COMPUTATIONAL STRUCTURAL BIOLOGY
STRUCTURE, SIMULATION, FUNCTION & PREDICTION

Lecture 7

Michael Levitt
Structural Biology, Stanford

http://csb.stanford.edu/class
BIOINFORMATICS II

Structure Comparison.
Structure and Sequence.
Structural Genomics.
Expression Patterns.
Discovering Drugs.
Diagnosing Disease.
Structure Comparison
Concept 7.1
STRUCTURE COMPARISON

Structure Superposition.
Gapped Superposition.
Structural.
STRUCTURE SUPERPOSITION

(1) Get an equivalence of points in structure A with structure B. For example, if these are two forms of the same proteins, then atom i of A is equivalenced to atom i of B.

(2) Superimpose the coordinates of of B on those of A using:

\[ r'_b = T \, r_b + t, \]

where \( T \) is a rotation matrix and \( t \) is a translation vector.

(3) Determine \( T \) and \( t \) to minimize \( \sum (r_A - r'_B)^2 \).


Dozens of papers have been written on this since 1926.
SIMPLE STRUCTURAL ALIGNMENT

Molecule A (fixed)

Molecule B

Superimpose B on A #1

Superimpose B on A #2

Superimpose B on A #3
DISTANCE AND SIMILARITY

Convert distance between CA atoms $i$ and $j$ to similarity of $i$ and $j$.

$$S_{ij} = 20/(1+d_{ij}^2/5)$$

If $d_{ij} = 0$, $S_{ij} = 20$

If $d_{ij}^2 = 5$, $S_{ij} = 10$

Many possible $S_{ij} = f(d_{ij})$. 

[Graph showing the relationship between $d_{ij}$ and $S_{ij}$]
STRUCTURAL IN ACTION

Subbiah et al., Current Biol. 3, 141 (1993).

STRUCTURAL ALIGNMENT

Easy. Two globins.
RMS = 1.6Å for 135

Harder.
RMS = 4.3Å for 67.

Very Hard.
Same topology for sub-domain (A. Murzin).
PROTEIN STRUCTURE SIMILARITY MEASURES

Structural alignment recognizes much more of CATH than sequence alignment does.
Structure and Sequence Concept 7.2
COMPARING SEQUENCE & STRUCTURE COMPARISON

Statistical Significance.
E-values.
Similarity measures.
SEQUENCE SCORE DISTRIBUTION

Contour number of occurrences with particular \((\text{Score},N)\) value.

- Include only those pairs of sequences for proteins in different SCOP class.
- None of these pairs should be significantly similar.
- As the protein length increases, the sequence score for a non-significant match increases as \(\log n\). (\(n\) is sequence length)
SEQUENCE SCORE DISTRIBUTION

Contour number of occurrences with particular (Score, N) value.

These “islands” are the true positives.

- Include all pairs of sequences.

SEQUENCE SCORE DISTRIBUTION

Contour number of occurrences with particular (Score, N) value.

- Show additional pairs in same class.
- Just pairs between SCOP classes.
SEQUENCE SCORE IS EXTREME VALUE DISTRIBUTION

- For the extreme value distribution,
  \[ P(Z) = \exp(-Z - \exp(-Z)) \]
  \[ \log_\text{10}(P(Z)) = -Z - \exp(-Z) \]

- For \( Z > 0 \):
  \[ \log_\text{10}(P(Z)) = -Z \]

- For \( Z < 0 \):
  \[ \log_\text{10}(P(Z)) = -\exp(-Z) \]
  where \( Z = (\text{score} - \text{mean})/\text{SD} \).
STRUCTURE SCORE DISTRIBUTION

Contour number of occurrences with particular (Score, N) value.

- Include only those pairs of sequences for proteins in different SCOP class.
- None of these pairs should be significantly similar.
- As the protein length increases, the Score for a non-significant match increases with N, the match length.
STRUCTURE SCORE DISTRIBUTION

Contour number of occurrences with particular \((\text{Score}, N)\) value.

- These "islands" are the true positives.
- Include all pairs of structures for proteins in SCOP.
- The distribution is not as clean as for sequences.
STRUCTURE SCORE DISTRIBUTION

Contour number of occurrences with particular \( (\text{Score}, N) \) value.

