US-SOMO Cluster Methods: Year One Perspective

Emre Brookes
Raminderjeet Singh
Marlon Pierce
Suresh Marru
Borries Demeler
Mattia Rocco

23 July 2013

XSEDE 2013
San Diego, California
Outline

• Background
  • Software
  • Science Gateway

• Community
  • Workshops
  • Followups

• Science Enabled
  • Salicylidene acylhydrazide (Byron et al)
  • Intrinsic Disorder of Single stranded DNA binding protein (Scott et al)
  • Smae-Diablo (Vachette et al)
  • Fibrinogen studies (Rocco et al)
  • Parsimonious models (Brookes)

• Future
US-SOMO

- Website: somo.uthscsa.edu
- GUI based: Windows, OSX, Linux, source code (GPL).
  - C++, Qt4
- Brookes et al, UltraScan Solution Modeler: Integrated Hydrodynamic Parameter and Small Angle Scattering Computation and Fitting Tools, *ACM XSEDE 2012*
  - “The overarching goal of our software is to provide an extensible general framework for generating collections of candidate structures from an initial structure or structures, modeling candidate structures under various experimental methods and conditions, and subsequently globally fitting and screening candidate structure's models against sets of experimental data”

XSEDE 2013 Brookes, E. *et. al.*
US-SOMO: Year one perspective
US-SOMO

- Bead model generation from atomic structures
- Hydrodynamic parameter calculation
- SAS
- Batch & cluster computations
  - Cluster submission directly from GUI
    - Apache Airavata middleware
    - IU quarry node
- DMD on cluster for expanding conformation space
  - Run on UTHSCA's Alamo cluster
• Discrete Molecular Dynamics

SAXS/SANS Plotting Functions

Load Atom Definition File: 1HEL s equi to 6 tm10000 m-9
Load Hybridization File: somo_atom
Load SAXS Coefficients File: somo_soms_atoms

SAXS/SANS I(q) Plotting Functions:
- SAXS: Full, H3, Fast, FoXS, Crysol
- SANS: Full, H3, Fast, Cryson

File suffix: fd

Compute SAXS Curve
- Load SAXS Curve
- Load Plotted
- Set Grid
- Clear SAXS Curve
- Legend

Load GNOM File
- Search
- Guinier Analysis
- Create standard output files
- Guinier plot
- q^2 range
- Standard: Kratky plot
- q range: 1

P(r) vs r Plotting Functions
- Bin size (Angstrom): 2
- Raw, SAXS, SANS, Normalize
- Residue contrib. range (Angstrom)
- Display

Compute P(r) Distribution
- Load P(r) Distribution
- Load Plotted
- Clear P(r) Distribution
- Legend

Model: 50
results_1qq_fd.csv /root/ultrascan/somo/cluster/results/dmd1/1HEL_s_equi
Model: 6
results_1qq_fd.csv /root/ultrascan/somo/cluster/results/dmd1/1HEL_s_equi
Model: 7
results_1qq_fd.csv /root/ultrascan/somo/cluster/results/dmd1/1HEL_s_equi
Model: 8
results_1qq_fd.csv /root/ultrascan/somo/cluster/results/dmd1/1HEL_s_equi
Model: 9
results_1qq_fd.csv /root/ultrascan/somo/cluster/results/dmd1/1HEL_s_equi

I(q) plot done
US-SOMO / Cluster Packaging

Create file for cluster jobs

<table>
<thead>
<tr>
<th>Grid from experimental data:</th>
<th>Add experimental data files</th>
<th>/root/ultrascan/somo/saxs/lyzexp.dat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clear experimental data files</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of jobs (maximum 51):</th>
<th>51</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Output base name (job identifier)</th>
<th>job</th>
</tr>
</thead>
</table>

- Package for parallel job submission
- Individual jobs for each grid
- DMD settings
- Advanced options

Create cluster job package
Submit jobs for processing
Check job status / Retrieve results
Extract results

File

Number of selected files: 51
Options summary: I(q) curves
### Submit jobs to cluster

**Available jobs**

<table>
<thead>
<tr>
<th>Name</th>
<th>Created</th>
<th>Size</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>me2-alamotar</td>
<td>Wed Nov 9 05:18:29 2011</td>
<td>112640 bytes</td>
<td></td>
</tr>
<tr>
<td>me3-lonestartar</td>
<td>Wed Nov 9 05:17:01 2011</td>
<td>112640 bytes</td>
<td></td>
</tr>
<tr>
<td>me3-ranger.tar</td>
<td>Wed Nov 9 05:17:01 2011</td>
<td>112640 bytes</td>
<td></td>
</tr>
<tr>
<td>testixtar</td>
<td>Thu Feb 9 14:37:19 2011</td>
<td>127954 bytes</td>
<td></td>
</tr>
</tbody>
</table>

**Systems**

- alamo
- lonestar
- lonestar:12-core
- lonestar-cuda
- ranger

Select all jobs | Remove selected jobs | Submit selected jobs

File

**THIS WINDOW IS UNDER DEVELOPMENT.**

---

XSEDE 2013 Brookes, E. *et al.*

US-SOMO: Year one perspective
US-SOMO / One possible “workflow”

- Starting with model(s)
- Expand conformation space
- Compute hydrodynamic parameters &/or SAXS curves
- Compare to experimental data
- Identify likely models
- Process locally or on clusters
Workshops

- 5 workshops so far
- 2x International Analytical Ultracentrifugation Symposium
- Harwell Research Complex, Oxfordshire
- La Trobe University, Melbourne
- UltraScan annual workshop, San Antonio

