Structure factor statistics and likelihood

Randy J Read

UNIVERSITY OF CAMBRIDGE

Cambridge Institute for Medical Research

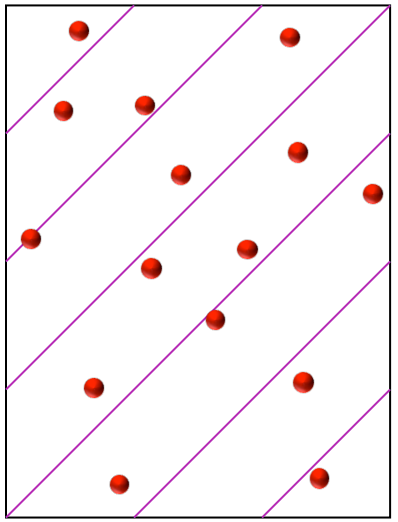
Principle of maximum likelihood

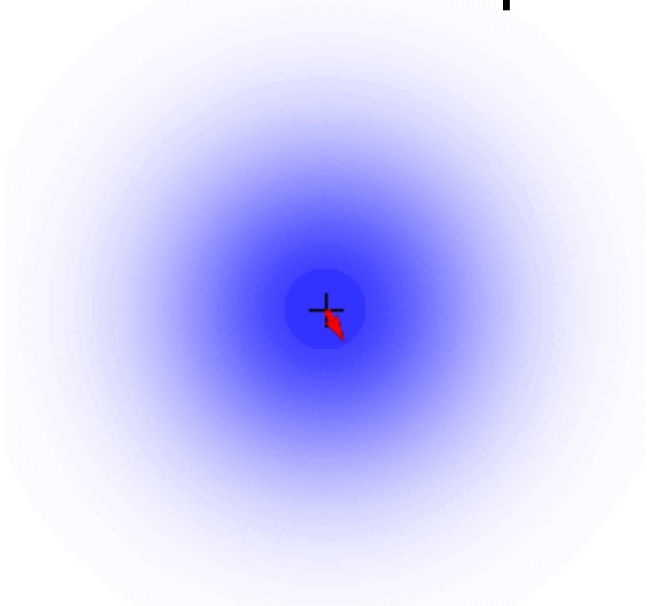
- Best model is most consistent with data
- Measure consistency by probabilities
- Optimise model by adjusting parameters in probability distribution

- Crystallographic likelihood is based on probability distributions of structure factors
  - univariate for maps, molecular replacement, refinement
  - multivariate for experimental phasing and other advanced applications

Structure factors with random atoms

- Assume atoms randomly scattered relative to Bragg planes
- Random walk in complex plane





Wilson distribution

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

Wilson distribution for space group P1

- Apply central limit theorem to real and imaginary parts of structure factor separately
  - sums of real and imaginary atomic contributions

Derivation of Wilson distribution: centroids

$$\mathbf{F} = \sum_{j=1}^{Natom} f_j \exp\left(2\pi i \mathbf{h}\cdot\mathbf{x}_j\right) = \sum_{j=1}^{Natom} f_j \cos\left(2\pi \mathbf{h}\cdot\mathbf{x}_j\right) + i\sum_{j=1}^{Natom} f_j \sin\left(2\pi \mathbf{h}\cdot\mathbf{x}_j\right)$$

$$= A + iB$$

$$\langle \mathbf{F} \rangle = \langle A + iB \rangle = \langle A \rangle + i\langle B \rangle$$

$$\langle A \rangle = \sum_{j=1}^{Natom} \left\langle f_j \cos\left(2\pi \mathbf{h}\cdot\mathbf{x}_j\right) \right\rangle$$

$$= 0, \text{ (assume random position relative to Bragg planes)}$$

$$\langle B \rangle = \sum_{j=1}^{Natom} \left\langle f_j \sin\left(2\pi h\cdot x_j\right) \right\rangle = 0$$

$$\langle \mathbf{F} \rangle = \mathbf{0}$$

Derivation of Wilson distribution: variances

$$\left\langle \left|\mathbf{F}-\mathbf{0}\right|^2 \right\rangle = \left\langle A^2 + B^2 \right\rangle = \left\langle A^2 \right\rangle + \left\langle B^2 \right\rangle$$

$$\left\langle A^2 \right\rangle = \sum_{j=1}^{Natom} \left\langle \left(f_j \cos\left(2\pi h \cdot x_j\right)\right)^2 \right\rangle, \text{(assume atoms uncorrelated)}$$

$$= \frac{1}{2} \sum_{j=1}^{Natom} f_j^2, \text{(assume random position relative to Bragg planes)}$$

$$= \Sigma_N / 2$$

$$\left\langle B^2 \right\rangle = \sum_{j=1}^{Natom} \left\langle \left(f_j \sin\left(2\pi h \cdot x_j\right)\right)^2 \right\rangle = \Sigma_N / 2$$