- Show additional pairs in same class.
- Just pairs between SCOP classes.
STRUCTURE SCORE follows extreme value

\[ \log_{10}(p) \]

Score

\[ \exp(-Z) \]

\[ N = 55 \]

\[ N = 75 \]

\[ N = 95 \]

\[ N = 115 \]

\[ N \text{ is length of matched region.} \]
STATISTICAL SIGNIFICANCE

- Few pairs have no significant structure match yet have a very significant sequence match.
- Many pairs have no significant sequence match yet have a very significant structure match.
SIMILARITY MEASURES

- The similarity measures that work best for both sequence and structure comparison are a sum of scores for each match.

- These are not the normally used measures.
  For sequence we like to use percent identity (%ID).
  For structure we like to use root mean square deviation (RMS).

- Neither %ID or RMS obey an Extreme value distribution for random matches.

- Neither %ID or RMS are reliable indicators of a significant match.
Structural Genomics
Concept 7.3
STRUCTURAL GENOMICS PROJECT

- Aim is to solve structures of all protein sequences.

- This is too much work.

- Solve enough structures so as to be able to model the rest.

- The number that needs to be solved depends on our modeling ability.
WHAT CAN ONE DO WITH STRUCTURAL GENOMICS

START

Template Search

Target – Template Alignment

Model Building

Model Evaluation

NO

OK?

YES

TARGET

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDO
VPERLIPERASFQMNDK

TEMPLATE

Comparative or Homology modeling


©Michael Levitt 04
STRUCTURAL GENOMICS & MODELING

- Solve enough structures so as to be able to model the rest.
- The number that needs to be solved depends on our ability to model (the radius of dashed circle).
EXPECTATIONS FROM STRUCTURAL GENOMICS

- This is for sequence that are clearly homologous.
- The fraction of folds that are new was 85% in 1987 and 30% in 1997. Today it is less.

Brenner & Levitt, Protein Engineering, 9, 197 (2000).
JCSG CENTER

- 8 new folds out of 57 deposited PDB files.
- Look at 6303 targets.
- UCSD, Scripps, Stanford.
BSGC CENTER

388 targets.
38 in PDB.
New Folds?

- UCB, UNC, LBL
  Stanford
MCSG CENTER

Consortium Members:

- Argonne National Laboratory
- Northwestern University
- Washington University School of Medicine
- University College London
- UT Southwestern Medical Center at Dallas
- University of Toronto
- University of Virginia

Active Targets: 1325
Crystallized: 318
In PDB: 112
New Folds: 15

The webpage is under construction in W.Minor Lab

1325 targets.
112 in PDB.
15 new folds.
NYSG CENTER

Mission Statement
To develop and use the technology for high-throughput structural and functional studies of proteins.

Participating Research Groups
- Albert Einstein College of Medicine
- Brookhaven National Laboratory
- Columbia University
- Structural Genomics, Inc
- The Rockefeller University
- University of California, San Francisco
- Weill Medical College of Cornell University

Public Target Information
- Public Target Progress Report
- Download Public Target Progress Report in XML Format

Progress Statistics
- Selected: 1713;
- Cloned: 750;
- Expressed: 688;
- Soluble: 626;
- Purified: 472;
- Crystallized: 178;
- Diffraction-quality Crystals: 106;
- Native diffraction-data: 83;
- Phasing diffraction-data: 83;
- Crystal Structure: 83;
- Deposited in PDB: 73;

- Columbia, BNL, Rockefeller, UCSF, SGX Inc.

1713 targets.
73 in PDB.
? new folds.
4077 targets.
43 in PDB.
? new folds.
<table>
<thead>
<tr>
<th>Region</th>
<th>Value 1</th>
<th>Value 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Genome</td>
<td>57</td>
<td>8</td>
</tr>
<tr>
<td>Berkeley</td>
<td>38</td>
<td>?</td>
</tr>
<tr>
<td>Midwest</td>
<td>112</td>
<td>15</td>
</tr>
<tr>
<td>New York</td>
<td>73</td>
<td>?</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>43</td>
<td>?</td>
</tr>
</tbody>
</table>

Estimate 15% new so about 50 new folds.
Expression Patterns
Concept 7.4
**MICROARRAY BASICS**

- Use position in array to distinguish immobilized molecules (A, B, C, D, E, F, G, H, I, J...).

- Use dye color (red or green) to distinguish source of soluble molecules (a, b, c, d, e, f, g, h, i, j...).

- Assume A only binds a, B only binds b, etc.

- Use fluorescence pattern to analyze.
OLIGO DNA MICROARRAYS

- In DNA arrays, A, B, C, D, etc are short single-stranded DNA oligomers of length ~25.
- Make DNA arrays by photo-lithography.
- Density can be very high (11 um features). 1,300,000 oligos per chip.
- Need several about 20 oligos per gene.

