Derivation of scattering intensities for macromolecular crystals

March 5, 2017

Contents

1 General expressions for Bragg and diffuse intensities

Consider a crystal made of $\mathcal{C}$ unit cells, each containing N atoms (encompassing one or many asymetric units). The total electron density of the crystal can be written as the sum over the electronic density of each atom, defined by the position of their nucleus R and the distribution function of their electrons g:

$$\begin{aligned} \rho_{total}(\vec{r}) &= \sum_{c}^{\mathcal{C}}\sum_{i}^{N}\rho_{ci}(\vec{r}) \\ \text{where } \rho_{ci}(\vec{r}) &= \int d\vec{r'}\ \delta(\vec{r'}-\vec{R}_{ci})g_i(\vec{r}-\vec{r'}) \end{aligned} \tag{1}$$

The wave $A(\vec{q})$ elastically scattered by the crystal can now be written. Noting f the Fourier transform of g, also known as the atomic form factor:

$$A(\vec{q}) = \sum_{c}\sum_{i} f_i(q)e^{-i\vec{q}.\vec{R}_{ci}} \tag{2}$$

The position of each nucleus can be decomposed over the vector from the origin of the coordinate system to the origin of unit cell c $\vec{u_c}$, the vector that describes the average position of atom i relative to origin of the unit cell $\vec{r_i}$ and $\vec{\delta_{ci}}$ the deviation to it. An argument needs to be made here about the meaning of "instantaneous" deviation in Moore's terminology (see [1]). More importantly, in the context of various experimental setups (synchrotron vs XFEL...), what is the average being made ? In the following, we assume that an ensemble needs to be performed, and we might be able to get there by emplying some kind of 'non-distinguishable cells' argument...

1.1 Statistical model

The total intensity scattered at $\vec{q}$ in reciprocal space can now be written as an ensemble average over all possible realizations of the deviations:

$$I(\vec{q}) = \langle|A(\vec{q})|^2\rangle = \sum_{c,c'} e^{-i\vec{q}\cdot(\vec{u_c}-\vec{u_c'})}\sum_{i,j} f_i f_j e^{-i\vec{q}(\vec{r_i}-\vec{r_j})}\langle e^{-i\vec{q}\cdot(\vec{\delta_{ci}}(t)-\vec{\delta_{c'j}}(t))}\rangle \tag{3}$$

Noting $\mathcal{P}$ the underlying probability distribution, the ensemble average over the $3N\mathcal{C}$-dimensional vector δ, of a function f of δ, is defined by

$$\begin{aligned} \langle f(\delta)\rangle &= Z^{-1}\int d\delta\ \mathcal{P}(\delta)f(\delta) \\ \text{where } Z &= \int d\delta\ \mathcal{P}(\delta) \end{aligned} \tag{4}$$

We now make the approximation that $\mathcal{P}$ is a multivariate Gaussian distribution, that is there is a $3N\mathcal{C}^2$ symmetric positive-definite matrix V^{-1} such that $\mathcal{P}(\delta) = e^{-\frac{1}{2}\delta^T V^{-1}\delta}$. At that stage, we shall recall an important property of Gaussian integrals involving polynomials (f would be a polynomial here)

$$\begin{aligned} \int d\delta\ e^{-\frac{1}{2}\delta^T V^{-1}\delta} f(\delta) &= Z\ e^{\frac{1}{2}\nabla^T V\nabla} f(\delta)|_{\delta=0} \\ &\simeq Z(1+\frac{1}{2}\nabla^T V\nabla + ...)f(\delta)|_{\delta=0} \end{aligned} \tag{5}$$

We now evaluate the first moments of δ for each of its $3N\mathcal{C}$ components that we index i for clarity

$$\begin{aligned} \langle\delta_i\rangle &= \sum_{k,l} V_{k,l}\partial_k\partial_l\ \delta_i = \sum_k V_{k,i}\partial_k 1 = 0 \\ \langle\delta_i\delta_j\rangle &= \sum_{k,l} V_{k,l}\partial_k\partial_l\ \delta_i\delta_j = \sum_k V_{k,i}\partial_k\delta_j = V_{j,i} \end{aligned} \tag{6}$$

We see that our Gaussian model results in a distribution of δ where each of its component average to zero, and the average correlation between two components allows to define the inverse of matrix V^{-1} that was not define initially. Re-indexing properly, and writing $\vec{\delta_{ci}}$ the 3D vector that describes the deviation of atom i in unit cell c, we completely define our statistical model by considering the matrix V whose $3x3$ $V_{ci,dj}$ encodes for the correlations in the displacement between the displacement vector δ_{ci} and δ_{dj}: $V_{ci,dj} = \langle \delta_{ci} \delta_{dj}^T \rangle$.

