conformational changes
  - Not addressed

Structural genomics: current scope

- USA/North America
  - 4 Production + 6 Specialized PSI-2 consortia
- Europe
  - Several initiatives organized as SPINE
- Japan + Asia
  - RIKEN
- Commercial sector
  - Target pharmaceutical customers

USA

Large-Scale Centers

- Joint Center for Structural Genomics
- Midwest Center for Structural Genomics
- New York Structural GenomiX Research Consortium
- Northeast Structural Genomics Consortium

Specialized Centers

- Accelerated Technologies Center for Gene to 3D Structure
- Center for Eukaryotic Structural Genomics
- Center for High-Throughput Structural Biology
- Center for Structures of Membrane Proteins
- Integrated Center for Structure and Function Innovation
- New York Consortium on Membrane Protein Structure

Homology Modeling Centers

- Joint Center for Molecular Modeling
- New Methods for High-Resolution Comparative Modeling

Resource Centers

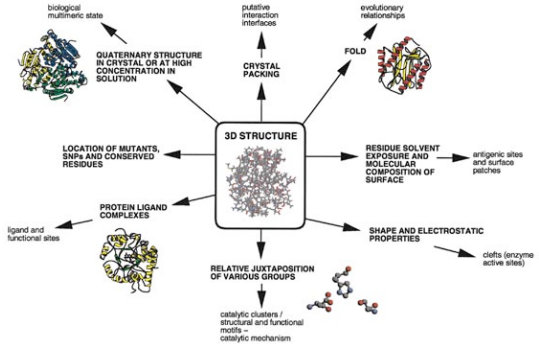
- PSI Materials Repository
- PSI Knowledgebase

Table 1 Current players in commercial structural genomics

Company name	Year founded	Location	Technology	URL
Experimental companies				
Astex	1998	Cambridge, UK	High throughput X-ray crystallography/focus on co-complexes	www.astex-technology.com
Integrative Proteomics	2000	Toronto, Canada	Automation for protein expression	www.integrativeproteomics.com
Structure-Function Genomics	1999	Piscataway, NJ	NMR, protein domain analysis and expression	www.monmouth.com/~spidersigns/SFG/index.html
Structural GenomiX	1999	San Diego, CA	High throughput X-ray crystallography and compound design	www.stromix.com
Syrrx	1999	La Jolla, CA	High throughput X-ray crystallography	www.syrrx.com
Modeling companies				
IBM (Blue Gene project)	2000		Computational protein folding	http://www.ibm.com/news/1999/12/06.phtml
Inpharmatica	1998	London, UK	Biopendium database	www.inpharmatica.com
Geneformatics	1999	San Diego, CA	'Fuzzy functional form' modeling for identifying active sites	www.geneformatics.com
Prospect Genomics	1999	San Francisco, CA	Homology modeling	not available
Protein Pathways	1999	Los Angeles, CA	Phylogenetic profiling, domain analysis, expression profiling	www.proteinpathways.com
Structural Bioinformatics	1996	San Diego, CA and Copenhagen, Denmark	Homology modeling, docking	www.strubix.com

From structure to function

- **Biochemical (molecular) function**
  - Possible to infer from structure in favorable cases
- **Biological (cellular) role (function)**
  - Requires additional data: expression, localization



From structure to function

- Comparison of structure with available structures
  - Structure is better conserved than sequence: can detect distant evolutionary relationships
  - E.g. DALI http://www2.ebi.ac.uk/dali
- Local structural motifs
  - E.g. helix-loop-helix binds DNA, EF hand binds Ca^{2+}, catalytic triad in proteinases
- Ab initio prediction of function
  - Active sites in clefts
  - Patch analysis or crystal packing to identify protein -protein interfaces
  - E.g. ProFunc http://www.ebi.ac.uk/thornton-srv /databases/ProFunc/
- Combine with other experimental data

Statistics from structural genomics

- 42 structures from structural genomics initiatives
  - 12 new fold
  - Functional information inferred for 75%
  - Additional new functions can be identified for proteins with "known" function

Source: Teichmann et al. (2001), Curr. Opin. Struct. Biol. 1, 354

Mj0226, M. jannaschii (Hwang KY et al (1999) Nature Struct Biol 6, 691)

- Partial structural similarity to nucleotide-binding proteins
- Biochemical analysis shows it is nucleotide triphosphatase

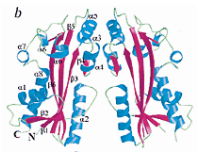
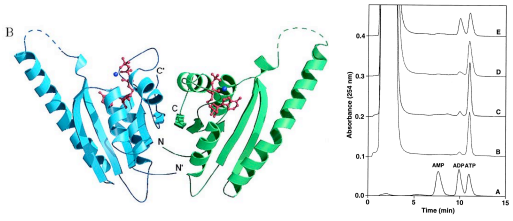


Table 2 Kinetic parameters of Mj0226 with various nucleotides

	K_{cat} (s^{-1})	K_m (mM)	K_{cat} / K_m
XTP	1009.37	0.10	10195.66
ITP	911.72	0.15	5998.16
GTP	97.65	1.11	87.66
dGTP	96.64	1.13	85.52
ATP	1.02	7.04	0.15
CTP	2.23	1.45	1.54
TTP	1.77	0.30	5.90

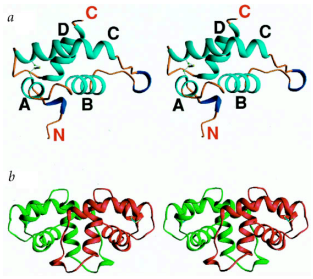
MJ0577, M. jannaschii (Zarembinski TI et al (1998) PNAS 95, 15189)

- Structure contains bound ATP
- Biochemical analysis shows ATPase activity in presence of cell extract, but not on its own



HheA, E. coli (Yang F et al (1998) Nature Struct Biol 5, 763)

- Structural similarity to a domain of Salmonella CheR
- No function could be inferred



Summary

- Protein function depends on its structure
- Structural genomics:
  - A systematic program of 3D structure determination aimed at developing a comprehensive view of protein structure universe
    - Experimentally determine representative protein structures
    - Computationally predict remaining protein structures
  - Goal
    - Infer functional information
    - Other benefits
  - Limitations
    - Technical limitations
    - Biochemical function can be inferred from structure in favorable cases, but biological role is more difficult to infer
    - Cooperation with other experimental methods required
  - Worldwide activity
- Bioinformatics
  - Integrative database required: link structural and functional information