Planning and carrying out automated structure determination using SAD phasing

SBGrid/NECat Phenix Workshop
Boston, November 10, 2016

Tom Terwilliger, Li-Wei Hung, Los Alamos National Laboratory
Randy Read, Airlie McCoy, University of Cambridge
Pavel Afonine, Paul Adams, Lawrence Berkeley National Laboratory
Steps in Single Wavelength Anomalous Diffraction (SAD) Structure Determination

• Plan the experiment
• Measure the data
• Scale the data
• Evaluate the accuracy of the anomalous differences
• Find the anomalous sub-structure
• Identify hand of sub-structure
• Calculate experimental phases and a map
• Improve the map with density modification
• Build and refine a model
Planning a SAD experiment

Will I find the sites of anomalously-scattering atoms?
Planning a SAD experiment

How many sites?
How many reflections?

What is the anomalously-scattering atom?
What is the wavelength?

Are the sites (on average) well ordered?
Are the data well-measured?
Maximizing the anomalous signal and the anomalous correlation

The **anomalous correlation** $CC_{\text{ano}}$ is a measure of the accuracy of each anomalous difference (correlation to ideal anomalous data from your structure).

The **anomalous signal** is a measure of how much total information per site is present in the anomalous differences (peak height in anomalous difference Fourier).
Anomalous difference Fourier with observed data and model phases

Peak height in anomalous difference Fourier at coordinates of anomalously-scattering atoms

Typical values of $S_{ano}$ for solved datasets: 10-20

Anomalous signal
What determines if I will find the sites?

- **Accuracy of data**
- **Anomalous signal**
- **Number of reflections**

\[
< S_{\text{ano}} > = CC_{\text{ano}} \frac{N^{1/2}_{\text{refl}}}{f^{1/2}n^{1/2}_{\text{sites}}}
\]

- **B-value for anomalous sub-structure**
- **Are sites ordered?**
- **Will I find sites?**

**Number of sites**
How big will my anomalous signal be?

Expected value of anomalous signal $S_{ano}$

$$< S_{ano} > = CC_{ano} \frac{N_{refl}^{1/2}}{f^{1/2} n_{sites}^{1/2}}$$

$f$ is 2\textsuperscript{nd} moment of the anomalous scattering factor

$(f$ is large if B-value for anomalously-scattering atoms is high)

$$f = \frac{< (f^h)^2 >}{< f^h >^2}$$

$$f^h \equiv f^* e^{-B (\sin^2 \theta_h / \lambda^2)}$$

Perfect data (20,000 reflections, 8 sites): $S_{ano} = (20000/8)^{1/2} = 50$

Good data (overall $CC_{ano}=0.36$ $f=2.0$): $S_{ano} = 12.6$
Checking our simple model for anomalous signal

\[ < S_{\text{ano}} > = CC_{\text{ano}} \frac{N_{\text{refl}}^{1/2}}{f^{1/2} n_{\text{sites}}^{1/2}} \]

- **\( CC_{\text{ano}} \):** Correlation of anomalous differences with model differences
- **\( S_{\text{ano}} \):** Peak height in model-phased difference Fourier

218 SAD datasets 1.2 – 4.5 Å

\[ y = 1.03x \]

\[ R^2 = 0.91 \]
CysZ multi-crystal sulfur-SAD data

Qun Liu, Tassadite Dahmane, Zhen Zhang, Zahra Assur, Julia Brasch, Lawrence Shapiro, Filippo Mancia, Wayne Hendrickson (2012). Science 336, 1033-1037

Data from 7 crystals collected at wavelength of 1.74 Å to resolution of 2.3 Å

Can anomalous signal tell us which merged datasets will be solved?
CysZ multi-crystal sulfur-SAD data
(Hyss LLG brute-force substructure determination)
CysZ multi-crystal sulfur-SAD data
(Hyss LLG brute-force substructure determination)
CysZ single-crystal sulfur-SAD data

Crystal 6   AutoSol $R/R_{free}=0.24/0.27$
phenix.plan_sad_experiment
Design an experiment that will give you enough anomalous signal

Choose I/sigI, estimate normalized errors and $CC_{\text{ano}}$

Anomalous signal

$< S_{\text{ano}} > = CC_{\text{ano}} \frac{N^{1/2}_{\text{refl}}}{f^{1/2} n^{1/2}_{\text{sites}}}$

Anomalous correlation

Number of reflections

Guess from sequence

Number of sites

Choose $d_{\text{min}}$, guess B

B-value for anomalous sub-structure
Will I solve my structure?

