Structure solution from weak anomalous data

CCP4 Study Weekend
Nottingham, UK
Jan. 8-9, 2015

Tom Terwilliger, Li-Wei Hung (Los Alamos National Laboratory)
Gábor Bunkóczi, Randy Read (Cambridge University)
Nat Echols, Ralf Grosse-Kunstleve, Paul Adams (Lawrence Berkeley National Laboratory)
James Holton (Lawrence Berkeley National Laboratory, UCSF)
David Akey, Janet Smith (University of Michigan)
Structure solution from weak anomalous data

Problems with weak signal

Quantifying the anomalous signal

Solving the anomalous sub-structure and phasing with weak signal

Estimating the anomalous signal in a dataset

Scaling and merging SAD data

Will I solve the anomalous sub-structure?
Problems with weak anomalous signal
Why would I have a weak anomalous signal?

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

Why is this a problem?

Sub-structure identification is difficult

Phasing is poor

Iterative density modification, model-building and refinement works poorly
Quantifying the anomalous signal
Anomalous differences
(Present for acentric reflections only)
\[ \Delta_{ano} \equiv F^+ - F^- \approx -2F_A \sin(\alpha) \]
Anomalous difference Fourier with observed data and model phases

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

Anomalous signal:
Peak height at coordinates of anomalously-scattering atoms

$$S_{\text{ano}} = \frac{\langle \rho_{\text{ano}}(x_j) \rangle}{\langle \rho_{\text{ano}}^2 \rangle^{1/2}}$$

Typical values of $S_{\text{ano}}$ for solved datasets: 10-20
Contributions to measured anomalous differences

\[ \Delta_{\text{obs}}^{\text{ano}} = \Delta_{\text{ano}} + \Delta_{\text{other}}^{\text{ano}} + \varepsilon \]

Measured

Sub-structure anomalous difference

Minor sites, C,N,O atoms...

Measurement errors
How similar are my anomalous differences to model differences?

Correlation of observed and sub-structure anomalous differences

Expected value of $CC_{ano}$

$CC_{ano} \equiv \frac{\langle \Delta_{ano,j} \Delta_{obs}^{ano,j} \rangle}{\sqrt{\langle \Delta_{ano} \rangle^2 / 2 \langle \Delta_{obs}^{ano} \rangle^2 / 2}}$

$\langle CC_{ano} \rangle \geq \frac{rms(\Delta_{ano})}{rms(\Delta_{obs}^{ano})}$

$CC_{ano}$ indicates how much of each anomalous difference is useful (on average)

Anomalous difference Fourier map at 2.5 $\sigma$ (with noise)
How big will my anomalous signal be?
(Based on our simple model for anomalous differences)

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$^{nd}$ moment of the anomalous scattering factor

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

Anomalous scattering factor

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

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
- **215 SAD datasets 1.2 – 4.5 Å**

The graph shows a linear relationship between anomalous signal \( S_{\text{ano}} \) and \( CC_{\text{ano}} \) N\(_{\text{refl}}^{1/2} / (f^{1/2} n_{\text{sites}}^{1/2}) \). The equation of the line is \( y = 1.03x \) with an \( R^2 = 0.91 \).
What affects the anomalous signal?

\[
\langle S_{\text{ano}} \rangle = CC_{\text{ano}} \frac{N_{\text{refl}}^{1/2}}{f^{1/2} n_{\text{sites}}^{1/2}}
\]
Solving the anomalous sub-structure with weak signal
The SAD likelihood function

The likelihood of measuring the observed anomalous data given a partial model

Most powerful source of information about the sub-structure before phases are known
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
Dual Space Sub-structure Completion

Fraction of sites found vs. Anomalous signal
LLG Sub-structure Search

Fraction of sites found vs. Anomalous signal
Anomalous signal indicates if a dataset can be solved
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

CysZ substructure solution
Single and merged datasets

Correct sites

Anomalous signal
CysZ multi-crystal sulfur-SAD data

Single datasets reprocessed by Z. Otwinowski

- Single datasets
CysZ single-crystal sulfur-SAD data
Crystal 6  AutoSol R/Rfree=0.24/0.27
Flavivirus NS1 multi-crystal sulfur-SAD data
Flavivirus NS1 multi-crystal sulfur-SAD data

Phasing with weak signal
Quality of phasing depends on the anomalous correlation ($CC_{ano}$).
Estimating the anomalous signal before and after collecting the data
What affects the anomalous signal?

\[ < S_{ano} > = CC_{ano} \frac{N_{refl}^{1/2}}{f^{1/2} n_{sites}^{1/2}} \]
Anomalous correlation decreases if the data are not accurately measured
Estimating the anomalous signal before collecting the data

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

**Anomalous signal**

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

**Anomalous correlation**

**Number of reflections**

Guess from sequence

**Number of sites**

Choose $d_{min}$, guess B

B-value for anomalous sub-structure
Estimating the anomalous signal before collecting the data

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

Choose $l/sigl$, estimate normalized errors and $CC_{\text{ano}}$

Anomalous correlation

Number of reflections

Guess from sequence

Number of sites

B-value for anomalous sub-structure

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

Predicted anomalous signal

\( y = 0.99x \)

\( R^2 = 0.58 \)
Scaling and merging SAD data
Anomalous correlation estimate from the measured data (deposited data)

\[ y = -0.67x^2 + 1.57x - 0.17 \]
\[ R^2 = 0.55 \]
Anomalous correlation estimate from the measured data
(after local scaling with *phenix.scale_and_merge*)

![Graph](image)

\[ y = 0.23x^2 + 0.72x - 0.05 \]

\[ R^2 = 0.80 \]
Improvement in anomalous correlation using local scaling with `phenix.scale_and_merge`
Estimating the anomalous signal after collecting the data

**Anomalous signal**

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

- Estimate from half-dataset anomalous CC, sigmas, skew
- Anomalous correlation
- Number of reflections
- Estimating the anomalous signal after collecting the data

**Number of sites**

**B-value for sub-structure from Wilson B**

- Guess from sequence
- From data
Estimating the anomalous signal after collecting the data

Anomalous signal $S_{\text{ano}}$

Estimated anomalous signal ($\text{phenix.anomalous\_signal}$)

$y = 1.06x$

$R^2 = 0.78$
Will I solve the anomalous substructure?
Will I solve the anomalous substructure?

*(Planning an experiment)*

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

**Anomalous signal**

**Anomalous correlation**

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

**Number of reflections**

Guess from sequence

**Number of sites**

Choose $d_{min}$, guess B

**B-value for anomalous sub-structure**

Estimated p(sub-structure) before collecting data *phenix.plan_sad_experiment*
Will I solve the anomalous substructure? (After collecting the data)

- Anomalous signal
- Estimate from half-dataset anomalous CC, sigmas, skew
- Anomalous correlation
- Number of reflections
- Number of sites
- B-value for anomalous sub-structure
- Estimate B for sub-structure from Wilson B

Graph: Estimated p(sub-structure) after collecting data phenix.anomalous_signal

Y-axis: Percentage substructure found
X-axis: Estimated p(sub-structure) after collecting data
Take-home message:
The anomalous signal is the key to solving your structure

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

An NIH/NIGMS funded Program Project

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

Los Alamos National Laboratory
Tom Terwilliger, Li-Wei Hung

University of Cambridge
Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner, Richard Mifsud

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

Phenix