

## Structural genomics

Bostjan Kobe

Professor of Structural Biology

SMMS and IMB

Room 76-452, 3365-2132, b.kobe@uq.edu.au

### Content lectures 1 and 2:

- 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](http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html)
  - Class:  $\alpha$ ,  $\beta$ ,  $\alpha/\beta$ ,  $\alpha+\beta$ , 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

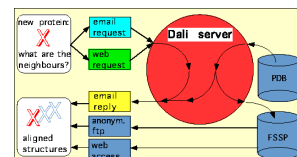
Murzin et al (1995). *J. Mol. Biol.* 247, 536-540.

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

## 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.

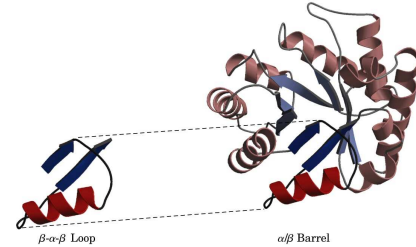
OUTPUT FROM DALI

STRID2 Z RMSD LALI LSEQ2 %IDE PROTEIN

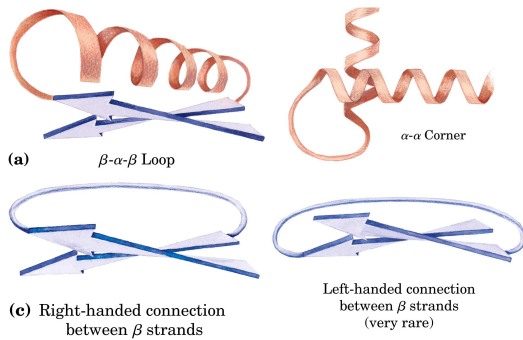
1bk5A	61.5	0.0	422	422	100	karyopherin alpha fragment (importin alpha, srp1p)
1bk5B	58.6	0.4	422	422	100	karyopherin alpha fragment (importin alpha, srp1p)
1bk6A	54.5	0.8	422	422	99	karyopherin alpha fragment (importin alpha, srp1p) biol
1bk6B	54.5	0.8	422	422	99	karyopherin alpha fragment (importin 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
lee4A	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

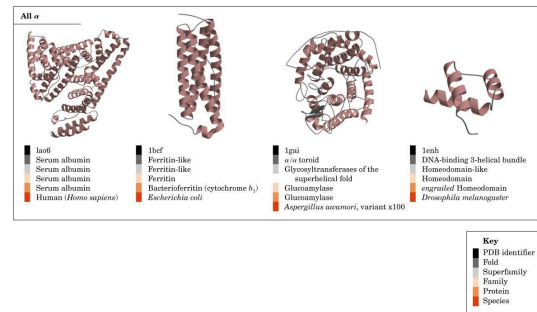
- A specific combination of smaller supersecondary structure motifs



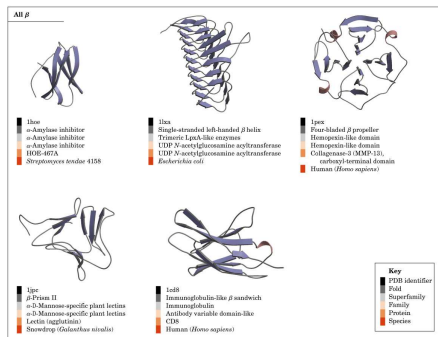
## Supersecondary structure motifs



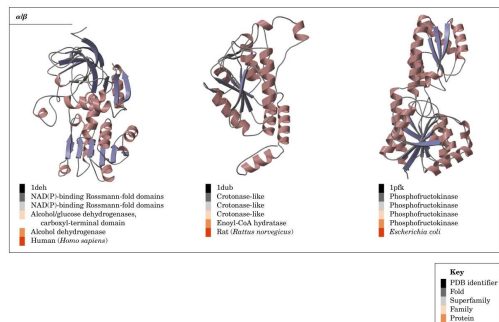
## Examples of protein structure (1)



## Examples of protein structure (2)



## Examples of protein structure (3)





## 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

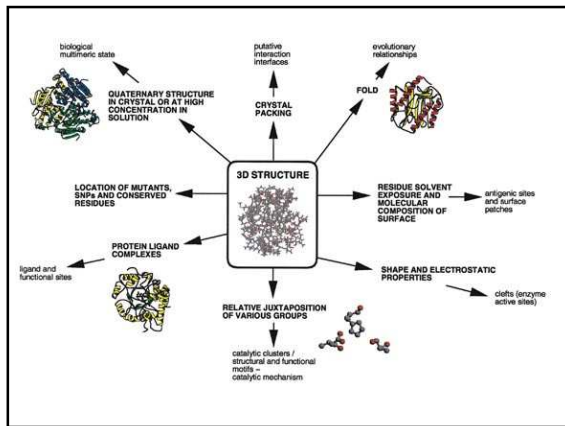
USA

QuickTime™ and a  
TIFF (LZW) decompressor  
are needed to see this picture.

Table 1 Current players in commercial structural genomics			
Company name	Year founded	Location	Technology
<b>Experimental companies</b>			
Astex	1998	Cambridge, UK	High throughput X-ray crystallography/focus on co-complexes
Integrative Proteomics	2000	Toronto, Canada	Automation for protein expression
Structure-Function Genomics	1999	Piscataway, NJ	NMR, protein domain analysis and expression
Structural Genomix	1999	San Diego, CA	High throughput X-ray crystallography and compound design
Syrrx	1999	La Jolla, CA	High throughput X-ray crystallography
<b>Modeling companies</b>			
IBM (Blue Gene project)	2000		Computational protein folding
Inpharmatica	1998	London, UK	Biospectrum database
Geneformatics	1999	San Diego, CA	'Fuzzy functional form' modeling for identifying active sites
Prospect Genomics	1999	San Francisco, CA	Homology modeling
Protein Pathways	1999	Los Angeles, CA	Phylogenetic profiling, domain analysis, expression profiling
Structural Bioinformatics	1996	San Diego, CA and Copenhagen, Denmark	Homology modeling, docking

## 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

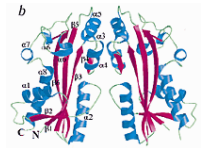
### Statistics from structural genomics

- 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)

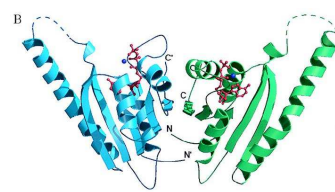
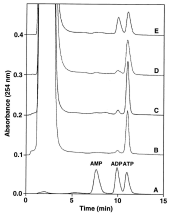
- Partial structural similarity to nucleotide-binding proteins
- Biochemical analysis shows it is nucleotide triphosphatase



	$K_{cat}$ ( $s^{-1}$ )	$K_m$ (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

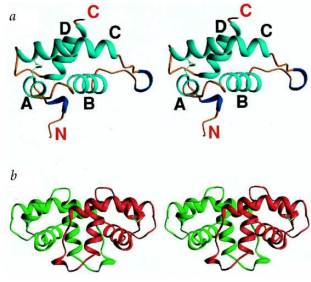
MJ0577, *M. jannaschii* (Zarembinski TI et al (1998) *PNAS* 95, 15189)

- 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