$$\left\langle \left|F\right|^2 \right\rangle = \Sigma_N$$

Derivation of Wilson distribution: joint distribution of *A* and *B*

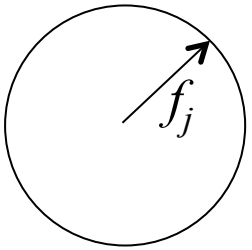
$$p(A) = \frac{1}{\sqrt{\pi \Sigma_N}} \exp\left(-\frac{A^2}{\Sigma_N}\right), \text{(Gaussian with mean of 0, variance of } \Sigma_N / 2)$$

$$p(B) = \frac{1}{\sqrt{\pi \Sigma_N}} \exp\left(-\frac{B^2}{\Sigma_N}\right)$$

$$p(\mathbf{F}) = p(A,B) = \frac{1}{\pi \Sigma_N} \exp\left(-\frac{A^2 + B^2}{\Sigma_N}\right) = \frac{1}{\pi \Sigma_N} \exp\left(-\frac{|\mathbf{F}|^2}{\Sigma_N}\right)$$

Alternative derivation of Wilson distribution

- Complex normal distribution
  - Gaussian for complex numbers
  - joint distribution of real and imaginary parts
  - Central Limit Theorem also applies

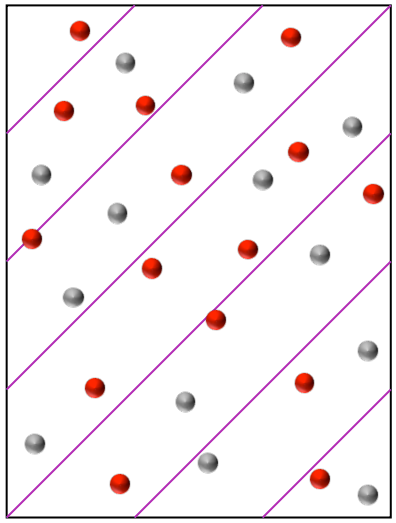


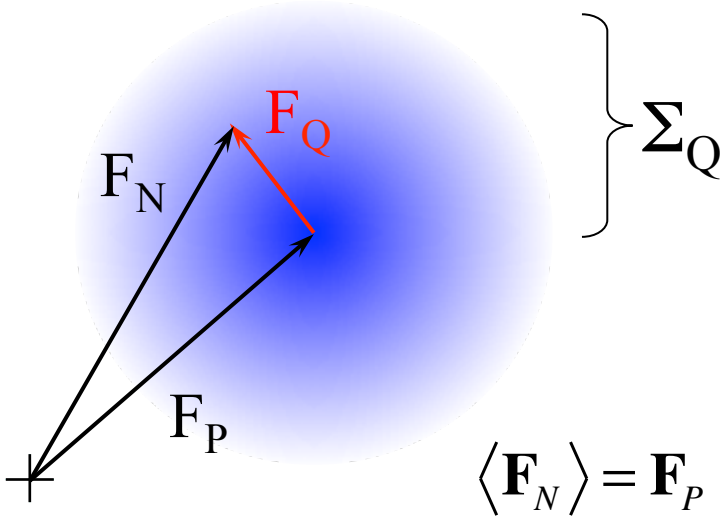
$$\langle \mathbf{F} \rangle = \sum_{j=1}^{Natom} \left\langle f_j \exp\left(2\pi i \mathbf{h}\cdot\mathbf{x}_j\right)\right\rangle = \mathbf{0}$$

$$\left\langle |\mathbf{F}|^2 \right\rangle = \left\langle \mathbf{F}\mathbf{F}^* \right\rangle = \sum_{j=1}^{Natom} \left\langle f_j \exp\left(2\pi i \mathbf{h}\cdot\mathbf{x}_j\right) f_j \exp\left(-2\pi i \mathbf{h}\cdot\mathbf{x}_j\right)\right\rangle = \sum_{j=1}^{Natom} f_j^2 = \Sigma_N$$

$$p(\mathbf{F}) = \frac{1}{\pi \Sigma_N} \exp\left(-\frac{|\mathbf{F}|^2}{\Sigma_N}\right), \text{(complex Gaussian with mean of 0, variance of } \Sigma_N)$$

Sim distribution: known and unknown atoms





$$\langle \mathbf{F}_N \rangle = \mathbf{F}_P$$

$$\left\langle \left(|\mathbf{F}_N| - \langle \mathbf{F}_N \rangle \right)^2 \right\rangle = \sum_{Q \text{ atoms}} f_j^2 = \Sigma_Q$$

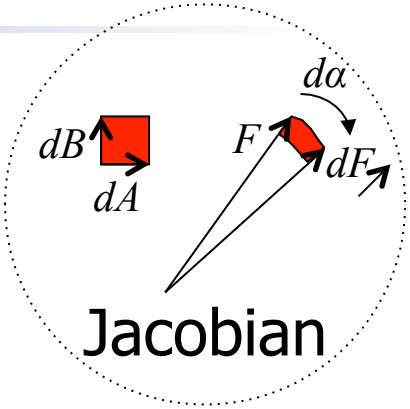
$$p(\mathbf{F}) = \frac{1}{\pi \Sigma_Q} \exp\left(-\frac{|\mathbf{F}_N - \mathbf{F}_P|^2}{\Sigma_Q} \right)$$

