Applying Parallel Computing to Quickly Find the Solution for Marginal-Quality Data

Zheng-Qing Fu

- SERCAT, APS, Argonne National Lab., Argonne, IL 60439
  - University Of Georgia, Athens, GA 30602

NECAT Workshop, APS Users Week, 05-06-2008
From gene to final structure, crystallographic analysis of protein structures is a complicated Multi-Step, Multi-Discipline, Costly, and Systematic Engineering Project.

Due to the vigorous research and advances in biochemistry, molecular biology, crystallography, detecting technology, computer hardware and software development, it is much easier to solve a biological macromolecular structure nowadays.

However, challenging cases are still there for one or the other reasons.
For Challenging Cases, Strategies are needed!
Ideal is to Quickly & Dynamically Monitor the Process On-the-fly

Parallel computing is good on solving complicated problems, and can help to approach the ideal.

Data Processing

Reduced Data: *.sca files

Molecular Information: *.fasta

Space Group Determination

HA/S Searching

Handedness Determination

Initial Experimental Phasing

Density Modification

Auto-Tracing

Density Fitting and Structure Refinement

1. Structure solving is a complicated intensive-computational process. Even experienced crystallographer can make careless mistakes, leading to the failure of the whole process.

Parallel computing can optimize the whole process, quickly identify and avoid such experimental mistakes that can happen at every step. It could improve the success rate and also accuracy of the final structure.
2. The Structural Biology in the post-genomics era challenges the investigators to identify and concentrate on ever larger, more difficult and biologically important macromolecules or complexes.

The current none-parallel, try-and-error practice to solve a structure is very time-consuming and tedious. It requires a lot working knowledge in crystallography, mastering various kinds of computational programs, and experiences in using different recipes.

Parallel computing would save a lot time and troubles, and let the structural biologists concentrate on the functional problems.
How parallel computing can help?

3. For challenging cases, it would take greater efforts, longer time and the expertise to find the right combination of programs and the parameters in order to produce a traceable map.

Very often, the question come up after a few unsuccessful trials. Have I tried enough? Or, should I give up, and go back to the wet lab to improve the crystal quality?

The parallel computing can quickly answer this question by automatically and systematically searching the whole program and parameter space.
4. It is feasible to build an automatic & high-efficient **Parallel Workflow Engine** for the structure-solving process.

**Crystallographically:** Due to the vigorous research and development of crystallographic methods, computational algorithms and programs in the last several decades, there are many different computer programs available for each of the steps to solve structures.

**Computationally:** In the last decade or so, advances in computer technology have made powerful computers, especially the high-performance Linux clusters, much cheaper and ever more affordable to a single lab.
What features such a parallel workflow engine should have?

1). Capable to search and deploy different advanced crystallographic methods, algorithms and computer programs to find the optimal combination of programs for the given data.

2). Capable to search and find the optimal combination of parameters for each program to produce the best solution.

3. Capable to utilize the powerful, high-performance computing facility such as Linux cluster.

4. Capable of parallel computing to find the best solution faster.

5. Capable of automatically and dynamically search and find a optimal path to the best solution.

6. Easy to use GUI.
Parallel Workflow Engine

to automatically, systematically and quickly search program and parameter spaces to find the best solution for given data.

Figure 1. The dark blocks represent parallel tasks dynamically generated from various crystallographic computing programs with different parameter settings. The tasks are distributed by workflow engine to the computing facility and run parallel. Upon completion, the workflow engine will harvest and analyze the results, and dynamically create and start another group of tasks for the next step. And so on, until the whole process finishes.

The ideal of scientific computing is to make software

**PUBS** (powerful, upgradable, beautiful, simple)

---

Virtual demo of SGXPRO

SERCAT, APS, Argonne National Lab., Argonne, IL 60439
- University Of Georgia, Athens, GA 30602
Automatic, Better and Faster Novel Structural Solutions

What needed:

1). Reduced Data File(s) *.sca

2). Sequence File *.fasta

An all-ALA sequence will be automatically generated if no sequence given, or you just want to build main chain.
Automatic, Better and Faster Novel Structural Solutions

Click to start “Novel Structure Solution” GUI
Automatic, Better and Faster Novel Structural Solutions

Read in Unit Cell from *.sca File
Automatic, Better and Faster Novel Structural Solutions

Read in real sequence or create all-ALA sequence. Then, click on ‘Estimate Possible Solvent Contents’
Automatic, Better and Faster Novel Structural Solutions

Select phasing method: SAD, MAD/SAD.
Load the *.sca file(s).
For MAD data, both MAD and SAD will be tried.

Click on ‘Save’
Automatic, Better and Faster Novel Structural Solutions

Green: Ready to Go!

Click to RUN
Automatic, Better and Faster Novel Structural Solutions

**Pink:** step finished.

**Green:** To Go

- solvent content estimation
- heavy-atom sites search (SHELXD, SOLVE)
- handedness test (ISAS)
- initial phasing (ISAS, SOLVE)
- density modification (ISAS, DM, SOLOMON)
- auto-tracing (RESOLVE, MAID, ARP/WARP)

**Featuring:**
- Parallel computing.
- Automatic programs/parameters space search.
- Dynamic tasks generation.
- Automatic inter-task communication.
- Automatic results harvesting and analysis.
Automatic, Better and Faster Novel Structural Solutions

Upon finished, the best solution is picked up and its density map with auto-traced model are automatically presented.
Automatic, Better and Faster Novel Structural Solutions

Easy to read summary of the top five solutions can be opened up in the built-in editor.
BLAST-Assisted Automatic MR Structural Solutions

What needed:

1). Sequence File  *.fasta
    or
    Your own model in PDB format file.

2). Reduced Data File  *.sca
Automatic, Better and Faster Novel Structural Solutions

Click to start “Molecular Replacement” GUI
BLAST-Assisted Automatic MR Structural Solutions

Browse in *.fasta sequence file
Automatic ‘BLAST’ search against the NCBI data base to generate a list of the sequence-homologues sorted by the LIS score.
BLAST-Assisted Automatic MR Structural Solutions

Select a possible homologue, which will be automatically downloaded from PDB site.
BLAST-Assisted Automatic MR Structural Solutions

Click on ‘Save’
You may have to manually check the PDB file(s) inside the ‘pdb’ directory to make sure it only contains the model you want.

Read in data from a *.sca file

File Edit Format Windows Help

Database: PDB, Protein-Drug, Protein-Protein, PDB, and PDB

Options for Advanced Blasting

Run Mode: Local

Search: Blast Search of Sequence-Homologue Structures and Molecular Replacement

Download PDB File: Browse

Reset Parameters: Reset All

Save Project: Save Project As

Program Config:

Model PDB File: /home/fzol/demo10/pdb/1h00.pdb

Tip 1: Select an entry above and click ‘Download PDB File’ to download the model from PDB. Then, ‘Add’ it to the MR Model list.

Tip 2: Before clicking on ‘Save’, edit the PDB file(s) to make sure only the models you want are included!

Model PDB File: /home/fzol/demo10/pdb/1h00.pdb

PDB ID: 1h00

Entry: 0.900

Space Group: P1

Cell: 78.089, 78.089, 78.089

Resolution: 99.00, 99.00, 99.00

Weight: 27.61

Value: 3.00

Save: /home/fzol/demo10/1h00_s.iso.sca
BLAST-Assisted Automatic MR Structural Solutions

Green: ready to Go!

Click to RUN
BLAST-Assisted Automatic MR Structural Solutions

Programs with different algorithms employed to find a better solution if any:

1). PHASER
2). EPMR
3). AMORE

Featuring:
- Parallel computing.
- Automatic programs/parameters space search.
- Dynamic tasks generation.
- Automatic inter-task communication.
- Automatic results harvesting and analysis.
BLAST-Assisted Automatic MR Structural Solutions

*Pink: tasks finished*
BLAST-Assisted Automatic MR Structural Solutions

Summary of results ready to be opened up with the built-in editor.
Other Useful Tools

1). Space Group Determination

Command:
spgr4d  D6_2_0001.x

Here D6_2_####.x are files from DENZO/HKL2000 integration.

... ...

Result:
Suggested Space Group:  P61 / P65
Other Useful Tools

2). Instant Solvent Content Estimation
Other Useful Tools

3). $f', f''$ vs. $\lambda$ / Energy Plot
Helpful to plan ahead MAD/SAD data collection at synchrotron beamlines

Click on the curve will point to the corresponding data values in the table on the left.
Other Useful Tools

4). Parallel SHELXD

A. Browse in the *.sca file

B. Give a number of sites to search.
   If not known, assign a reasonably large number. Only the good sites will be picked up based on the analysis by the built-in filters.
Other Useful Tools

4). Parallel SHELXD

Parallel SHELXD would allow users to quickly (in a minute on a small 16-node Linux cluster) find if there is any anomalous signal or the anomalous signal is good enough to find heavy atoms.
Other Useful Tools

4). Parallel SHELXD

Searching heavy atom sites is the first step in novel-structure-solving process. The correlation coefficients CC-ALL/CC-WEAK above 35.0/above 10.0 may indicate a possible solution.

If not, you might have to check your data collection (wavelength, beam intensity, exposure time, redundancy …), data processing etc. to identify and solve those experimental problems that you can’t do anything about after your data collection is done.
Virtual demo of SGXPRO

The End
A Recent Case

Molecule: 95 AAs, 4S
Space Group: P4₂
Diffraction: to 2.4Å

1ˢᵗ Batch Data: 360°, Nred=13, no solution!

