Likelihood and SAD phasing in *Phaser*
Concept of likelihood

- Likelihood with dice

Roll a seven. Which die?
- \( p(4) = p(6) = 0 \)
- \( p(8) = 1/8 \)
- \( p(10) = 1/10 \)
Principle of maximum likelihood

- Best model is most consistent with data
- Measure consistency by probabilities
- Optimise model by adjusting parameters in probability distribution
Least squares and likelihood

- Most experiments have multiple sources of error: Gaussian error in observations
  - Central Limit Theorem
- Likelihood for Gaussians = least squares
Least-squares line fitting

\[ p(A_{j,\text{obs}}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(A_{j,\text{obs}} - A_{j,\text{calc}})^2}{2\sigma^2} \right] \]

\[ -\ln \prod_j p(A_{j,\text{obs}}) \propto c + \sum_j \frac{(A_{j,\text{obs}} - A_{j,\text{calc}})^2}{\sigma^2} \]
Why not least squares in crystallography?

- Gaussian error for observations
- Error in predicting observation generally includes difference between structure factors
  - this is Gaussian in *phased* difference
  - *e.g.* $F$ vs. $F_C$ from model, $F_P$ vs. $F_{PH}$
- Phased error usually dominates
  - elimination of unknown phase changes probabilities
Applying likelihood to crystallography

- Find probability distribution for observations
  - start from structure factor probabilities
  - eliminate unknown phase angles
- Adjust parameters to optimise likelihood
- Applications:
  - calculating model phase probabilities
  - structure refinement
  - experimental phasing (isomorphous/anomalous)
  - likelihood-based molecular replacement
The Central Limit Theorem

- Probability distribution of a sum of independent random variables tends to be Gaussian
  - regardless of distributions of variables in sum
- Conditions:
  - sufficient number of independent random variables
  - none may dominate the distribution
- Centroid (mean) of Gaussian is sum of centroids
- Variance of Gaussian is sum of variances
Effect of atomic errors (or differences)

- Atomic errors give “boomerang” distribution of possible atomic contributions
- Portion of atomic contribution is correct
Structure factor with coordinate errors

- Same direction as the sum of the atomic $f$
  - but shorter by $0 < D < 1$
  - $D = f(\text{resolution})$

- Central Limit Theorem
  - Many small atoms
  - Gaussian distribution for the total summed $F$
  - $\sigma_{\Delta} = f(\text{resolution})$
Probability distribution for related structure factors

• Fraction of calculated structure factor correlated to true structure factor: $D$

$$p(F; F_C) = \frac{1}{\pi \varepsilon \sigma_\Delta^2} \exp \left( - \frac{|F - DF_C|^2}{\varepsilon \sigma_\Delta^2} \right)$$

• (Sim, Luzzati, Srinivasan...)

• Takes form of complex normal distribution
Amplitude probability distribution

- Integrate over unknown phase angle to get Rice (Luzzati, Sim, Srinivasan) distribution
SAD: single-wavelength anomalous diffraction

- Most popular way to solve structures by experimental phasing (over 70% and rising)
- Can be done with intrinsic S and CuKα X-rays
- SAD phasing theory is very good
- Easy to automate
- Can be very fast
  - Can be done from single dataset
- May need multiple crystals
  - And careful data processing
Anomalous scattering

Anomalous adj.
Deviating from the normal or common order, form or rule

“Anomalous scattering” is absolutely normal while “normal scattering” occurs only as an ideal, over simplified model, which can be used as a first approximation when studying scattering problems”

IUCR Pamphlet “Anomalous Dispersion of X-rays in Crystallography”
S. Caticha-Ellis (1998)
Anomalous scattering

- Anomalous scattering is due to the electrons being tightly bound (particularly in K & L shells)
- In classical terms, the electrons scatter as though they have resonant frequencies
Driven Mechanical Oscillator

MIT Physics Lecture Demonstration Group

https://www.youtube.com/watch?v=aZNnwQ8HJHU
Diffraction with anomalous scatterers

- SAD: single-wavelength anomalous diffraction
Diffraction with anomalous scatterers

- SAD: single-wavelength anomalous diffraction
Harker construction for SAD phasing
Collections of structure factors

- Distribution for one is complex normal
  - 2D Gaussian centered on expected value
  - one variance describes size of uncertainty
- Distribution for several is multivariate complex normal
  - single variance is replaced by covariance matrix
  - covariance matrix describes uncertainties and correlations among structure factors
- SAD: 2 observed structure factors, and 2 calculated structure factors
SAD likelihood function

- Factor joint probability into two parts
  \[ p(F^+_o, F^-_o; H^+, H^-) = p(F^+_o; F^-_o, H^+, H^-) p(F^-_o; H^-) \]
- Integrate out unknown phases, \( \alpha^+ \) and \( \alpha^- \)
Intuitive understanding of SAD phasing

