THE IMPACT OF STRUCTURAL GENOMICS:
Expectations and Outcomes

John-Marc Chandonia and Steven E. Brenner
STRUCTURAL GENOMICS

- Similar to structural biology in the methods used
  - X-ray crystallography, NMR
- Attempts to determine the structure of all the proteins of a given organism
- Focuses on High Throughput Screening (HTS)
  - Robotics, data processing software, sensitive detectors
- Economy of scale
PROTEIN STRUCTURE INITIATIVE (PSI)

- Focuses on decreasing the cost and time associated with 3-D protein structure determination using structural genomics
- 10 year, 764 million budget
- Two phases
  - Phase 1 (2000-2005)
    - To develop methods that streamline the determination of protein structures
  - Phase 2 (2005-2010)
    - To use methods developed in Phase 1 to determine a large number of protein structures, and to continue to streamline the processes in structural genomics
GOALS

- **Measures of success**
  - Biological importance and difficulty
  - Novel structures
    - First protein in a family can have its structure and function evaluated, then its properties can be translated to:
      - Create comparative models
      - Find new evolutionary relationships between proteins
FINDING NEW STRUCTURES BY SEQUENCE COMPARISON

- BLAST and PSI-BLAST were used to determine sequence similarity, to help remove bias introduced by Pfam
  - Pfam does not include many species specific proteins
  - The number of novel structures has decreased over the last 15 years
    - 20% in 1990, 10% in 2005
  - According to the analysis, SG structures accounted for 44% of the total number of new structures in 2004
c) Overlap between PSI-BLAST and Pfam

First structural representative of a Pfam family

Previously characterized Pfam family

Not in Pfam

Year
A Novelty of Structural Genomics Targets, by direct sequence comparison with earlier structures

Fraction Novel, by Center

Highest Novelty Level:
- >95% ID
- 95% ID
- 30% ID
- BLAST (E=10^-4)
- BLAST (E=10^-2)
- PSI-BLAST (E=10^-2)

Center

More Novel

More Sequence-Similar
IMPACT OF STRUCTURAL GENOMICS ON COVERAGE OF PROTEIN FAMILIES

- Pfam families (since 2000)
  - 20.4% of structures reported by PSI represented new families
  - 5% of structures reported by non-SG labs represented new families
  - Structure determination of first structures in a Pfam family by non-SG has decreased, while SG centers have made up the difference

- SG centers now account for half of new structurally characterized families, but only make up 20% of new structures
B Pfam families with a first representative solved, per month

Monthly Totals

1-year Moving Average

All

non-SG

SG

PSI

New Pfam Families Solved

Year

90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05
# Novel Structures Solved

<table>
<thead>
<tr>
<th>Group or SG center</th>
<th>Targets and nonidentical chains</th>
<th>New Pfam families (total family size)</th>
<th>Novel structures (30% ID)</th>
<th>New SCOP folds</th>
<th>New SCOP fold or superfamily</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SG centers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berkeley Structural Genomics Center (BSGC)</td>
<td>57 (57 chains)</td>
<td>22 (5757)</td>
<td>41</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Center for Eukaryotic Structural Genomics (CESG)</td>
<td>48 (48 chains)</td>
<td>7 (387)</td>
<td>28</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Joint Center for Structural Genomics (JCSG)</td>
<td>186 (187 chains)</td>
<td>32 (4875)</td>
<td>92</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Midwest Center for Structural Genomics (MCSG)</td>
<td>224 (229 chains)</td>
<td>55 (5512)</td>
<td>163</td>
<td>18</td>
<td>25</td>
</tr>
<tr>
<td>Northeast Structural Genomics Consortium (NESGC)</td>
<td>159 (159 chains)</td>
<td>52 (4811)</td>
<td>108</td>
<td>15</td>
<td>26</td>
</tr>
<tr>
<td>New York Structural Genomics Research Consortium (NYSGRC)</td>
<td>166 (171 chains)</td>
<td>27 (3982)</td>
<td>90</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Southeast Collaboratory for Structural Genomics (SECSG)</td>
<td>67 (67 chains)</td>
<td>6 (1079)</td>
<td>25</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Structural Genomics of Pathogenic Protozoa Consortium (SGPP)</td>
<td>26 (26 chains)</td>
<td>1 (19)</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>TB Structural Genomics Consortium (TB)</td>
<td>99 (99 chains)</td>
<td>9 (3938)</td>
<td>42</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>PSI centers (total of 9 centers above)</strong></td>
<td>1032 (1043 chains)</td>
<td>211 (30,360)</td>
<td>597</td>
<td>48</td>
<td>74</td>
</tr>
<tr>
<td>Japanese center (RIKEN)</td>
<td>686 (718 chains)</td>
<td>50 (6860)</td>
<td>289</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Other international SG (total, excluding all centers above)</td>
<td>169 (183 chains)</td>
<td>33 (5877)</td>
<td>69</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td><strong>Non-SG groups (since 2000)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-SG structural biology (total)</td>
<td>17,096 (23,747 chains)</td>
<td>928 (249,171)</td>
<td>2,521</td>
<td>269</td>
<td>478</td>
</tr>
<tr>
<td>Steitz group</td>
<td>46 (559 chains)</td>
<td>23 (4190)</td>
<td>31</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>Huber group</td>
<td>185 (273 chains)</td>
<td>8 (679)</td>
<td>38</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Iwata group</td>
<td>14 (54 chains)</td>
<td>14 (7960)</td>
<td>20</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
IMPACT OF STRUCTURAL GENOMICS ON THE STRUCTURAL CLASSIFICATION OF PROTEINS

- SCOP classifications
  - Family
    - Clear common evolutionary origin, one family member can be used to construct comparative models
  - Superfamily
    - Groups of families that have similar structure or functions that imply a common evolutionary origin
  - Fold
    - Superfamilies that share similar secondary structures, but have little evidence of common evolutionary origin
Fig. S4: Time Course of Results for PSI Centers
70% of Non-SG structures in the last 10 years represent a new experiment on a protein with a previously determined structure.

The percentage of domains that represent a new family in SCOP has decreased from 9.6% in 1995 to 4.4% in 2004.

- Structural biologist tend to work on known proteins more often.
RESULTS VS. EXPECTATIONS

For PSI, the number of domains that represented a new SCOP domain or superfamily was 16%

- Higher than non-SG average of 4%, but not at the target of 40%
- Can use BLAST or PSI-BLAST to avoid finding structures for homologs of known proteins
- Higher sequence novelty correlates with higher rates of discovery for new folds, superfamilies and families
B Novelty of Structural Genomics Targets in SCOP

Fraction New (SCOP category), by Center

- Experiment
- Species
- Protein
- Family
- Superfamily
- Fold

Center

BSGC  CESG  JCSG  MCSG  NESGC  NYS3RC  SECSG  SGPP  TB  PSI average  RIKEN  Other Inst  SG  StrBlO (no sequence similarity to earlier structures)
COST EFFECTIVENESS

- Average costs for determining a structure with <95% sequence similarity
  - Non-SG: $250,000 to $300,000
  - PSI: $211,000 (overall)
    - 2004-2005: $138,000
      - Most productive center (MCSG): $67,000
      - Adjusted for protein size and composition
        - Larger complexes are more difficult
        - 66% to 85% of non-SG
COST EFFECTIVENESS

- Costs per novel structure
  - 30% sequence ID
    - Non-SG: $532,000 to $1.9 million
    - PSI: $364,000
  - New Pfam family
    - Non-SG: $1.5 to $5.5 million
    - PSI: $1.0 million
  - SCOP superfamily or fold
    - Non-SG: $2.0 to $7.3 million
    - PSI: $2.2 million

- Since most structural biology labs focus on proteins with similar sequences to those already solved, the cost associated is much higher for new structures
  - Due to focus on function
COMPARISONS WITH LEADING STRUCTURAL BIOLOGISTS

- Large Structural Biology labs
  - Good at solving large, challenging complexes
  - Robert Huber: proteasome, DNA primase, light harvesting complexes
  - So Iwata: photosystem II complex
  - Tom Steitz: protein-nucleic acid complexes
  - Budget
    - $1.5 vs. $5.7 million for PSI centers
    - Also comparable in cost efficiency at solving novel structures

<table>
<thead>
<tr>
<th>Non-5G groups (since 2000)</th>
<th>Non-5G structural biology (total)</th>
<th>17,096 (23,747 chains)</th>
<th>928 (249,171)</th>
<th>2,521</th>
<th>269</th>
<th>478</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steitz group</td>
<td>46 (559 chains)</td>
<td>23 (4190)</td>
<td>31</td>
<td>7</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Huber group</td>
<td>185 (273 chains)</td>
<td>8 (679)</td>
<td>38</td>
<td>5</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Iwata group</td>
<td>14 (54 chains)</td>
<td>14 (7960)</td>
<td>20</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>
CONFOUNDING FACTORS

- Factors not taken into consideration
  - Many SG centers collaborate with structural biology groups
    - Causes some of the cost protein structure determination and materials to be shifted onto structural biologists
  - SG centers included structures in the lists solved before their funding in 2000
  - However, a lot of capital was invested into new technologies that may not have given a return yet
  - SG centers had to put money into additional costs
    - Computation, data reporting and analysis
  - Many structural biology projects benefited from prior work on the proteins they study, which is important for more complicated projects
EVALUATION OF STRUCTURE IMPACT

- Can be crudely determined by number of subsequent citations
  - 104 SG structures published between 2001 and 2002
    - 11 average, 4 median
    - The two most cited (107 and 61) describe the overall work of a SG center
  - Randomly selected 104 non-SG publications
    - 21 average, 11.5 median
    - However, novel structures were cited more often
  - This may have to do with the fact that traditional structural biologists usually biochemically characterize their protein as well
CRITICISM

- Gives little functional information of determined proteins
  - Open vs. closed Bcr-Abl tyrosine kinase
- Very pricey
  - Budget could fund 100-200 traditional structural biology labs
  - Uses public money
- Author bias
  - Chandonia and Brenner are affiliated with BSCG, a PSI lab
IS IT WORTH IT?

- Depends on different factors
  - SG tends to focus on smaller, easier to determine structures
  - Relatively little functional information may be known about a given protein
  - However, it allows for a greater rate of discovery of novel protein families
- Good for overall knowledge, not necessarily for specific problems