Simulate experiment with 
*phenix.plan_sad_experiment* based on:

- \( \frac{I}{\sigma} \) (errors in measurement)
- Anomalously-scattering atom, wavelength (\( f" \))
- Sequence (other atoms)
- Resolution of data
- Number of sites
Expected anomalous signal (S)

300 residues, 5 Se atoms

\[ \lambda = 0.9792 \, \text{Å} \quad d_{\text{min}} = 3 \, \text{Å} \]

\[ \langle S_{\text{ano}} \rangle = C \frac{C_{\text{ano}}}{f^{1/2}} \frac{N^{1/2}}{n_{\text{sites}}} \]

Anomalous signal depends on I/sigl
Anomalous signal depends on $f''$ (S vs Se)

$$< S_{ano} >= C C_{ano} \frac{N_{refl}^{1/2}}{f^{1/2} n_{sites}^{1/2}}$$
Estimating the anomalous signal before collecting the data

\[ \langle S_{ano} \rangle = CC_{ano} \frac{N^{1/2}}{f^{1/2} n^{1/2}} \]

- Choose \( l/sigl \), estimate normalized errors and \( CC_{ano} \)
- Anomalous correlation
- Number of reflections
- B-value for anomalous sub-structure
- Choose \( d_{\text{min}} \), guess B

\[ y = 0.99x \]
\[ R^2 = 0.58 \]

Predicted anomalous signal
\( \text{(phenix.plan_sad_experiment)} \)
Estimating the anomalous signal after collecting the data
Finding the anomalous sub-structure
Using the SAD likelihood function to find sites

“The likelihood of measuring the observed anomalous data given a potential sub-structure”
Using the SAD likelihood function to find the anomalous sub-structure

Start with guess about the anomalous sub-structure

From anomalous difference Patterson
Random
Any other source

Find additional sites that increase the likelihood

LLG completion based on log-likelihood gradient maps*
Iterative addition of sites

Related to using an anomalous difference Fourier—but better

LLG sub-structure searches in HySS

Test cases

164 SAD datasets from PDB (largely JCSG MAD data)

Using peak, remotes, inflection as available to include data with low anomalous signal
Finding anomalous substructure with LLG completion

- **Guess 2-site solutions**
- **Peaks from Patterson**
- **Extrapolation**
  - Direct methods
  - Phaser LLG completion
- **Scoring**
  - Correlation Phaser LLG

- **Range of resolution**
- **Variable number of Patterson solutions**

- **Adjustable LLGC_SIGMA**
  (cut-off for peak height)

- Use LLG score to compare solutions
- Terminate early if same solution found several times
- Run quick direct methods first
LLG Sub-structure Search

Anomalous signal indicates if a dataset can be solved
Optimizing scaling and merging of SAD data

(phenix.scale_and_merge)
Why F+, F- differ from one crystal to another

Errors in measurement ($\sigma_{\text{obs}}$)

Crystals really are different ($\sigma_{\text{crystal}}$)
Optimizing estimates of $F^+, F^-$

Crystal 1
$F^+, F^-$

Local scaling to reduce systematic errors

Use of $\sigma_{\text{crystal}}$ in weighting

Crystal 2
$F^+, F^-$
Applying inter-dataset variances in weighting

Weighting for data from an individual crystal:

$$\sigma^2_{\text{total}} \approx \sigma^2_{\text{obs}} + \sigma^2_{\text{crystal}}$$
Improvement in anomalous correlation using local scaling in *phenix.scale_and_merge*
<table>
<thead>
<tr>
<th>Set</th>
<th>PDB</th>
<th>Atom</th>
<th>Expt</th>
<th>Resolution (Å)</th>
<th>Sites</th>
<th>Molecule</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>#3</td>
<td>3TRZ</td>
<td>ZN</td>
<td>SAD</td>
<td>2.8</td>
<td>12</td>
<td>Lin28/let-7d microRNA complex</td>
<td>(Nam et al., 2011)</td>
</tr>
<tr>
<td>#97</td>
<td>1XBN</td>
<td>FE</td>
<td>MAD</td>
<td>2.6</td>
<td>1</td>
<td>bacterial nitric oxide sensor</td>
<td>(Nioche et al., 2004)</td>
</tr>
<tr>
<td>#111</td>
<td>4TSO</td>
<td>BA</td>
<td>SAD</td>
<td>2.6</td>
<td>1</td>
<td>Fluorescent RNA aptamer</td>
<td>(Warner et al., 2015)</td>
</tr>
<tr>
<td>#123</td>
<td>3M1C</td>
<td>SE</td>
<td>SAD</td>
<td>2.7</td>
<td>9</td>
<td>herpesvirus fusion regulator complex gH-gL</td>
<td>(Chowdary et al., 2010)</td>
</tr>
</tbody>
</table>
### 3TRZ (12 ZN SAD)

