Structure solution from weak anomalous data

Diffraction Methods in Structural Biology

Gordon Research Conference
Bates College, Maine
July 30, 2014

Gábor Bunkóczi, Airlie McCoy, Randy Read (Cambridge University)
Nat Echols, Ralf Grosse-Kunstleve, Paul Adams, James Holton (Lawrence Berkeley National Laboratory)
Tom Terwilliger (Los Alamos National Laboratory)
Structure solution from weak anomalous data

Problems with weak signal
Quantifying the anomalous signal
Solving the anomalous sub-structure with weak signal
Solving structures with weak signal
Estimating the anomalous signal from the data
Weak anomalous signal

Reasons:

Few anomalous scatterers, sulfur SAD, weak diffraction, wavelength far from peak

Consequences:

Substructure identification is difficult
Phasing is poor
Iterative density modification, model-building and refinement works poorly
**Quantifying the anomalous signal**

**$CC_{ano}$: How accurate are the anomalous differences?**

Anomalous differences measured with errors $\varepsilon_j$

\[ \Delta_{obs, j}^{ano} = \Delta_{ano, j} + \varepsilon_j \]

Correlation of observed and true anomalous differences

\[ CC_{ano} \equiv \frac{\langle \Delta_{ano, j}\Delta_{obs, j}^{ano} \rangle}{\langle \Delta_{ano}^2 \rangle^{1/2} \langle \Delta_{obs, ano}^2 \rangle^{1/2}} \]

Fraction of observed anomalous differences that is noise

\[ E_{ano}^2 = \frac{\langle \sigma_{ano}^2 \rangle}{\langle \Delta_{obs, ano}^2 \rangle} \]

Expected value of $CC_{ano}$

\[ CC_{ano} \sim [1 - E_{ano}^2]^{1/2} \]
Anomalous signal $S_{\text{ano}}$: How accurate are maps based on the anomalous differences?

Anomalous difference Fourier with model phases

$\rho(x) = \frac{1}{V} \sum_h \Delta_{\text{ano},h} e^{i(\varphi_h^c - \pi)^2} e^{-2\pi i(h.x)}$

Peak height at coordinates of anomalously-scattering atoms

$S_{\text{ano}} \equiv \frac{\langle \rho(x_k) \rangle}{\langle \rho^2 \rangle^{1/2}}$

Expected value of signal $S_{\text{ano}}$

$S_{\text{ano}} \sim CC_{\text{ano}} \frac{N_{\text{refl}}^{1/2}}{N_{\text{sites}}^{1/2} \left(\frac{5}{4}\right)^{1/2}}$
Example of anomalous signal $S_{ano}$

Holton Challenge data

bl831.als.lbl.gov/~jamesh/challenge/anom/

Simulated diffraction data from 3dk0 to 1.8 Å
(useful to 2.3 Å)

0% to 100% occupancy of Se in selenomethionine

"Impossible.mtz" has fraction Se of 0.21
21% SeMet incorporation
22% SeMet incorporation

http://bl831.als.lbl.gov/~jamesh/powerpoint/anomalous_challenge.pptx
Example of anomalous signal $S_{ano}$

Holton Challenge data

Anomalous signal

Anomalous signal $S_{ano}$ vs. Fraction Se

impossible.mtz

Difficult

Easy
Finding the anomalous sub-structure with weak anomalous signal

Current approaches

Dual-space methods (*Shelxd, HySS, Crunch2*)

*Difference Fourier (Solve)*

Limitation of these approaches

*Anomalous differences are only approximately proportional to the structure factors for anomalously-scattering atoms*
Finding the anomalous sub-structure with weak anomalous signal

Most powerful source of information about substructure before phases are known is the SAD likelihood function:

*The likelihood of measuring the observed anomalous data given a partial model*
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 a difference Fourier—but much better

Using LLG completion and dual-space completion in HySS

- 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
Using LLG completion 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
Setting up test data on 165 datasets

- `phenix.fetch_pdb 2o7t`
- `phenix.python $PHENIX/phenix/phenix/autosol/sad_data_from_pdb.py 2o7t`
- Splits out each wavelength (peak, edge, remote etc) for MAD and run separately
- Run HySS with dual-space methods or LLG completion
Direct methods vs LLG completion