Sim distribution for amplitudes

- Likelihood function is the probability of the observations
  - but only the intensity (or amplitude) is measured
  - phase component has to be eliminated
- Change variables from real and imaginary to amplitude and phase
- Integrate over all possible values of (unknown) phase

Sim distribution for amplitudes

$$p(\mathbf{F}_N) = \frac{1}{\pi \Sigma_Q} \exp\left(-\frac{|\mathbf{F}_N - \mathbf{F}_P|^2}{\Sigma_Q} \right), \; dA\, dB = F\, dF\, d\alpha$$



$$p(F_N, \alpha_N) = \frac{F_N}{\pi \Sigma_Q} \exp\left(-\frac{F_N^2 + F_P^2 - 2F_N F_P \cos(\alpha_N - \alpha_P)}{\Sigma_Q} \right)$$

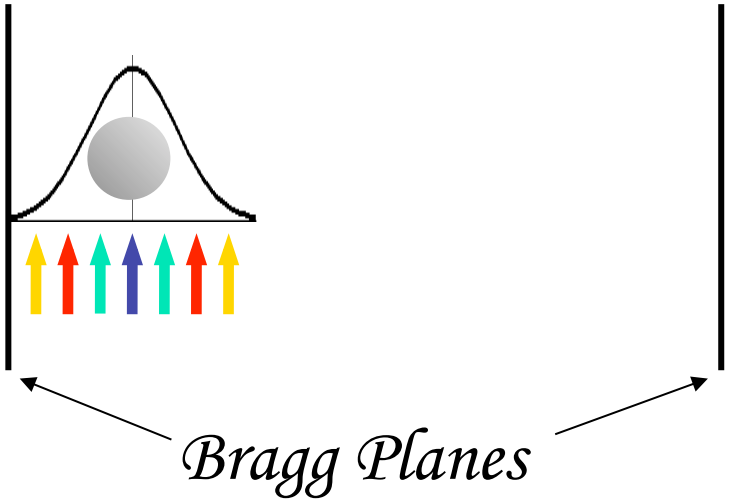
$$p(F_N) = \int_0^{2\pi} p(F_N, \alpha_N)\, d\alpha$$

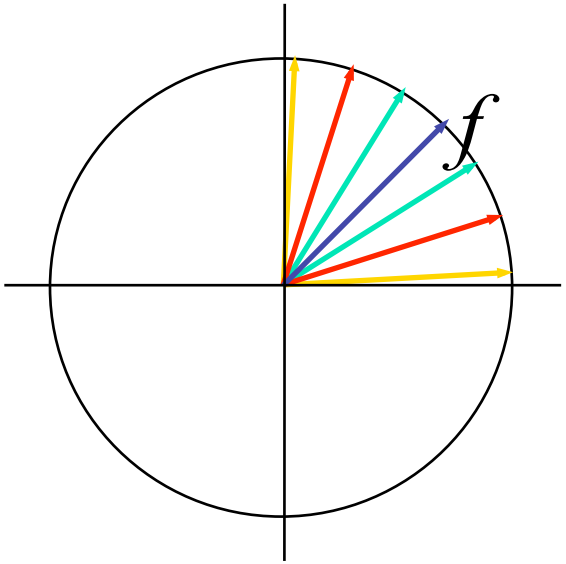
$$= \frac{F_N}{\pi \Sigma_Q} \exp\left(-\frac{F_N^2 + F_P^2}{\Sigma_Q} \right) \int_0^{2\pi} \exp\left(\frac{2F_N F_P \cos(\alpha_N - \alpha_P)}{\Sigma_Q} \right) d\alpha$$

$$= \frac{2F_N}{\Sigma_Q} \exp\left(-\frac{F_N^2 + F_P^2}{\Sigma_Q} \right) \mathrm{I}_0\left(\frac{2F_N F_P}{\Sigma_Q} \right)$$

Effect of atomic errors (or differences)

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





Effect of atomic errors (or differences)

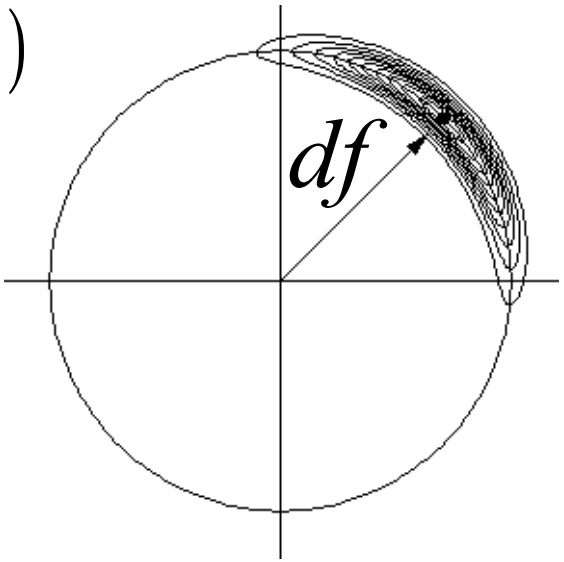
- Work out centroid and variance for the contribution from a single atom