2ⁿᵈ Batch Data from same crystal: another 360°, Nred=12.5

1ˢᵗ + 2ⁿᵈ Batches Combined: Solution! with 62.1% automatically built.
### Some of the “difficult” Structures Recently Solved On-Site

<table>
<thead>
<tr>
<th>Name</th>
<th>NaaAU</th>
<th>HA</th>
<th>Nha</th>
<th>SpaceGroup &amp; UC(a,b,c)</th>
<th>Nw</th>
<th>Reso(Å)</th>
<th>Built%</th>
<th>Time(’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssos_03</td>
<td>225x1</td>
<td>Se</td>
<td>3</td>
<td>P61 (105.9, 105.9, 51.9) SAD 2.1</td>
<td>92.0</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_04</td>
<td>277x1</td>
<td>Se</td>
<td>8</td>
<td>P212121 (53.6, 59.8, 69.7) SAD 1.6</td>
<td>96.8</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_05</td>
<td>214x1</td>
<td>Se</td>
<td>4</td>
<td>P212121 (34.7, 61.0, 84.4) SAD 1.8</td>
<td>94.4</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_06</td>
<td>138x2</td>
<td>Se</td>
<td>9</td>
<td>P212121 (57.2, 94.7, 102.0) SAD 2.6</td>
<td>82.6</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_01</td>
<td>?</td>
<td>Se</td>
<td>?</td>
<td>P212121 (49.2, 56.7, 98.8) MAD ?</td>
<td>?</td>
<td>?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_02</td>
<td>1370x1</td>
<td>Se</td>
<td>12</td>
<td>P21 (77.9, 123.8, 84.8) SAD 3.6</td>
<td>52.2</td>
<td>58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_07</td>
<td>311x4</td>
<td>Se</td>
<td>8</td>
<td>P6122 (250.9, 250.9, 121.0) SAD 2.4</td>
<td>92.4</td>
<td>43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_08</td>
<td>241x2</td>
<td>Se</td>
<td>10</td>
<td>P61 (122.1, 122.1, 109.1) MAD 2.6</td>
<td>87.1</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_09</td>
<td>245x2</td>
<td>Se</td>
<td>12</td>
<td>F23 (206.5, 206.5, 206.5) SAD 2.7</td>
<td>77.5</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_10</td>
<td>*850?</td>
<td>Se</td>
<td>9</td>
<td>P4232 (155.2, 155.2, 155.2) SAD 3.0</td>
<td>*397</td>
<td>38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_11</td>
<td>258x2</td>
<td>S</td>
<td>16</td>
<td>P43212 (88.3, 88.3, 126.8) SAD 2.0</td>
<td>95.3</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_12</td>
<td>95</td>
<td>S</td>
<td>4</td>
<td>P42 (53.5, 53.5, 41.3) SAD 2.4</td>
<td>62.1</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_13</td>
<td>350x2</td>
<td>S</td>
<td>44?</td>
<td>P21 (42.0, 169.8, 69.3) SAD 2.1</td>
<td>?</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_m1</td>
<td>420x1</td>
<td>x</td>
<td>x</td>
<td>P212121 (75.5, 87.7, 152.6) MR 2.0</td>
<td>x</td>
<td>360</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_m2</td>
<td>311x1</td>
<td>x</td>
<td>x</td>
<td>P4132 (135.7, 135.7, 135.7) MR 1.9</td>
<td>x</td>
<td>185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ssos_m3</td>
<td>163x1</td>
<td>x</td>
<td>x</td>
<td>P3121 (68.8, 68.8, 79.8) MR 1.8</td>
<td>x</td>
<td>49</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

… …
Example ssos_07

Dozens of SAD/MAD data sets collected from 4 trips in half years:
1). 08/24/2006 (2.30Å); 2). 10/11/2006 (2.37Å)
3). 12/18/2006 (2.60Å); 4). 02/01/2007 (2.60Å)

The structure was finally solved on-site in early March with one of the data sets collected on October 10, 2006: 311aa x 4.
First no solution, Then the SGXPRO helped quickly guiding to an unexpected structure solution through intensive automatic and systematic program and parameter search.

850aa  17 Se-Met
3.0Å
P213   a=b=c=155.227
63% Solvent

560aa   9 Se-Met?
51% Solvent
Phased, but Unsolved Case!
350x2 AAs, P2₁, 2.1Å, 44S?    November, 2007
Acknowledgment

Dr George Wu, Dr. Dongsheng Che, Dr. Jizhen Zhao, Dr. Feng Sun, Dr. Haijin Yan, Dr. Jay Zhang, Dept. of Computer Sciences, UGA
Dr. Joe Zhu, Dept. of Physics, UGA

Dr. Bi-Cheng Wang, Dr. John Rose, UGA
Dr. John Chrzas, Dr. Jim Fait, Dr. Zhongmin Jin, John Gonczy, UGA, SER-CAT, APS, Argonne National Lab.
Dr. Andy Howard, Illinois Institute of Technology

Dr. George Sheldrick, Univ. of Göttingen, Germany