164 SAD datasets from PDB

HySS Direct Methods

Fraction of sites found vs Anomalous signal
Direct methods vs LLG completion
164 SAD datasets from PDB

HySS LLG Completion

Fraction of sites found vs Anomalous signal
Holton Challenge data
Correct sites found vs anomalous signal $S_{ano}$

Anomalous signal needed to find sites

- HySS-LLG-brute-force
- HySS-LLG
- Shelxd (100000 tries)
- Shelxd (1000 tries)
- Crunch2
- SOLVE
- HySS (direct methods)
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 1.74 Å

Only merged data could be solved

What is the minimum number of crystals that could have been used?
CysZ multi-crystal sulfur-SAD data

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Anomalous signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5.22</td>
</tr>
<tr>
<td>1</td>
<td>5.66</td>
</tr>
<tr>
<td>4</td>
<td>5.81</td>
</tr>
<tr>
<td>2</td>
<td>5.87</td>
</tr>
<tr>
<td>6</td>
<td>6.23</td>
</tr>
<tr>
<td>7</td>
<td>6.63</td>
</tr>
<tr>
<td>3</td>
<td>6.77</td>
</tr>
<tr>
<td>56</td>
<td>7.13</td>
</tr>
<tr>
<td>561</td>
<td>7.94</td>
</tr>
<tr>
<td>67</td>
<td>8.22</td>
</tr>
<tr>
<td>273</td>
<td>9.02</td>
</tr>
<tr>
<td>2734</td>
<td>9.03</td>
</tr>
<tr>
<td>27345</td>
<td>9.07</td>
</tr>
<tr>
<td>27346</td>
<td>9.28</td>
</tr>
<tr>
<td>273456</td>
<td>9.41</td>
</tr>
<tr>
<td>2734561</td>
<td>9.63</td>
</tr>
</tbody>
</table>
CysZ multi-crystal sulfur-SAD data

- **LLG (brute-force)**
- **Shelxd (100000 tries)**

![Graph showing the comparison between LLG and Shelxd for CysZ multi-crystal sulfur-SAD data. The x-axis represents the anomalous signal, and the y-axis represents the number of correct sites. The graph displays two sets of data points, one for LLG and one for Shelxd.](image-url)
CysZ multi-crystal sulfur-SAD data

Correct sites vs. Number of crystals included

- Blue diamonds: LLG (brute-force)
- Red crosses: Shelxd (100000 tries)
CysZ multi-crystal sulfur-SAD data

Merge 6-7

Correct sites vs. Number of crystals included

- LLG (brute-force)
- Shelxd (100000 tries)
CysZ multi-crystal sulfur-SAD data
Merge of crystals 6, 7
AutoSol/Autobuild R/Rfree=0.22/0.26
CysZ multi-crystal sulfur-SAD data

Substructure solution with single datasets

- LLG Single datasets
- Shelxc/d Single datasets

Correct sites vs. Anomalous signal
CysZ multi-crystal sulfur-SAD data
(The minimum number of datasets for this structure is 1)

Data reprocessed by Z. Otwinowski

- LLG single datasets
- Shelxc/d single datasets

Correct sites vs Anomalous signal
Structure determination with weak anomalous signal

AutoSol:
Substructure solution, phasing, density modification, preliminary model-building

AutoBuild
Iterative model-building, refinement, density modification

Parallel AutoBuild
Parallel runs of AutoBuild with map averaging and picking best models
Structure solution with phenix.autosol

- Experimental data, sequence, anomalously-scattering atom, wavelength(s)
- Find heavy-atom sites with direct methods (HYSS dual-space)
- Calculate phases (Phaser)
- Improve phases, find NCS, build model
Structure solution with phenix.autosol: enhancements for weak SAD data

1. Experimental data, sequence, anomalously-scattering atom, wavelength(s)
2. Find heavy-atom sites with direct methods (HYSS LLG completion)
3. Calculate phases (Phaser)
4. Improve phases, find NCS, build model
5. Use map and model in LLG completion
AutoSol structure solution
164 SAD datasets from PDB
(including inflection/remote datasets not previously used as SAD data)
AutoSol structure solution
164 SAD datasets from PDB

Phenix AutoSol/AutoBuild (optimized)
AutoBuild model-building
164 SAD datasets from PDB
Holton Challenge data

Starting point: known sites.