To be continued ...

Here it is assumed that V is identical when evaluated for equivalent atoms in different unit cells: $V_{ci,ci} = V_{c'i,c'i}$. Then, the exponential decay term in the equation for $I(\vec{q})_{Bragg}$ can be evaluated as[1]:

$$\langle e^{-i\vec{q}\cdot\vec{\delta}_{ci}} \rangle = \int P(\vec{\delta}_{ci})\, e^{-i\vec{\delta}_{ci}\cdot\vec{q}}\, d\vec{\delta}_{i,c} = e^{-\frac{1}{2}\vec{q}^{\intercal} V_{ci,ci} \vec{q}} \tag{7}$$

and the equation for $I(\vec{q})_{Bragg}$ becomes:

$$\begin{aligned} I(\vec{q})_{Bragg} &= \sum_{c,c'} e^{-i\vec{q}\cdot(\vec{u_c} - \vec{u_c'})} \sum_{i,j} f_i f_j e^{-i\vec{q}(\vec{r_i} - \vec{r_j})} e^{-\frac{1}{2}\vec{q}^{\intercal} V_{ci,ci} \vec{q}} e^{-\frac{1}{2}\vec{q}^{\intercal} V_{c'j,c'j} \vec{q}} \\ &= \sum_{c,c'} e^{-i\vec{q}\cdot(\vec{u_c} - \vec{u_c'})} \sum_{i,j} f_i f_j e^{-i\vec{q}(\vec{r_i} - \vec{r_j})} e^{-\frac{1}{2}(\langle(\vec{\delta}_i\cdot\vec{q})^2\rangle + \langle(\vec{\delta}_j\cdot\vec{q})^2\rangle)} \end{aligned} \tag{8}$$

where the final exponential term corresponds to the Debye-Waller or conventional B factors. The first sum over unit cells takes the form of the Dirac comb in the limit that the number of unit cells, N, is large, and is order N^2 at reciprocal lattice points.

The ‘diffuse scattering’ is the residual scattering described by equation 1 that cannot be accounted for by $I(\vec{q})_{Bragg}$:

$$\begin{aligned} I(\vec{q})_{diffuse} &= \langle I(\vec{q})_{total} - I(\vec{q})_{Bragg} \rangle \\ &= \sum_{c,c'} e^{-i\vec{q}\cdot(\vec{u_c} - \vec{u_c'})} \sum_{i,j} f_i f_j e^{-i\vec{q}(\vec{r_i} - \vec{r_j})} [\langle e^{-i\vec{q}\cdot(\vec{\delta}_{ci} - \vec{\delta}_{c'j})} \rangle - e^{-\frac{1}{2}(\langle(\vec{\delta}_i\cdot\vec{q})^2\rangle + \langle(\vec{\delta}_j\cdot\vec{q})^2\rangle)}] \end{aligned} \tag{9}$$

Note that:

$$\langle e^{-i\vec{q}\cdot(\vec{\delta}_{ci} - \vec{\delta}_{c'j})} \rangle = \int e^{-\frac{1}{2}\vec{q}^{\intercal} V_{ci,c'j} \vec{q} - i\vec{q}^{\intercal}(\vec{\delta}_{ci} - \vec{\delta}_{c'j})} d\vec{\delta} = e^{-\frac{1}{2}\vec{q}^{\intercal}[V_{ci,c'i} + V_{cj,c'j} - 2V_{ci,c'j}]\vec{q}} \tag{10}$$

Thus,

$$\begin{aligned} I(\vec{q})_{diffuse} &= \sum_{c,c'} e^{-i\vec{q}\cdot(\vec{u_c} - \vec{u_c'})} \sum_{i,j} \tilde{f}_i \tilde{f}_j e^{-i\vec{q}(\vec{r_i} - \vec{r_j})} [e^{\vec{q}^{\intercal} V_{ci,c'j} \vec{q}} - 1] \\ &= \sum_{c,c'} e^{-i\vec{q}\cdot(\vec{u_c} - \vec{u_c'})} \sum_{i,j} f_i f_j e^{-i\vec{q}(\vec{r_i} - \vec{r_j})} e^{-\frac{1}{2}(\langle(\vec{\delta}_i\cdot\vec{q})^2\rangle + \langle(\vec{\delta}_j\cdot\vec{q})^2\rangle)} [e^{\langle(\vec{\delta}_{ci}\cdot\vec{q})(\vec{\delta}_{c'j}\cdot\vec{q})\rangle} - 1] \end{aligned} \tag{11}$$