*Scaling of data with phenix.scale_and_merge*

<table>
<thead>
<tr>
<th>Resolution</th>
<th>CC1/2_ano</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.0</td>
<td>0.64</td>
</tr>
<tr>
<td>5.5</td>
<td>0.44</td>
</tr>
<tr>
<td>5.0</td>
<td>0.25</td>
</tr>
<tr>
<td>4.5</td>
<td>0.17</td>
</tr>
<tr>
<td>4.0</td>
<td>0.09</td>
</tr>
<tr>
<td>3.5</td>
<td>-0.01</td>
</tr>
<tr>
<td>3.0</td>
<td>-0.01</td>
</tr>
<tr>
<td>2.8</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

*phenix.scale_and_merge XDS_ASCII.HKL*
### 3TRZ (12 ZN SAD)
**Analysis of anomalous data with phenix.anomalous_signal**

<table>
<thead>
<tr>
<th>Resolution</th>
<th>CC1/2_ano</th>
<th>CCano*</th>
<th>Anomalous Signal</th>
<th>P(substr)</th>
<th>FOM*</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.0</td>
<td>0.63</td>
<td>0.66</td>
<td>12</td>
<td>76</td>
<td>0.3</td>
</tr>
<tr>
<td>5.5</td>
<td>0.59</td>
<td>0.66</td>
<td>13</td>
<td>81</td>
<td>0.3</td>
</tr>
<tr>
<td>5.0</td>
<td>0.51</td>
<td>0.63</td>
<td>15</td>
<td>91</td>
<td>0.3</td>
</tr>
<tr>
<td>4.5</td>
<td>0.42</td>
<td>0.61</td>
<td>16</td>
<td>97</td>
<td>0.3</td>
</tr>
<tr>
<td>4.0</td>
<td>0.31</td>
<td>0.55</td>
<td>17</td>
<td>99</td>
<td>0.3</td>
</tr>
<tr>
<td>3.5</td>
<td>0.15</td>
<td>0.39</td>
<td>14</td>
<td>86</td>
<td>0.3</td>
</tr>
<tr>
<td>3.0</td>
<td>0.06</td>
<td>0.26</td>
<td>11</td>
<td>73</td>
<td>0.2</td>
</tr>
<tr>
<td>2.8</td>
<td>0.05</td>
<td>0.23</td>
<td>10</td>
<td>65</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**phenix.anomalous_signal**  
 data=scaled_data.mtz scaled_data.mtz  
half_dataset_a=half_dataset_a.mtz  
half_dataset_b=half_dataset_b.mtz  
seq_file=seq.dat  
atom_type=ZN  
sites=12
3TRZ (12 ZN SAD)
Structure contains translational non-crystallographic symmetry

phenix.xtriage scaled_data.mtz

Peak in native Patterson function:
(0,0,1/3) 36% of origin
→ Strong translational NCS present

(Note: Phaser SAD LLG scoring does not yet account for tNCS)
3TRZ Hyss direct methods substructure determination

phenix.hyss nproc=48 scaled_data.mtz 12 ZN wavelength=1.2549 resolution=4.2
3TRZ Hyss LLG substructure determination (brute force)

phenix.hyss nproc=48 scaled_data.mtz 12 ZN wavelength=1.2549 \ resolution=4.2 rescore=phaser-complete strategy=brute_force
3TRZ (12 ZN SAD)
Structure solution with phenix.autosol

phenix.autosol nproc=48 sites=12 atom_type=zn
data=scaled_data.mtz lambda=1.2549
seq_file=seq_PROTEIN.dat direct_methods_only=true
3TRZ (12 ZN SAD)
Protein model building with phenix.autobuild
(594 of 939 residues correctly built with rmsd=0.54 A R/Rfree=0.34/0.37)
3TRZ (12 ZN SAD)
Protein model building with phenix.autobuild
(594 of 939 residues correctly built with rmsd=0.54 Å R/Rfree=0.34/0.37)