Calculate phases, carry out iterative density modification, model-building and refinement.

Final map correlation vs anomalous signal-to-noise.
Estimating the anomalous signal from the data

Gold standards for the anomalous information:

Correlation of true and observed differences:

Peak height in model-phased Difference Fourier:

Relationship between $CC_{ano}$ and $S_{ano}$
Checking the relationship between $CC_{ano}$ and $S_{ano}$

$S_{ano} \sim CC_{ano} \frac{N^{1/2}_{\text{refl}}}{N^{1/2}_{\text{sites}} \left(\frac{5}{4}\right)^{1/2}}$

$CC_{ano}$: Correlation of anomalous differences with model differences

$S_{ano}$: Peak height in model-phased difference Fourier
Checking the relationship between $CC_{ano}$ and $S_{ano}$

**Equation:**

$$S_{ano} \sim CC_{ano} \frac{N^{1/2}}{N_{sites}^{1/2} \left(\frac{5}{4}\right)^{1/2}}$$

**Explanation:**

$CC_{ano}$: Correlation of anomalous differences with model differences

$S_{ano}$: Peak height in model-phased difference Fourier
Estimating the anomalous correlation $CC_{ano}$ from the data

$CC_{ano}$ estimates based on simple theory:

Estimated from experimental uncertainties and anomalous differences

Estimated from half-dataset correlation of experimental anomalous differences

$$E_{ano}^2 = \frac{\langle \sigma_{ano}^2 \rangle}{\langle \Delta_{ano}^{2,obs} \rangle}$$

$$CC_{ano} \sim [1 - E_{ano}^2]^{1/2}$$

$$CC_{ano}^* = \left[ \frac{2CC_{ano}^{\text{half - dataset}}}{1 + CC_{ano}^{\text{half - dataset}}} \right]^{1/2}$$

$$CC_{ano} \sim CC_{ano}^*$$
Estimating $CC_{\text{ano}}$ from experimental uncertainties and anomalous differences

$$E^2_{\text{ano}} = \frac{<\sigma^2_{\text{ano}}>}{<\Delta^2_{\text{ano,obs}}>},$$

$$CC_{\text{ano}} \sim [1 - E^2_{\text{ano}}]^{1/2}$$

$$(1-<\sigma^2>/<\Delta^2_{\text{ano}}>)^{1/2}$$

$y = 0.8617x$

$R^2 = 0.71607$
Estimating $CC_{ano}$ from the half-dataset anomalous correlation.

\[ CC_{ano} = \left[ \frac{2CC_{half\_dataset}}{1 + CC_{ano}} \right]^{1/2} \]

\[ CC_{ano} \sim CC^*_{ano} \]
Skew of anomalous difference Patterson

Anomalous difference Patterson for 2a3n (14 Se sites, 1.3 Å)
Contours at +/-4σ. Positive pink, negative blue

Model anomalous differences
Skew of anomalous difference Patterson

Anomalous difference Patterson for 2a3n (14 Se sites, 1.3 Å)
Contours at 4σ. Positive blue, negative green

Measured anomalous differences
Skew of anomalous difference Patterson

Anomalous difference Patterson for 2a3n (14 Se sites, 1.3 Å)
Contours at 4σ.

Model (pink) and experimental (blue) anomalous differences
Anomalous difference Patterson for 2a3n (14 Se sites, 1.3 Å)
Contours at 4σ. Positive blue, negative pink.