Expected value of $F^*$ ($H^*$)

$$P(F_0^-, \alpha_0^-; H^-)$$

Expected difference between $F^*$ and $F^*$

$$P(F_0^+; F_0^-, H^+, H^-)$$
Intuitive understanding of SAD phasing

Total likelihood is integral of the product of the two distributions under the black circle

Expected value of $F^{-*} (H^{-*})$

Expected difference between $F^+$ and $F^{-*}$
SAD log-likelihood gradient (LLG) map

- Compute derivative of log-likelihood with respect to heavy atom structure factor
- Fourier transform gives map of where likelihood target would like to see changes in anomalous scatterer model
- Very sensitive to minor sites
  - picks up sites identified as water molecules in refined structures determined by halide soaks
  - tutorial with data for lysozyme iodide soak
Locating anomalous scatterers in model solved by MR

- Structure of thyroxine-binding globulin
  - thyroxine doesn’t bind where most people expected
- Thyroxine contains 4 iodine atoms
  - $f'' \approx 3e$ for $\lambda = 0.979\text{Å}$
- Compare conventional model-phased anomalous difference map with *Phaser* LLG map
$\Delta_{\text{anor}} 3.5\sigma$

mol 1

mol 2

LLG, $5.5\sigma$
Combining MR and SAD

- CuKα data to 1.9Å on hen egg-white lysozyme
  - can’t find sulfurs with HySS or SHELXD
- Solve by MR with goat alpha-lactalbumin (40% identical)
- Use MR model as “substructure” for SAD
  - look for S atoms in LLG map (finds all 10 S, 5-9 Cl⁻)
  - phases automatically combine MR and SAD
- Automated fitting with density-modified map
  - tutorial with these data is available
Breakdown of Friedel’s law

- Friedel’s law breaks down for mixture of scatterers differing in real:anomalous ratio
Nitrate reductase (Natalie Strynadka)

- Integral membrane protein, 1976 residues
  - contains 21 Fe atoms, 1 Mo, 118 S, 5 P (146 total)
  - solved using combination of Fe-MAD, MIRAS
SAD phasing of nitrate reductase

- Fe peak SAD data only
  - find 11 “Fe” sites with phenix.hyss
    - several are super-sites of Fe$_4$S$_4$ clusters
  - phase and complete adding Fe, Mo, S with Phaser
    - total of 57 sites: 20 Fe, 6 Mo, 31 S
    - superatoms are resolved, 51 of 57 are identified correctly
      - identity based initially on height of peak in LLG map, revised based on refined occupancy
      - correct hand indicated by number of sites, LLG score
Iterative model-building and phasing

- Improve phases by density modification
- Build with ARP/wARP (or Resolve)
  - 1607 residues, 1368 docked in sequence
- LLG completion from ARP/wARP model
  - 105 sites, 92 correctly identified
- Repeat DM and ARP/wARP
  - 1813 residues, 1775 docked in sequence
Automation of SAD phasing

- Functions are all available from Python
  - used for SAD in PHENIX AutoSol GUI
  - used as optional substructure completion method in phenix.hyss

- Log-likelihood-gradient completion
  - look for one or several types of scatterer
    - start from MR model (atoms or density) or partial substructure
  - analyse map to add sites, make atoms anisotropic
  - delete atoms that fade away
  - repeat to convergence
Practical aspects of SAD phasing in *Phaser*

- Provide information about cell content
  - sequence, molecular weight, percent solvent...
  - used to put data on absolute scale
    - occupancies are reasonably accurate
- Provide information about $f''$ values
  - wavelength (table lookup) or measured
  - refined by default if near edge
- Try both hands if uncertain
  - separate completion if mixture of atom types
SAD phasing in Phenix

- **AutoSol GUI**
  - finds sites with HySS
    - sensitivity of HySS improved by using *Phaser* to complete and score substructures
  - automatically uses *Phaser* for phasing if SAD data
  - tests both hands, chooses best hand
  - carries out Resolve density modification and model-building
SAD phasing in CCP4

- ccp4i GUI
  - modes for SAD phasing or MR+SAD
  - SAD phasing pipeline
    - find substructure with Hyss or SHELXD
    - phase both hands
    - density modification with parrot
    - quick model-building with buccaneer
- MR+SAD
  - provide MR model
  - only phase in hand of MR solution
Background information

  - plus papers cited here

- [http://www-structmed.cimr.cam.ac.uk/Course](http://www-structmed.cimr.cam.ac.uk/Course)
Contributors

- Experimental phasing
  - Airlie McCoy, Laurent Storoni

- PHENIX collaboration
  - Ralf Grosse-Kunstleve, Nigel Moriarty, Paul Adams
  - Tom Terwilliger

- CCP4 SAD pipeline
  - Kevin Cowtan