$$\left\langle f_j \exp\left(2\pi i \mathbf{h}\cdot\left(\mathbf{x}_j+\delta_j\right)\right)\right\rangle = \left\langle \exp\left(2\pi i \mathbf{h}\cdot\delta_j\right)\right\rangle f_j \exp\left(2\pi i \mathbf{h}\cdot\mathbf{x}_j\right)$$

$$= d_j f_j \exp\left(2\pi i h\cdot x_j\right)$$

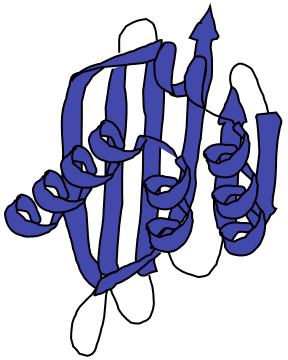
$$= d_j \mathbf{f}_j$$

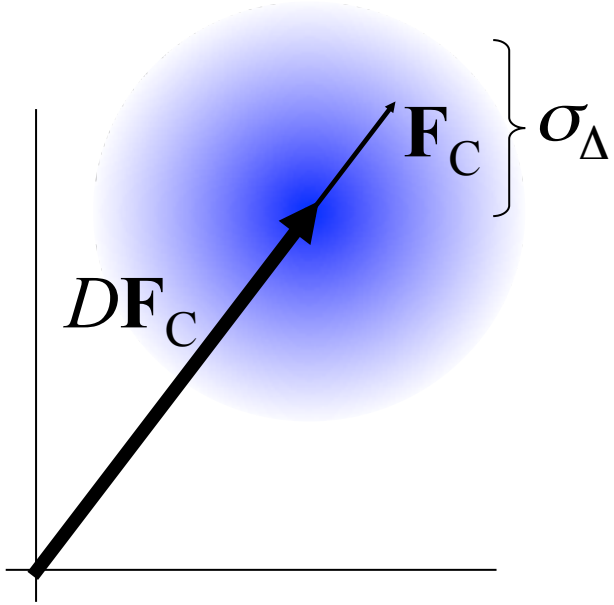
$$\left\langle \left|\mathbf{f}_j \exp\left(2\pi i \mathbf{h}\cdot\delta_j\right) - d_j \mathbf{f}_j\right|^2 \right\rangle = f_j^2\left(1-d_j^2\right)$$

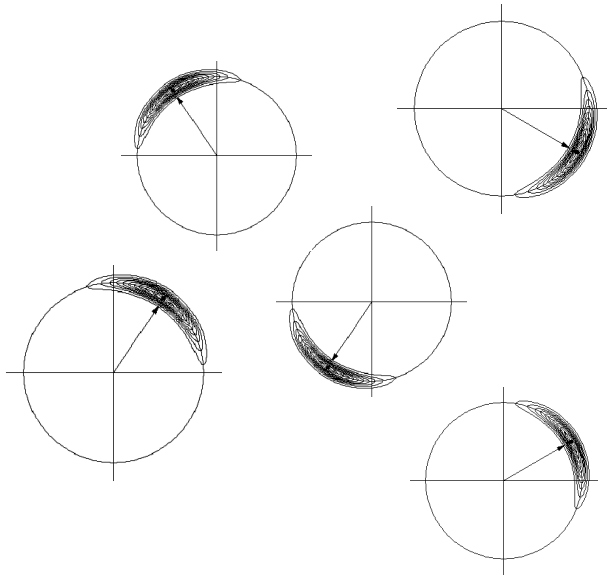


Structure factor with coordinate errors

- Luzzati (1952)
  - all atoms subject to same errors
- Complex normal distribution

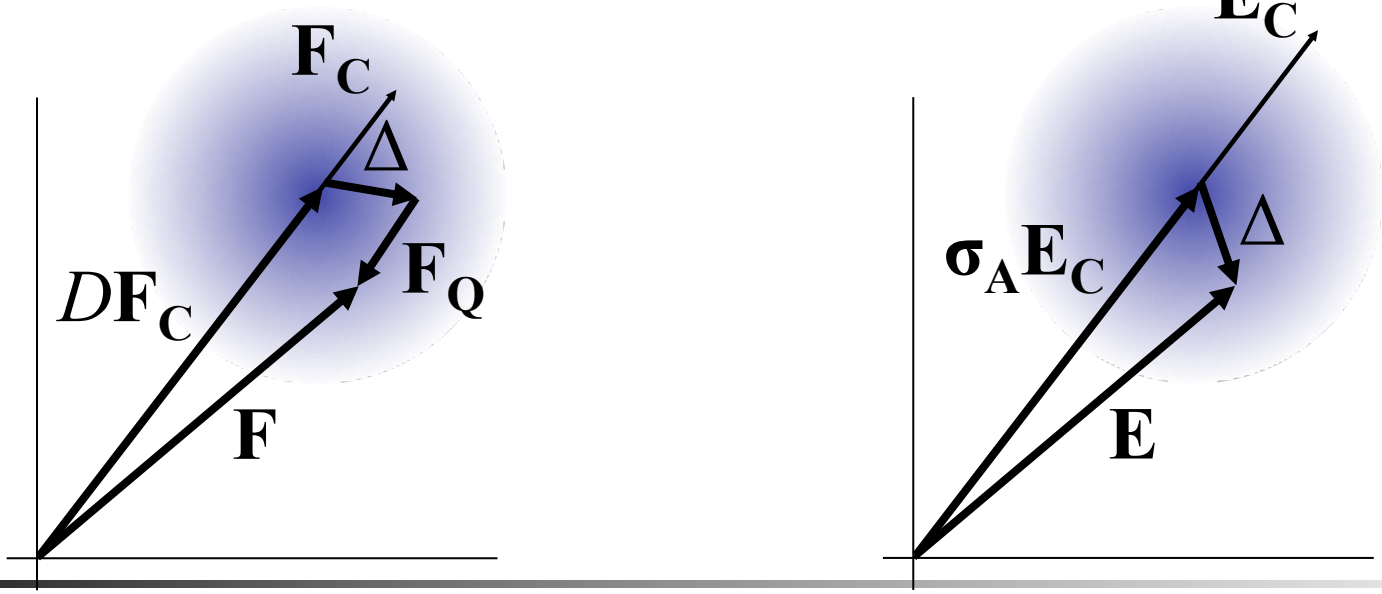






Srinivasan and colleagues: σ_A

- Effects of errors and incompleteness are equivalent in terms of E-values
  - variance of error in E-value is $1-\sigma_A^2$
  - conservation of scattering power



Advanced applications of likelihood

- Refinement and MR involve pairs of structure factors with one observation: $\mathbf{F}_O$ and $\mathbf{F}_C$
- Other applications involve larger collections
  - MIR: $\mathbf{F}_P$, $\mathbf{F}_{PH1}$, $\mathbf{H}_1$, $\mathbf{F}_{PH2}$, $\mathbf{H}_2$, ...
  - SAD: $\mathbf{F}^+$, $\mathbf{F}^-$, $\mathbf{H}^+$, $\mathbf{H}^-$
- Need joint distributions of collections of structure factors

Likelihood for more structure factors

- σ_A can be interpreted as complex covariance between normalised structure factors
- New applications based on multivariate complex normal distribution
  - difficulty is integrating out more than one phase!
  - can always isolate one phase by factoring probability distribution
    - see paper on SAD likelihood target

The normal (Gaussian) distribution

- Gaussian distribution for one variable

$$\mathrm{p}(x) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left(-\frac{\left(x - \langle x \rangle\right)^2}{2\sigma^2} \right)$$

$$= \frac{1}{\left(2\pi\Sigma\right)^{1/2}} \exp\left[-\tfrac{1}{2}\left(x - \langle x \rangle\right)\Sigma^{-1}\left(x - \langle x \rangle\right) \right]$$

- Multivariate normal distribution

$$\mathrm{p}(\mathbf{x}) = \frac{1}{\left|2\pi\Sigma\right|^{1/2}} \exp\left[-\tfrac{1}{2}\left(\mathbf{x} - \langle \mathbf{x} \rangle\right)^T \Sigma^{-1}\left(\mathbf{x} - \langle \mathbf{x} \rangle\right) \right] \text{, where}$$

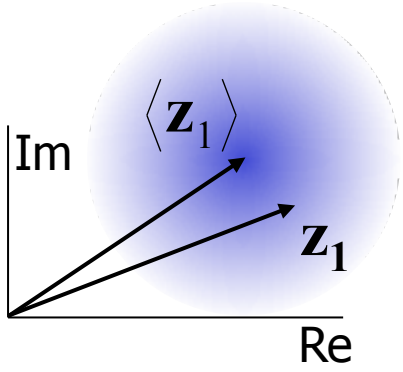
$$\text{elements of } \Sigma \text{ given by } \sigma_{ij} = \left\langle \left(x_i - \langle x_i \rangle\right)\left(x_j - \langle x_j \rangle\right) \right\rangle$$

Multivariate complex normal distribution

- Complex normal

$$p(\mathbf{z}_1) = \frac{1}{\pi\Sigma}\exp\left[-\frac{|\mathbf{z}_1 - \langle \mathbf{z}_1 \rangle|^2}{\Sigma}\right]$$

$$= \frac{1}{\pi\Sigma}\exp\left[-(\mathbf{z}_1 - \langle \mathbf{z}_1 \rangle)^* \Sigma^{-1} (\mathbf{z}_1 - \langle \mathbf{z}_1 \rangle)\right]$$



- Multivariate complex normal distribution
  - Hermitian covariance matrix

$$p(\mathbf{z}) = \frac{1}{|\pi\boldsymbol{\Sigma}|}\exp\left[-(\mathbf{z} - \langle \mathbf{z} \rangle)^H \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \langle \mathbf{z} \rangle)\right] \text{ , where}$$

$$\text{elements of } \boldsymbol{\Sigma} \text{ given by } \boldsymbol{\sigma}_{ij} = \left\langle (\mathbf{z}_i - \langle \mathbf{z}_i \rangle)(\mathbf{z}_j - \langle \mathbf{z}_j \rangle)^* \right\rangle$$