Mask. Need 4 mask for each added nucleotide (A, G, C, T)

Silicon Chip

CDNA EXPRESSION ARRAYS

- In cDNA arrays, A, B, C, D, etc are single-stranded cDNA molecules of length ~500 or more.
- Make cDNA arrays by spotting or printing.
- Density lower than DNA chip. More uniform
  - 25,000 genes per chip.
  - Spot cDNA with robot or ink-jet printer. Make DNA from sequence.

MORE EXPRESSION ARRAY CARTOONS

Oligo DNA Chip hybridizes RNA

cDNA chip also hybridizes RNA

A Comparative Hybridization Experiment
**Experimental setup**

1. Obtain two samples of liver cells; apply the drug to one sample. Then, from each sample, collect molecules of messenger RNA (mRNA)—the mobile copies of genes and the templates for protein synthesis in cells.

2. Transcribe the mRNA into more stable complementary DNA (cDNA) and add fluorescent labels—green to cDNAs derived from untreated cells, red to those from treated cells.

3. Apply the labeled cDNAs to the chip. Binding occurs when cDNA from a sample finds its complementary sequence of bases on the chip (detail at right). Such binding means that the gene represented by the chip DNA was active, or expressed, in the sample.

**Examples of reactions**
Computational analysis.

5 Put the chip in a scanner. Have a computer calculate the ratio of red to green at each spot (to quantify any changes in gene activity induced by the drug) and generate a color-coded readout.

6 Determine whether any genes responded strongly to the drug in ways known to promote or reflect liver damage. Or compare the overall expression pattern produced by strong responders with the patterns produced when those genes react to known liver toxins (right). Close similarity would indicate that the new candidate was probably toxic as well. In the diagram, each box represents a single gene’s response to a compound.

CLEANING UP Oligo Chips

Filter bad Spots:
Small spots.
Smudgy spots.
Non-round spots.

Background correction:
Raw data has large positional dependence.
Remove artifactual smudges.

NOISY BACKGROUND

Before

Raw data has large positional dependence

After

Remove artifactual smudges

Positional normalization removes most of the artifact.
Discovering Drugs
Concept 7.5
DRUG BINDING

Drugs bind to proteins by lock and key fit.
Structure based Drug Design

Hydrophobic residue.

Hydrophobic Pocket

Design drug to fit pocket.
DRUG DOCKING

DOCK developed in 1981 by Kuntz and co-workers at UCSF.

Fit into active site by geometrical criteria


http://dock.compbio.ucsf.edu/
AUTODOCK

- Use simulated annealing to fit into active site with simple energy function.

- Olson group at Scripps in 1989.


http://www.scripps.edu/pub/olson--web/people/gmm/
AUTODOCK MOVIES

Benzamidine binding to Trypsin.
Gives energies in increasing order.

http://www.scripps.edu/pub/olson-web/people/gmm/
AUTODOCK MOVIES

Biotin binding to Streptavidin.
Gives energies in increasing order.

http://www.scripps.edu/pub/olson-web/people/gmm/
DRUGS FROM SEQUENCE

(1) Chromosome
(2) DNA Sequence
(3) Protein Sequence
(4) Multi Sequence Alignment
(5) Model Structure
(6) Dock Drug

From Gene to Drug in six “easy” steps.
Diagnosing Disease Concept 7.6
PHARMACOGENOMICS

Little data so far.

Lots of genes.

http://pharmgkb.org/
PROTEIN 2-D GELS

Map Selection: LYMPHOMA_HUMAN

Spots corresponding to known proteins are highlighted in red. Please click on one of them or select the same map in another format:

- LYMPHOMA_HUMAN large, spots highlighted
- LYMPHOMA_HUMAN small, spots not highlighted
- LYMPHOMA_HUMAN large, spots not highlighted

For high resolution gel image, see reference or download the tiff file.

- Separate proteins in tissue by pK (charge) and Molecular weight (size).
- Need to have a enough protein.
- Look for differences relative to normal tissue.
**Expression Array Diagnosis**

- Cluster genes by similarity of their expression pattern in the different tumors.
- Find that this clusters tumors.
- Proliferating cancers are separated from others by the genes that they express.


EXPRESSION ARRAY DIAGNOSIS

- Cluster genes by similarity of expression pattern.
- Find that this clusters tumors.


Red: over-expressed gene in tumor.
Green: under-expressed in tumor.