Randomized anomalous differences
Estimating $CC_{ano}$ from skew of the anomalous difference Patterson

$$y = 1.3329x + 0.1419$$

$R^2 = 0.68298$

$CC_{ano} \sim \text{skew}_{\text{Patterson}}^{1/2}$
Estimating the anomalous correlation $CC_{ano}$

Estimated fraction of observed anomalous differences that is noise

$$E^2_{ano} = \frac{\langle \sigma^2_{ano} \rangle}{\langle \Delta^2_{ano,obs} \rangle}$$

$$CC_{ano} \sim [1 - E^2_{ano}]^{1/2}$$

Half-dataset CC of anomalous differences

$$CC^*_{ano} = \left[ \frac{2CC^{\text{half \_dataset}}_{ano}}{1 + CC^{\text{half \_dataset}}_{ano}} \right]^{1/2}$$

$$CC_{ano} \sim CC^*_{ano}$$

Skew of anomalous difference Patterson

$$CC_{ano} \sim skew^{1/2}_{\text{Patterson}}$$
Estimating the anomalous signal $S_{\text{ano}}$

Estimation of $S_{\text{ano}}$ requires the value of $CC_{\text{ano}}$ and the number of sites

$$S_{\text{ano}} \sim CC_{\text{ano}} \frac{N_{\text{refl}}^{1/2}}{N_{\text{sites}}^{1/2} \left( \frac{5}{4} \right)^{1/2}}$$

Use \textit{phenix.autosol} estimate of number of sites

Based on sequence, asymmetric unit volume

Guess of number of NCS copies

Guess of number of sites for atoms other than S, Se (typically 1-2 per 100 residues)
Estimating $S_{ano}$ from skew of the anomalous difference Patterson

$CC_{ano} \sim skew_{Patterson}^{1/2}$

$S_{ano} \sim CC_{ano} \frac{N_{refl}^{1/2}}{N_{sites}^{1/2} \left(\frac{5}{4}\right)^{1/2}}$

\[ y = 1.4589x + 5.4687 \]

$R^2 = 0.71737$
Estimating $S_{ano}$ from all 3 measures of anomalous correlation

$CC_{ano} \sim \left[ 1 - E_{ano}^2 \right]^{1/2}$

$CC_{ano} \sim \left[ \frac{2CC_{ano}^{\text{half \_dataset}}}{1 + CC_{ano}^{\text{half \_dataset}}} \right]^{1/2}$

$CC_{ano} \sim \text{skew}_{\text{Patterson}}^{1/2}$

$S_{ano} \sim CC_{ano} \frac{N_{\text{refl}}^{1/2}}{N_{\text{sites}}^{1/2} \left( \frac{5}{4} \right)^{1/2}}$

$y = 1.0917x - 0.9473$

$R^2 = 0.7222$
Using the anomalous signal $S_{\text{ano}}$ and correlation $CC_{\text{ano}}$

**What do we expect:**

*Finding sites may be most closely related to map quality ($S_{\text{ano}}$)*

*Experimental phase quality may be most closely related to the accuracy of the anomalous differences ($CC_{\text{ano}}$)*
Can I find the substructure:
Using the anomalous signal $S_{ano}$ to guess

*Best possible case: using known signal $S_{ano}$*
Can I find the substructure:
Using the anomalous correlation $CC_{ano}$

*Best possible case: using true $CC_{ano}$*
Can I find the substructure:
Using the anomalous signal $S_{ano}$ to guess

*Best possible case: using known signal $S_{ano}$*
Can I find the substructure:
Using the anomalous signal $S_{ano}$ to guess

Real-world case: $S_{ano}$ estimated from the data
How good will the phasing be: Could we use the anomalous signal $S_{ano}$?
How good will the phasing be: Using the anomalous correlation $CC_{ano}$ to guess

Real-world case: $CC_{ano}$ estimated from the data
Anomalous signal and anomalous correlation are useful measures of quality and can be estimated from the data.

Likelihood-based methods for finding the anomalous substructure are powerful even with weak signal.

Structures can be solved with weak signal.
The PHENIX Project

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Nat Echols, Ralf Grosse-Kunstleve, Nigel Moriarty, Nicholas Sauter, Peter Zwart

Cambridge University

Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner, Richard Mifsud

Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Duke University

Jane & David Richardson, Jeff Headd, Vincent Chen, Chris Williams, Bryan Arendall, Swati Jain, Bradley Hintze, Lizbeth Videau

An NIH/NIGMS funded Program Project