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

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

Threading

Sequence-structure compatibility (fold recognition)

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.

Model structure coverage in sequence space

Adopted from Vitkup et al., 2001

Adopted from Redfern and Orengo, 2005
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
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

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

Yee et al., Acc. Chem. Res. 2003, 36, 183-189
NIGMS Protein Structure Initiative

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected</td>
<td>11214</td>
<td>21872</td>
<td>42726</td>
<td>74637</td>
</tr>
<tr>
<td>Cloned</td>
<td>5465</td>
<td>11277</td>
<td>23237</td>
<td>45353</td>
</tr>
<tr>
<td>Expressed</td>
<td>2860</td>
<td>6115</td>
<td>13602</td>
<td>25536</td>
</tr>
<tr>
<td>Purified</td>
<td>1505</td>
<td>2823</td>
<td>5291</td>
<td>8398</td>
</tr>
<tr>
<td>Crystallized</td>
<td>336</td>
<td>1161</td>
<td>1876</td>
<td>3199</td>
</tr>
<tr>
<td>Diffraction</td>
<td>96</td>
<td>438</td>
<td>767</td>
<td>1651</td>
</tr>
<tr>
<td>Crystal structure</td>
<td>87</td>
<td>314</td>
<td>545</td>
<td>1260</td>
</tr>
<tr>
<td>PDB</td>
<td>76</td>
<td>247</td>
<td>569</td>
<td>1488</td>
</tr>
</tbody>
</table>

Adopted from O'Toole et al., 2004

Targets by genome

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of targets</th>
<th>% Of all targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caenorhabditis elegans</td>
<td>4674</td>
<td>17.4</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>3900</td>
<td>14.5</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>3257</td>
<td>12.1</td>
</tr>
<tr>
<td>Pyrococccus furiosus</td>
<td>2179</td>
<td>8.1</td>
</tr>
<tr>
<td>Thermotoga maritima</td>
<td>1890</td>
<td>6.9</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>1476</td>
<td>5.5</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>1272</td>
<td>4.7</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>1254</td>
<td>4.7</td>
</tr>
<tr>
<td>Bacillus subtilis</td>
<td>1220</td>
<td>4.5</td>
</tr>
<tr>
<td>Bacillus steaotermophilus</td>
<td>784</td>
<td>2.8</td>
</tr>
</tbody>
</table>

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

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, (XUE, and tab delimited formats)
- Download target data document type definition.
- The XUE document is organized according to the recommendations of the International Task Force on Target Tracking.
- Target Status Query Feature.
- Summary Reports

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

<table>
<thead>
<tr>
<th>BISS</th>
<th>BSSC</th>
<th>ESSG</th>
<th>ESIO</th>
<th>ESIP</th>
<th>JCSG</th>
<th>MESS</th>
<th>MGSP</th>
<th>NGSG</th>
<th>NYSXRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPFG</td>
<td>RIKEN</td>
<td>SJF</td>
<td>SCCG</td>
<td>SIC</td>
<td>SICG</td>
<td>SICG</td>
<td>SICG</td>
<td>SICG</td>
<td>SICG</td>
</tr>
</tbody>
</table>

Using the Target Search Form: