Structural Biology on the Grid
SBGrid Research Coordination Network
Harvard Medical School
Summary

SBGrid Background
Portal Infrastructure
Molecular Replacement
Objectives and Priorities
SBGrid

- Consortium of structural biology labs
  - Involved in structure determination of (primarily) proteins
  - X-Ray crystallography, NRM, EM
- 87 member labs across the US
  - 28 at Harvard
- 90 software packages
- Modest local cluster resource
  - 300 cores across several clusters (Intel, Mac, AMD)
- Now developing web-based portal interfaces to key apps

Ian Stokes-Rees, http://sbgrid.org
Welcome to SBGrid

Structural Biology Grid (SBGrid) is a computing collaboration of several X-ray crystallography, NMR and electron microscopy laboratories. Participating laboratories include groups primarily at Harvard Medical School, Harvard University and Yale Medical School, but our alumni often remain members after becoming principal investigators at other institutions.

Explore SBGrid.org

- See what SBGrid can do for you
- See the labs currently affiliated with SBGrid
- Browse the software that SBGrid configures for its affiliates
- Learn more about the benefits of Grid Computing
- See how SBGrid is leveraging the power of OS X for scientific applications
- Download publically-available software released by SBGrid

Latest News

- SBGrid User Summit: Quo Vadis Structural Biology?
  SAVE THE DATE! The first ever SBGrid user summit ‘Quo Vadis Structural Biology?’ will take place on May 5th and 6th 2008 in Boston, Massachusetts. We will have a number of talks and workshops focused (read more...)
- Grid computing specialist joins SBGrid
- Experimental cluster acquired under NSF grant
- Apple Worldwide Developers Conference
- New Consortium Members
- SBGrid Receives Taplin Funds for Discovery Award for Grid Computing

Ian Stokes-Rees, http://sbgrid.org
SBGrid Services

- Application optimization
- Application packaging and automated distribution/update
- Help desk
- Centralized license management

Soon:
- Grid portal for SB applications
- Gateway to OSG
Motivation for Grid

Because **computational requirements** continue to be a bottleneck

Because **complexity of tools** impedes quality and efficiency of scientific investigation

Because some affiliated labs don’t have large compute clusters available to them

Because **new computationally intensive methods** are being developed
Portal Infrastructure
Software Stack

Client
Browser
Globus
X.509

Grid Portal
Tomcat
Grid Sphere
OGCE
Globus
Condor
MyProxy

Cluster
Globus
Condor

DB
File

Cluster
SGE
PBS
Torque

Cluster
Globus

Cluster
OSG

Ian Stokes-Rees, http://sbgrid.org
Molecular Replacement
Macromolecular Crystallography

CRYSTAL → RECIPROCAL LATTICE → ELECTRON DENSITY MAP → MODEL

Ian Stokes-Rees, http://sbgrid.org
Fourier series for electron density is a sum of contributions from individual reflections.

Ian Stokes-Rees, http://sbgrid.org
Phase Problem

$$F_{HKL}$$

- **Amplitudes**
  - can be measured
  - $\sim$ sq rt of intensity

- **Frequency**
  - Fixed and known from X-ray source

- **Phase**
  - Unknown!

Ian Stokes-Rees, http://sbgrid.org
# Heavy Atom Derivatives

Many heavy atom derivatives needed: time consuming


Ian Stokes-Rees, [http://sbgrid.org](http://sbgrid.org)

---

**Table S1 | Data collection and refinement statistics**

<table>
<thead>
<tr>
<th></th>
<th>Native</th>
<th>Xenon</th>
<th>Ta$<em>2$Br$</em>{14}$</th>
<th>SelenoMet</th>
<th>2'-1-ADP</th>
<th>EMP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Collection</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>C2</td>
<td>C2</td>
<td>C2</td>
<td>C2</td>
<td>C2</td>
<td>C2</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>161.279</td>
<td>161.391</td>
<td>160.895</td>
<td>161.399</td>
<td>160.621</td>
<td>161.510</td>
</tr>
<tr>
<td></td>
<td>103.955</td>
<td>103.839</td>
<td>104.163</td>
<td>103.886</td>
<td>105.045</td>
<td>104.074</td>
</tr>
<tr>
<td></td>
<td>181.013</td>
<td>181.948</td>
<td>181.937</td>
<td>181.644</td>
<td>181.245</td>
<td>181.546</td>
</tr>
<tr>
<td>α, β, γ (°)</td>
<td>90.000</td>
<td>90.000</td>
<td>90.000</td>
<td>90.000</td>
<td>90.000</td>
<td>90.000</td>
</tr>
<tr>
<td></td>
<td>97.987</td>
<td>97.471</td>
<td>97.849</td>
<td>97.501</td>
<td>98.021</td>
<td>97.672</td>
</tr>
<tr>
<td></td>
<td>90.000</td>
<td>90.000</td>
<td>90.000</td>
<td>90.000</td>
<td>90.000</td>
<td>90.000</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>30-3.0</td>
<td>30-3.3</td>
<td>30-3.8</td>
<td>30-3.3</td>
<td>30-3.1</td>
<td>30-3.3</td>
</tr>
<tr>
<td>$R_{	ext{sym}}$ or $R_{	ext{merge}}$</td>
<td>8.1 (56.9)</td>
<td>10.9 (45.7)</td>
<td>10.8 (27.8)</td>
<td>8.1 (35.1)</td>
<td>7.9 (45.9)</td>
<td>9.9 (53.3)</td>
</tr>
<tr>
<td>l/$σ$</td>
<td>20.9 (2.18)</td>
<td>14.5 (2.00)</td>
<td>13.3 (3.96)</td>
<td>16.7 (2.83)</td>
<td>19.5 (1.99)</td>
<td>15.0 (2.40)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.1 (87.9)</td>
<td>98.0 (85.9)</td>
<td>95.2 (73.9)</td>
<td>97.6 (79.2)</td>
<td>98.0 (79.9)</td>
<td>100 (100)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>7.4 (5.3)</td>
<td>7.1 (5.1)</td>
<td>6.6 (5.5)</td>
<td>6.0 (4.4)</td>
<td>6.2 (5.1)</td>
<td>6.4 (5.9)</td>
</tr>
</tbody>
</table>
Molecular Replacement

Homologous or incomplete model:

Combining model phases with experimental intensities will reveal details of missing elements

Typically 30% identity and 1/3 of a structure required.
Cross-Rotation Function

NUMBER OF RF PEAKS : 10

<table>
<thead>
<tr>
<th>SOL_RF</th>
<th>THETA</th>
<th>PHI</th>
<th>CHI</th>
<th>ALPHA</th>
<th>BETA</th>
<th>GAMMA</th>
<th>RF</th>
<th>RF/SIGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>146.34</td>
<td>-139.17</td>
<td>155.19</td>
<td>55.63</td>
<td>65.55</td>
<td>153.97</td>
<td>0.1453E+07</td>
<td>3.69</td>
</tr>
<tr>
<td>2</td>
<td>132.29</td>
<td>165.77</td>
<td>54.08</td>
<td>56.82</td>
<td>39.30</td>
<td>265.27</td>
<td>0.1409E+07</td>
<td>3.57</td>
</tr>
<tr>
<td>3</td>
<td>126.82</td>
<td>159.77</td>
<td>57.42</td>
<td>51.60</td>
<td>45.23</td>
<td>272.06</td>
<td>0.1395E+07</td>
<td>3.54</td>
</tr>
<tr>
<td>4</td>
<td>147.99</td>
<td>-141.48</td>
<td>160.47</td>
<td>49.99</td>
<td>62.99</td>
<td>152.96</td>
<td>0.1387E+07</td>
<td>3.52</td>
</tr>
<tr>
<td>5</td>
<td>154.58</td>
<td>-138.22</td>
<td>162.21</td>
<td>50.18</td>
<td>148.05</td>
<td>0.1285E+07</td>
<td>3.26</td>
<td></td>
</tr>
</tbody>
</table>

Chi = 180.0

NUMBER OF PEAKS : 50

<table>
<thead>
<tr>
<th>ALPHA</th>
<th>BETA</th>
<th>GAMMA</th>
<th>THETA</th>
<th>PHI</th>
<th>CHI</th>
<th>RF</th>
<th>RF/SIGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>49.88</td>
<td>82.51</td>
<td>104.56</td>
<td>41.97</td>
<td>62.66</td>
<td>160.85</td>
<td>7748.</td>
</tr>
<tr>
<td>2</td>
<td>102.03</td>
<td>27.25</td>
<td>29.60</td>
<td>14.88</td>
<td>126.21</td>
<td>133.08</td>
<td>4409.</td>
</tr>
<tr>
<td>3</td>
<td>141.78</td>
<td>36.81</td>
<td>105.22</td>
<td>158.24</td>
<td>-71.72</td>
<td>116.84</td>
<td>4346.</td>
</tr>
<tr>
<td>4</td>
<td>50.12</td>
<td>44.24</td>
<td>104.89</td>
<td>22.60</td>
<td>62.61</td>
<td>156.88</td>
<td>3881.</td>
</tr>
<tr>
<td>5</td>
<td>148.79</td>
<td>85.37</td>
<td>27.42</td>
<td>42.70</td>
<td>150.68</td>
<td>177.22</td>
<td>3812.</td>
</tr>
</tbody>
</table>

Chi = 180.0

model with 30% identity, maybe?

1-24 hours per model

Peak?

YES!

Ian Stokes-Rees, http://sbgrid.org
Translation Function

<table>
<thead>
<tr>
<th>IX</th>
<th>IY</th>
<th>IZ</th>
<th>XFRAC</th>
<th>YFRAC</th>
<th>ZFRAC</th>
<th>XORT</th>
<th>YORT</th>
<th>ZORT</th>
<th>DENS</th>
<th>DENS/SIGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>55</td>
<td>79</td>
<td>0.333</td>
<td>0.457</td>
<td>0.410</td>
<td>16.45</td>
<td>34.12</td>
<td>50.41</td>
<td>0.2875</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>45</td>
<td>79</td>
<td>0.335</td>
<td>0.379</td>
<td>0.410</td>
<td>16.54</td>
<td>28.31</td>
<td>50.41</td>
<td>0.1590</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>50</td>
<td>79</td>
<td>0.334</td>
<td>0.417</td>
<td>0.411</td>
<td>16.51</td>
<td>31.18</td>
<td>50.55</td>
<td>0.1518</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>55</td>
<td>62</td>
<td>0.331</td>
<td>0.458</td>
<td>0.322</td>
<td>16.35</td>
<td>34.23</td>
<td>50.55</td>
<td>0.1356</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>55</td>
<td>79</td>
<td>0.218</td>
<td>0.455</td>
<td>0.410</td>
<td>10.78</td>
<td>34.03</td>
<td>50.45</td>
<td>0.1347</td>
</tr>
</tbody>
</table>

Obvious solution with a 100% identical model

<table>
<thead>
<tr>
<th>IX</th>
<th>IY</th>
<th>IZ</th>
<th>XFRAC</th>
<th>YFRAC</th>
<th>ZFRAC</th>
<th>XORT</th>
<th>YORT</th>
<th>ZORT</th>
<th>DENS</th>
<th>DENS/SIGMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>27</td>
<td>25</td>
<td>0.262</td>
<td>0.224</td>
<td>0.131</td>
<td>12.98</td>
<td>16.72</td>
<td>16.15</td>
<td>0.1388</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>43</td>
<td>42</td>
<td>0.114</td>
<td>0.361</td>
<td>0.219</td>
<td>5.64</td>
<td>26.98</td>
<td>26.90</td>
<td>0.1180</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>37</td>
<td>83</td>
<td>0.118</td>
<td>0.307</td>
<td>0.434</td>
<td>5.86</td>
<td>22.95</td>
<td>53.38</td>
<td>0.1151</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>49</td>
<td>56</td>
<td>0.218</td>
<td>0.409</td>
<td>0.294</td>
<td>10.80</td>
<td>30.53</td>
<td>36.14</td>
<td>0.1136</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>27</td>
<td>32</td>
<td>0.238</td>
<td>0.227</td>
<td>0.168</td>
<td>11.78</td>
<td>16.93</td>
<td>20.61</td>
<td>0.1125</td>
</tr>
</tbody>
</table>

Possible solution - needs further testing

Ian Stokes-Rees, http://sbgrid.org
Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data.