where $\tilde{f}_i$ is the atomic form factor decayed by its Debye-Waller factor, $e^{-\frac{1}{2}\vec{q}^{\intercal} V_{ci,c'i} \vec{q}}$. The first version of equation 8 uses the interatomic covariance matrix, V, but could be equivalently written using the interatomic displacement covariance matrix, U, noting that $U = V_{ci,c'i} + V_{cj,c'j} - 2V_{ci,c'j} = \langle(\vec{\delta}_i - \vec{\delta}_j^{\intercal})^2\rangle$. We can then compute $I(\vec{q})_{diffuse}$ as the total intensity modulated by this single matrix:

$$I(\vec{q})_{diffuse} = \sum_{c,c'} e^{-i\vec{q}\cdot(\vec{u_c} - \vec{u_c'})} \sum_{i,j} f_i f_j e^{-i\vec{q}(\vec{r_i} - \vec{r_j})} [e^{-\frac{1}{2}q^{\vec{T}} U_{ij} \vec{q}} - 1] \tag{12}$$

Assuming there is no coupling between unit cells c and c’, then the covariance matrix V (or displacement covariance matrix U) is nonzero only for $c = c'$, in which case the sum over unit cells evaluates to N, the number of unit cells.

2 The form of diffuse scattering in specific disordered regimes

2.1 Translational disorder of the asymmetric unit

Suppose that individual protein molecules are displaced from their mean lattice positions, and that these displacement sample a Gaussian distribution. Also initially assume that the cell is P1, i.e. that there is one copy of the

[1] The term $\langle e^{-i\vec{\delta_i}\cdot\vec{q}} \rangle$ can alternatively be evaluated by expansion through a power series followed by averaging each term over the atomic displacement distribution function: $\langle e^{-i\vec{\delta_i}\cdot\vec{q}} \rangle = 1 + (-i\vec{\delta_i}\cdot\vec{q}) + \frac{(-i\vec{\delta_i}\cdot\vec{q})^2}{2!} + ... = 1 - \frac{1}{2}(\vec{\delta_i}\cdot\vec{q})^2 + ...$. (The average over the first order term is 0.) If the displacement distribution function is Gaussian, $\langle e^{-i\vec{\delta_i}\cdot\vec{q}} \rangle = e^{-\frac{1}{2}\langle(\vec{\delta_i}\cdot\vec{q})^2\rangle}$.

protein per unit cell. Then the expression for the diffuse intensity in equation 8 can be simplified by recognizing that $\forall (i,j)\, \vec{\delta_i} = \vec{\delta_j}$:

$$\begin{aligned} I(\vec{q})_{diffuse} &= \sum_{c,c'} e^{-i\vec{q}\cdot(\vec{u_c}-\vec{u_c'})} \sum_{i,j} f_i f_j e^{-i\vec{q}(\vec{r_i}-\vec{r_j})} e^{-(\vec{\delta}\cdot\vec{q})^2} [e^{\langle(\vec{\delta}\cdot\vec{q})^2} - 1] \\ &= \sum_{c,c'} e^{-i\vec{q}\cdot(\vec{u_c}-\vec{u_c'})} \sum_{i,j} f_i f_j e^{-i\vec{q}(\vec{r_i}-\vec{r_j})} [1 - e^{-(\vec{\delta}\cdot\vec{q})^2}] \\ &= N|F_{P1}(\vec{q})|^2 [1 - e^{-(\vec{\delta}\cdot\vec{q})^2}] \end{aligned} \tag{13}$$

where $|F_{P1}(\vec{q})|^2$ is the unit cell structure factor and N is the number of unit cells, respectively. For non-P1 cells, where individual protein molecules coincide with the asymmetric unit, equation 11 can be expressed as:

$$I(\vec{q})_{diffuse} = N \sum_{m=1}^{M} |F_{asu}(R_m\vec{q})|^2 [1 - e^{-(\vec{\delta}\cdot\vec{q})^2}] \tag{14}$$

where R_m is the m'th symmetry operator for the space group under consideration. This equation agrees with the derivation of Ayyer *et. al.* [2]