Deriving conditional Gaussian probability

- Start with large joint distribution

  $p\left(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots y_n\right)$

- Fix known or model (y) terms
  - partition covariance matrix

$$\begin{bmatrix} \text{data} - \text{data} & \text{data} - \text{model} \\ \text{data} - \text{model} & \text{model} - \text{model} \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

  - update covariances and expected values for remaining variables

$$\Sigma_{11}^{'} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$\begin{bmatrix} \langle x_1 \rangle \\ \vdots \\ \langle x_m \rangle \end{bmatrix} = \Sigma_{12}\Sigma_{22}^{-1} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Re-deriving Srinivasan distribution

- Start from joint distribution of $\mathbf{E}_O$ and $\mathbf{E}_C$
- Fix $\mathbf{E}_C$, manipulate covariance matrix
  - turn joint distribution into conditional

$$\begin{bmatrix} \langle \mathbf{E}_O \mathbf{E}_O^* \rangle & \langle \mathbf{E}_O \mathbf{E}_C^* \rangle \\ \langle \mathbf{E}_O \mathbf{E}_C^* \rangle^* & \langle \mathbf{E}_C \mathbf{E}_C^* \rangle \end{bmatrix} = \begin{bmatrix} 1 & \sigma_A \\ \sigma_A & 1 \end{bmatrix} \longrightarrow \begin{array}{l} \langle \mathbf{E}_O \rangle_{\mathbf{E}_C} = \sigma_A \mathbf{E}_C \\ \mathrm{var}\left(\langle \mathbf{E}_O \rangle_{\mathbf{E}_C} \right) = 1 - \sigma_A^2 \end{array}$$

Molecular replacement with an ensemble

- Start from joint distribution of structure factors, including true and all models
- Conditional distribution turns collection of models into a statistically-weighted ensemble-average structure factor

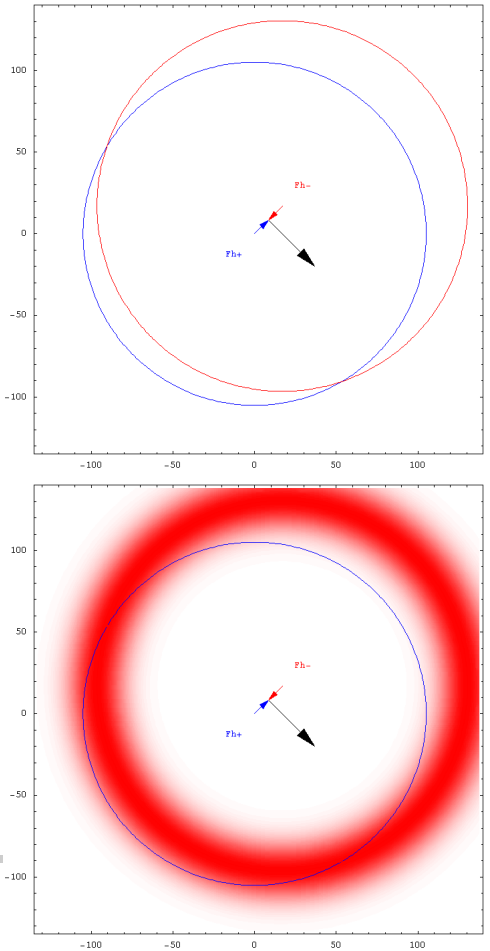
$$\boldsymbol{\Sigma} = \left[\begin{array}{c|ccc} \Sigma_N & D_1\Sigma_{P_1} & \cdots & D_n\Sigma_{P_n} \\ \hline D_1\Sigma_{P_1} & \Sigma_{P_1} & \cdots & \left\langle \mathbf{F}_{C_1}\mathbf{F}_{C_n}^{*} \right\rangle \\ \vdots & \vdots & \ddots & \vdots \\ D_n\Sigma_{P_n} & \left\langle \mathbf{F}_{C_1}^{*}\mathbf{F}_{C_n} \right\rangle & \cdots & \Sigma_{P_n} \end{array}\right]$$

SAD: probabilities for Friedel pair

- Start from joint distribution of true and calculated structure factors
- Use standard manipulations to get conditional probability:

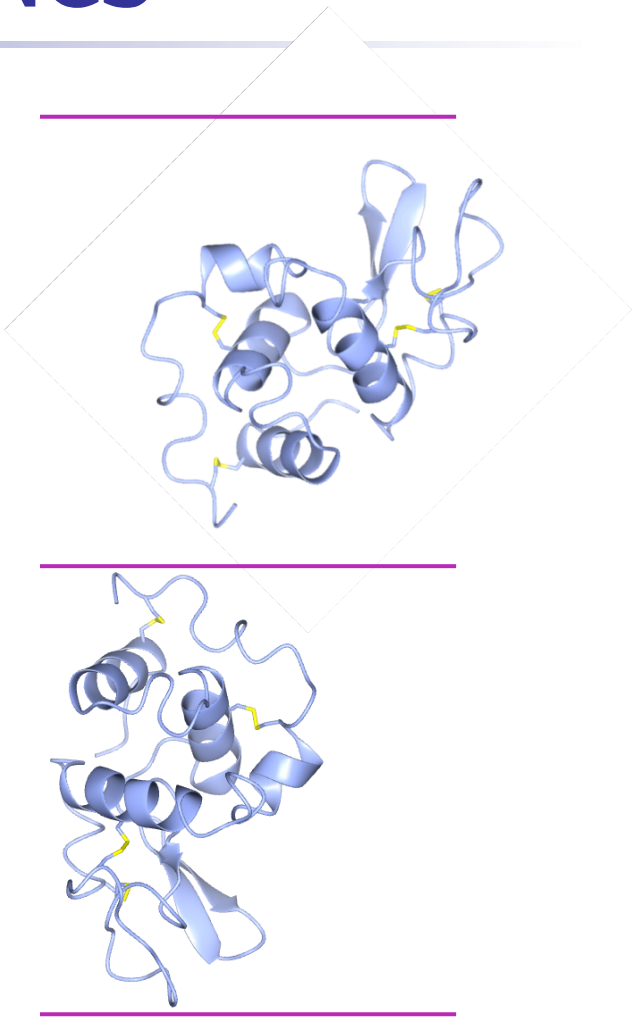
$$p\left(\mathbf{F}_O^+,\mathbf{F}_O^-,\mathbf{H}^+,\mathbf{H}^-\right)\rightarrow p\left(\mathbf{F}_O^+,\mathbf{F}_O^-;\mathbf{H}^+,\mathbf{H}^-\right)$$

$$\left[\begin{array}{cc|cc} \Sigma_N & \left\langle \mathbf{F}_O^+\mathbf{F}_O^- \right\rangle & \left\langle \mathbf{F}_O^+\mathbf{H}^{+*} \right\rangle & \left\langle \mathbf{F}_O^+\mathbf{H}^- \right\rangle \\ \left\langle \mathbf{F}_O^+\mathbf{F}_O^- \right\rangle^* & \Sigma_N & \left\langle \mathbf{F}_O^-\mathbf{H}^+ \right\rangle^* & \left\langle \mathbf{F}_O^{-*}\mathbf{H}^- \right\rangle \\ \hline \left\langle \mathbf{F}_O^{+*}\mathbf{H}^+ \right\rangle & \left\langle \mathbf{F}_O^-\mathbf{H}^+ \right\rangle & \Sigma_H & \left\langle \mathbf{H}^+\mathbf{H}^- \right\rangle \\ \left\langle \mathbf{F}_O^+\mathbf{H}^- \right\rangle^* & \left\langle \mathbf{F}_O^-\mathbf{H}^{-*} \right\rangle & \left\langle \mathbf{H}^+\mathbf{H}^- \right\rangle^* & \Sigma_H \end{array}\right]$$



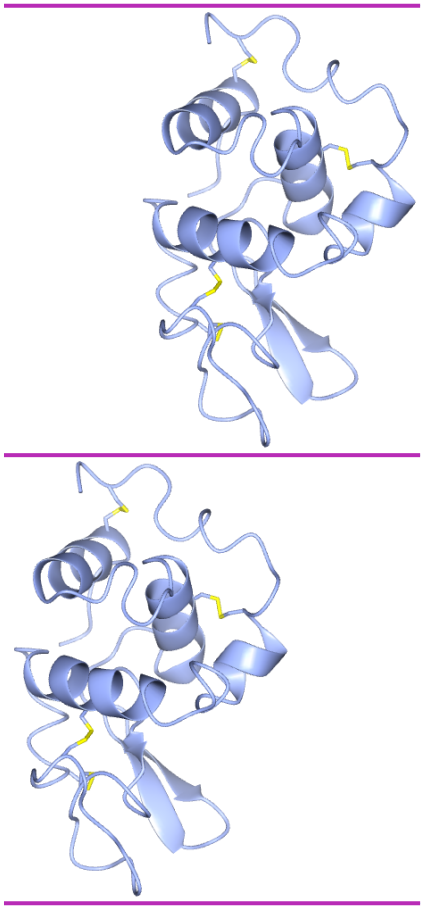
Dealing with translational NCS

- Diffraction from copies in different orientations is uncorrelated
  - zero covariances



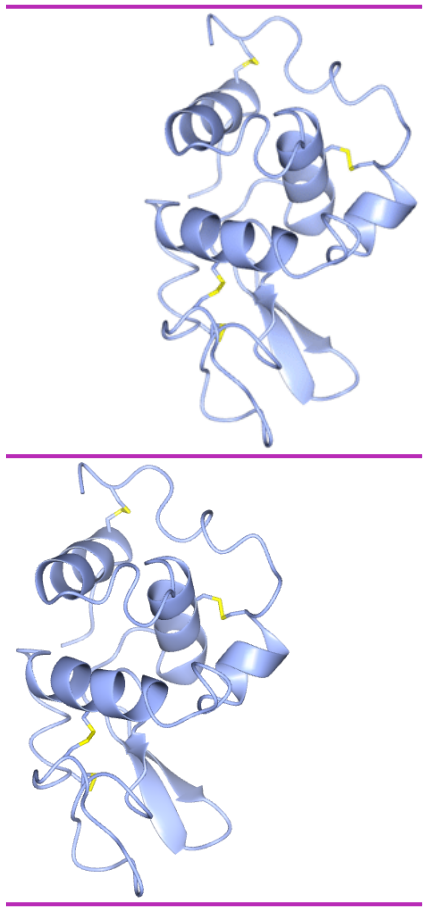
Dealing with translational NCS

- Diffraction from copies in different orientations is uncorrelated
  - zero covariances
- Diffraction from copies in the same orientation is correlated
  - covariances are modulated
- Add covariances to get expected intensity



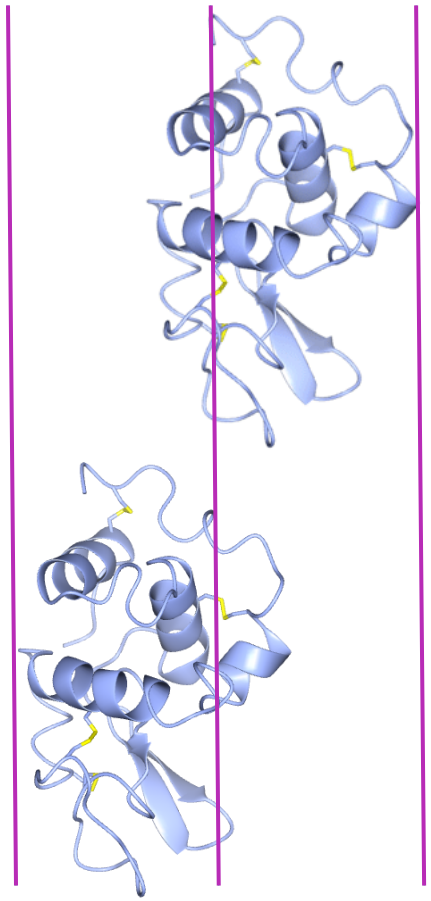
Effect of rotation on translational NCS

- Rotation parallel to diffraction vector has no effect



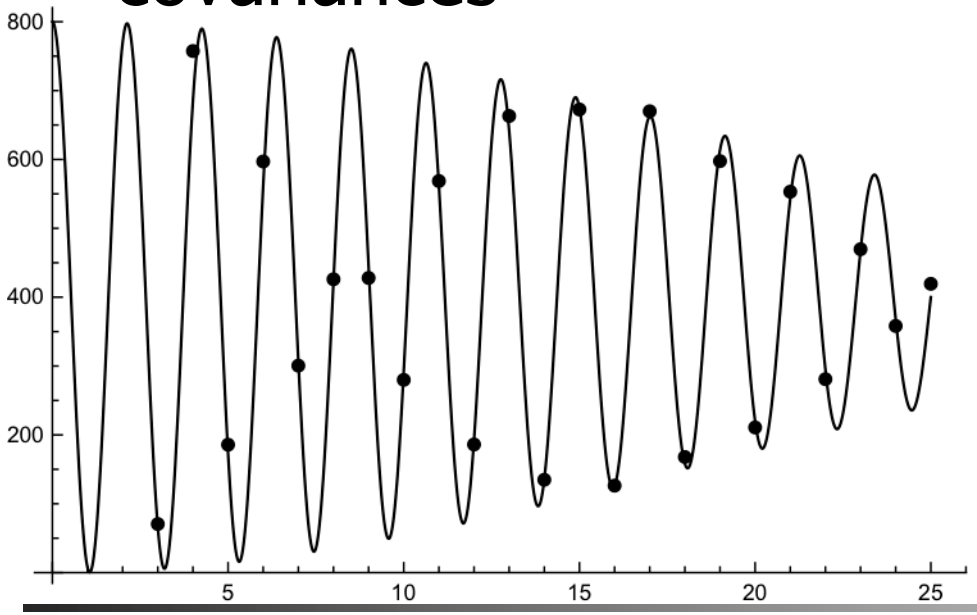
Effect of rotation on translational NCS

- Rotation parallel to diffraction vector has no effect
- Rotation around other axes reduces correlation
- Random coordinate differences between copies reduce correlation



Accounting for translational NCS

- Model effect of translation combined with small rotation and random differences between copies
- Also model *vs.* data covariances



Hyp-1:
Sliwiak, Jaskolski, Dauter, McCoy, Read
(unpublished)

Other potential applications

- SIR likelihood target
  - SIR phasing
  - joint refinement of native and liganded structures
- Fast translation function for SAD target
- Understanding of solvent flattening

Acknowledgements

- Phaser: Airlie McCoy, Laurent Storoni, Gábor Bunkóczi, Rob Oeffner
- Hyp-1: Mariusz Jaskolski, Joanna Sliwiak, Zbyszek Dauter