Protein Structure Analysis

Iosif Vaisman

2004

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.

Structural genomics data integration

Target selection

- Realm of interest
- Family exclusion - impossible
- Family exclusion - known
- Prioritization
- Selection
- Analysis and interpretation

M. thermoautotrophicum
structural genomics project

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

Adopted from Vitkup et al., 2001
**M. thermoautotrophicum** structural genomics project: selected results

### Table: 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>4974</td>
<td>17.1</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>Pseudomonas aeruginosa</td>
<td>2179</td>
<td>8.1</td>
</tr>
<tr>
<td>Thermotoga maritima</td>
<td>1980</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>Staphylococcus aureus</td>
<td>1230</td>
<td>4.5</td>
</tr>
<tr>
<td>Bacteroides vulgatus</td>
<td>754</td>
<td>2.8</td>
</tr>
</tbody>
</table>

*Adopted from O'Toole et al., 2004*

### Diagram: Structural genomics shortcuts

- **GENOMIC SEQUENCING PROJECT**
  - Gene annotation (e.g., cloning)
  - Protein expression (e.g., purification)
  - Biochemical characterization (e.g., nucleic acid extraction)

### Diagram: Target size distribution and program progress

#### n targets

- **a** Target sizes in (b) at 24/4/2003
  - n targets: 74637
  - Selected: 11214
  - Cloned: 5465
  - Expressed: 2860
  - Purified: 1505
  - Crystallized: 336
  - Diffraction: 96
  - Crystal structure: 87
  - PDB: 76

**NIGMS Protein Structure Initiative**

<table>
<thead>
<tr>
<th>Date</th>
<th>Selected</th>
<th>Cloned</th>
<th>Expressed</th>
<th>Purified</th>
<th>Crystallized</th>
<th>Diffraction</th>
<th>Crystal structure</th>
<th>PDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/4/2001</td>
<td>11214</td>
<td>5465</td>
<td>2860</td>
<td>1505</td>
<td>336</td>
<td>96</td>
<td>87</td>
<td>76</td>
</tr>
<tr>
<td>12/4/2002</td>
<td>21872</td>
<td>11277</td>
<td>13602</td>
<td>2823</td>
<td>1161</td>
<td>438</td>
<td>314</td>
<td>247</td>
</tr>
<tr>
<td>12/4/2003</td>
<td>42726</td>
<td>23327</td>
<td>25586</td>
<td>5291</td>
<td>1876</td>
<td>767</td>
<td>545</td>
<td>569</td>
</tr>
<tr>
<td>12/1/2004</td>
<td>74637</td>
<td>45553</td>
<td>25536</td>
<td>8398</td>
<td>3199</td>
<td>1651</td>
<td>1260</td>
<td>1488</td>
</tr>
</tbody>
</table>