<table>
<thead>
<tr>
<th>Event</th>
<th>Date</th>
<th>Attendees</th>
<th>Duration</th>
<th>Cluster usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAUCS #1</td>
<td>26 March '12</td>
<td>12</td>
<td>8 h</td>
<td>Demo only</td>
</tr>
<tr>
<td>IAUCS #2</td>
<td>27 March '12</td>
<td>18</td>
<td>8 h</td>
<td>Demo only</td>
</tr>
<tr>
<td>HRC</td>
<td>5 October '12</td>
<td>6</td>
<td>6 h</td>
<td>Demo only Follow up usage</td>
</tr>
<tr>
<td>LU</td>
<td>15-16 November '12</td>
<td>17</td>
<td>12 h</td>
<td>Yes, by all attendees</td>
</tr>
<tr>
<td>UltraScan</td>
<td>3-4 June '13</td>
<td>19</td>
<td>12 h</td>
<td>Yes*</td>
</tr>
</tbody>
</table>
Science Enabled / Fibrinogen

- Mattia Rocco et al.
- Fibrinogen is an important component of the coagulation cascade, as well as a major determinant of blood viscosity and blood flow
- A centrosymmetric dimer made by 3 pairs of chains
- US-SOMO/DMD simulations of the conformational variability for comparison to experimental data
Science Enabled / Salicylidene acylhydrazides

- Kate Beckman et al, University of Glasgow
- Salicylidene acylhydrazides inhibit virulence of E. coli O157
- Tpx
- US-SOMO used for hydrodynamic parameter calculations & DMD

XSEDE 2013 Brookes, E. et. al.
US-SOMO: Year one perspective

Images credit: Olwyn Byron
Science Enabled / Smac-Diablo

- Group of M. Milani, University of Milano
- Smac-Diablo a dimeric protein involved in apoptosis (programmed cell death)
- chain: 192 residues, MM= 21.8 kDa
- Final refinement of model by addition of N- and C-termini using US-SOMO/DMD

Image credit: Patrice Vachette
Science Enabled / Single Stranded DNA Binding Protein (SSB)

- Mathew Green et al, University of Nottingham
- SSB binds individual strands of DNA
- Critical role in DNA metabolism: Replication, recombination & repair
- Intrinsically disordered
- US-SOMO/DMD used to create conformations to screen against SAXS data
## SE: Parsimonious spatial models

### SAXS Curve

![SAXS Curve](image)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1CTS_1-db_1sa-10_1</td>
<td>7.530e-07</td>
<td>2.20</td>
<td>5.69</td>
<td>5.69</td>
<td>5.69</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>1CTS_1-db_2sa-10_1</td>
<td>7.480e-07</td>
<td>2.34</td>
<td>7.62</td>
<td>5.04</td>
<td>5.04</td>
<td>1.51</td>
<td>1.51</td>
<td>1.00</td>
</tr>
<tr>
<td>1CTS_1-db_3sa-10_1</td>
<td>7.410e-07</td>
<td>2.34</td>
<td>7.42</td>
<td>5.06</td>
<td>5.06</td>
<td>1.47</td>
<td>1.47</td>
<td>1.00</td>
</tr>
<tr>
<td>1CTS_1-db_4sa-10_1</td>
<td>6.910e-07</td>
<td>2.39</td>
<td>7.39</td>
<td>8.62</td>
<td>5.52</td>
<td>1.34</td>
<td>0.86</td>
<td>1.56</td>
</tr>
<tr>
<td>1CTS_1-db_5sa-10_1</td>
<td>7.430e-07</td>
<td>2.31</td>
<td>7.43</td>
<td>6.37</td>
<td>5.06</td>
<td>1.47</td>
<td>1.17</td>
<td>1.26</td>
</tr>
<tr>
<td>1CTS_1-db_6sa-10_1</td>
<td>6.900e-07</td>
<td>2.38</td>
<td>7.32</td>
<td>8.50</td>
<td>5.52</td>
<td>1.33</td>
<td>0.86</td>
<td>1.54</td>
</tr>
<tr>
<td>1CTS_1-db_7sa-10_1</td>
<td>7.390e-07</td>
<td>2.32</td>
<td>7.48</td>
<td>5.12</td>
<td>6.09</td>
<td>1.23</td>
<td>1.46</td>
<td>0.84</td>
</tr>
<tr>
<td>1CTS_1-so</td>
<td>7.098e-07</td>
<td>2.32</td>
<td>7.46</td>
<td>6.31</td>
<td>5.52</td>
<td>1.35</td>
<td>1.18</td>
<td>1.14</td>
</tr>
</tbody>
</table>

XSEDE 2013 Brookes, E. *et al.*
US-SOMO: Year one perspective
Future

- EPSRC/NSF: *SI2-CHE*: **CCP-SAS** – Collaborative Computing consortium for advanced analysis of structural data in chemical biology and soft condensed matter
  - UK Lead PI Steve Perkins, UCL; US Lead PI Paul Butler, NIST
    - Open source
    - SASSIE
    - Perkins code
  - Framework for scientific application development
    - generate GUI & web based versions from a single set of minimal text based description files
      - GUI required for secure data
      - needs to be simple enough for a scientist to wrap their own applications
      - Current base python, c, fortran, c++ (qt)
    - local & “cloud/grid” computations
    - big data
    - if something already exists will adopt/adapt
      - otherwise, we will have to create
  - ½ a postdoc
Acknowledgments

- To E. Brookes
  - NIH NIGMS 1K25GM090154 / PI
  - NSF OCI-1032742 (Pierce) / co-PI
  - NSF CHE-1265817 / PI
  - NSF XSEDE

Funded by a grant from the National Institute of General Medical Sciences of the National Institutes of Health

XSEDE 2013 Brookes, E. et. al.
US-SOMO: Year one perspective