phenix.autobuild nproc=5 seq_file=seq.dat
data=AutoSol_run_1_/overall_best_refine_data.mtz
map_file=AutoSol_run_1_/overall_best_denmod_map_coeffs.mtz
ncs_file=AutoSol_run_1_/overall_best_ncs_file.ncs_spec
ha_file=AutoSol_run_1_/overall_best_ha_pdb.pdb
3TRZ (12 ZN SAD)
(RNA model-building R/Rfree=0.36/0.40)

phenix.autobuild nproc=5
seq_file=seq_RNA.dat
data=AutoSol_run_1/
overall_best_refine_data.mtz
map_file=AutoBuild_run_3/
overall_best_denmod_map_coeffs.mtz chain_type=RNA
solvent_fraction=0.66
input_lig_file_list=AutoBuild_run_3/placed.pdb
1XBN (Fe MAD)
Analysis of anomalous data with phenix.autosol

phenix.autosol mad.eff

Parameters file:

autosol {
    seq_file = seq.dat
    sites = 1
    atom_type = Fe
    wavelength {
        data = e1.HKL
        lambda = 1.738729
    }
    wavelength {
        data = e2.HKL
        lambda = 1.624747
    }
    wavelength {
        data = e3.HKL
        lambda = 1.740630
    }
}
1XBN (Fe MAD)
Analysis of anomalous data with SOLVE in phenix.autosol  
(AutoSol_run_1_/solve_2.prt)

**Correlation of anomalous differences between wavelengths**

<table>
<thead>
<tr>
<th>Resolution (Å)</th>
<th>CC 1 vs 2</th>
<th>CC 1 vs 3</th>
<th>CC 2 vs 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>0.84</td>
<td>0.81</td>
<td>0.75</td>
</tr>
<tr>
<td>4.0</td>
<td>0.56</td>
<td>0.56</td>
<td>0.38</td>
</tr>
<tr>
<td>3.7</td>
<td>0.26</td>
<td>0.37</td>
<td>0.2</td>
</tr>
<tr>
<td>3.5</td>
<td>0.21</td>
<td>0.33</td>
<td>0.03</td>
</tr>
<tr>
<td>3.3</td>
<td>0.07</td>
<td>0.26</td>
<td>-0.01</td>
</tr>
<tr>
<td>3.2</td>
<td>0.06</td>
<td>0.35</td>
<td>-0.02</td>
</tr>
<tr>
<td>3.0</td>
<td>-0.04</td>
<td>0.23</td>
<td>-0.03</td>
</tr>
<tr>
<td>2.9</td>
<td>-0.01</td>
<td>0.3</td>
<td>0.12</td>
</tr>
<tr>
<td>2.8</td>
<td>0.13</td>
<td>0.16</td>
<td>-0.02</td>
</tr>
<tr>
<td>2.6</td>
<td>-0.07</td>
<td>0.24</td>
<td>-0.07</td>
</tr>
</tbody>
</table>
1XBN (Fe MAD)

Analysis of anomalous data with phenix.autobuild

$R/R_{free} = 0.35/0.41$

**Command:**

```
phenix.autobuild nproc=5 seq_file=seq.dat
data=AutoSol_run_1_overall_best_refine_data.mtz
map_file=AutoSol_run_1_overall_best_denmod_map_coeffs.mtz
ha_file=AutoSol_run_1_overall_best_ha_pdb.pdb
model=AutoSol_run_1_overall_best.pdb
```
4TS0 (Ba SAD)

Analysis of anomalous data with phenix.autosol; phenix.autobuild
47 of 87 residues built with rmsd 0.54 Å. R/Rfree=0.43/0.48

phenix.autosol sites=1 atom_type=BA
data=XDS_ASCII.HKL lambda=1.5498
seq_file=seq.dat mad_ha_list='P F'

phenix.autobuild nproc=5 seq_file=seq.dat
data=AutoSol_run_1_overall_best_refine_data.mtz
map_file=AutoSol_run_1_overall_best_denmod_map_coeffs.mtz

phenix.autobuild nproc=5 seq_file=seq.dat
data=AutoBuild_run_1_overall_best_denmod_map_coeffs.mtz
Analysis of anomalous data with phenix.autosol; phenix.autobuild
698 of 865 residues built with rmsd 0.51 Å. R/Rfree=0.29/0.34
The Phenix Team

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Nigel Moriarty, Nicholas Sauter, Oleg Sobolev, Billy Poon

Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Cambridge University

Randy Read, Airlie McCoy, Gabor Bunkóczi, Robert Oeffner

Duke University

Jane & David Richardson, Christopher Williams, Bradley Hintze

An NIH/NIGMS funded Program Project