This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found here.

A narrated tutorial illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: Small Interfering RNA (siRNA)

Double-stranded RNA is often a sign of trouble. Our transfer RNA and ribosomes do contain little hairpins that are double-stranded, but most of the free forms of RNA, messenger RNA molecules in particular, are single strands. Many viruses, however, form long stretches of double-stranded RNA as they replicate their genomes. When our cells find double-stranded RNA, it is often a sign of an infection, and they mount a vigorous response that often leads to death of the entire cell. However, plant and animal cells also have a more targeted defense that attacks the viral RNA directly, termed RNA interference.

- More ...
- Previous Features
Molecular Replacement Structures

Advanced Keyword Query for: molecular replacement

Crystal structure of rGdd(CGCGCG) forming hexamer Z-DNA duplex with 5'-(rG) overhang

Release Date: 25-Feb-2003  Exp. Method: X Ray Diffraction
Resolution: 1.54 Å
Classification: DNA
Polymer: 1  Molecule: 5'-R(G)D(*CP*GP*CP*GP*CP*G)-3'  Chains: A,B
Authors: Rees, B., Sundaralingam, M.

21583
Antibodies can help with crystallization of soluble proteins, RNA and membrane proteins (by making rigid crystal packing contacts)

Once crystallized antibodies can be placed in the unit cell by molecular replacement and will provide phasing information about the studied molecule (RNA, membrane protein, etc)

855 models of antibody are deposited in the Protein Data Bank. Due to hinge variability it is difficult to predict which antibody molecule will provide a clear molecular replacement solution.

Solution: try all of them
MolRep: Antibodies (II)

855 models - which one to use?


crystals, diffraction images were obtained by microcrystallography. The structure of the $\beta_2$AR365–Fab5 complex was solved by molecular replacement, using separate constant and variable Fab domain structures as search models. Coordinates and structure factors are deposited in the Protein Data Bank (accession codes 2R4R for $\beta_2$AR365–Fab5 and 2R4S for $\beta_2$AR24/365–Fab5).
Example from Harrison Lab, Harvard Medical School

Our roadmap:

- Expand the Antibody Library to incorporate new structures
- Setup computations through a portal
- Configure molecular replacement applications with more advanced options (e.g. rigid body refinement).


Ian Stokes-Rees, http://sbgrid.org
CASE 2:

Blind Molecular Replacement

Welcome to SCOP: Structural Classification of Proteins.
1.73 release (November 2007)

34494 PDB Entries, 1 Literature Reference, 97178 Domains. (excluding nucleic acids and theoretical models).
Folds, superfamilies, and families [statistics here].
New folds, superfamilies [families].
List of obsolete entries and their replacements.

Ian Stokes-Rees, http://sbgrid.org
Alternative: Experimental Phasing

Transglycosylase

Lysozyme

PHAGE

3 monthsToo Slow!

Ian Stokes-Rees, http://sbgrid.org
Objectives and Priorities
Grid Computing for Biologists

 Ease of use is number one concern
 - Portal infrastructure
 - Single Sign On
 - Single point of access

 Integration of diverse resources
 - Local VDT/OSG-based clusters
 - Local SGE clusters
 - OSG
 - Member labs’ compute resources

 Secure processes and data

 Data availability
 - WebDAV
 - SCP/SFTP
 - HTTP(S)

 Storage management
 - Meta-data facilities
 - File catalogue

 Advanced Users
 - APIs
 - Scripting
 - Workflows
Thank you! Questions?

Ian Stokes-Rees, Research Associate
SBGrid, Harvard Medical School
ijstokes@crystal.harvard.edu

http://sbgrid.org

Check out our website and email us with any questions.