2.2 Unstrained conformational disorder

Suppose that displacements occur at the atomic rather than individual protein level, and that these displacements are uncorrelated between atoms. Here we refer to this type of disorder as 'strained.' Then, the sum over atom pairs in equation 8 will only be nonzero when $i = j$:

$$\begin{aligned} I(\vec{q})_{diffuse} &= \sum_{c,c'} e^{-i\vec{q}\cdot(\vec{u_c}-\vec{u_c'})} \sum_{i,i} f_i f_i e^{-i\vec{q}(\vec{r_i}-\vec{r_i})} e^{-\frac{1}{2}(\langle(\vec{\delta_i}\cdot\vec{q})^2\rangle + \langle(\vec{\delta_i}\cdot\vec{q})^2\rangle)} [e^{\langle(\vec{\delta}_{ci}\cdot\vec{q})(\vec{\delta}_{c'i}\cdot\vec{q})\rangle} - 1] \\ &= N \sum f_i^2 e^{-\langle(\vec{\delta_i}\cdot\vec{q})^2\rangle} [e^{(\vec{\delta_i}\cdot\vec{q})^2} - 1] = N \sum f_i^2 [1 - e^{-(\vec{\delta_i}\cdot\vec{q})^2}] \end{aligned} \tag{15}$$

In the above equation, $\vec{\delta_i}$ is related to the anisotropic displacement parameters (ADPs). Specifically, $\langle \vec{\delta_i}\vec{\delta_i}^\intercal \rangle$ forms the 3 x 3 symmetric tensor conventionally termed U, whose eigenvectors and eigenvalues correspond to the atomic displacement axes and the mean square displacements, respectively[2]. The isotropic B factor equivalent can be estimated from the trace of U: $B_{iso} = 8\pi^2 \langle \vec{\delta}^2 \rangle = \frac{8}{3}\pi^2 tr(W)$, where W is the diagonalized U.

2.3 Strained conformational disorder

Suppose that the displacements between different atoms are correlated. Here we will refer to such models of disorder as 'strained.' In the anisotropic case, $I(\vec{q})_{diffuse}$ can be evaluated from equations 8 or 9. In the case that disorder is assumed to be isotropic (or the model only has sufficient resolution to describe isotropic disorder), the (n x n x 3 x 3) anisotropic displacement covariance matrix can be reduced to the (n x n) correlation matrix, C, by the following relationship for the atom pair (i,j):

$$C_{ij} = \frac{\langle \vec{\delta}_i^\intercal \cdot \vec{\delta_j} \rangle}{\sqrt{\langle \vec{\delta}_i^\intercal \cdot \vec{\delta_i} \rangle \langle \vec{\delta}_j^\intercal \cdot \vec{\delta_j} \rangle}} = \frac{Tr(\langle \vec{\delta_i}\vec{\delta}_j^\intercal \rangle}{\sqrt{Tr(\langle \vec{\delta_i}\vec{\delta}_i^\intercal \rangle) Tr(\langle \vec{\delta_j}\vec{\delta}_j^\intercal \rangle)}} = \frac{Tr(\langle \vec{\delta_i}\vec{\delta}_j^\intercal \rangle}{b_i^{\frac{1}{2}} b_j^{\frac{1}{2}}} \tag{16}$$

where $b_i = \langle \vec{\delta_i}\vec{\delta}_i^\intercal \rangle$, and thus related to the Debye Waller factor: specifically, $b = \frac{B_{iso}}{8\pi^2}$, where B_{iso} is the values of the conventional B factor computed during crystallographic refinement. Then, the expression for $I(\vec{q})_{diffuse}$ in equation 8 simplifies to:

$$I(\vec{q})_{diffuse} = \sum_i \sum_j f_i f_j e^{i\vec{q}(\vec{r_i}-\vec{r_j})} [e^{-\frac{1}{2}\vec{q}^2(b_i+b_j-2C_{ij}\sqrt{b_i b_j})} - 1] \tag{17}$$

3 Inference of the Correlation Matrix

Let us assume that we have many measurements of I_{ensemble} and wish to infer a model of the form just discussed. To do this, we may minimize the least squares loss function

$$\mathcal{O}(\{C_{kk'}\}) = \sum_{\{\mathbf{q}\}} \left| I_{\text{esb}}^{\text{obs}}(\mathbf{q}) - I_{\text{esb}}^{\text{calc}}(\mathbf{q}; \{C_{kk'}\}) \right|^2$$

[2] The six unique values of U are listed in the ANISOU records of PDB files in the order U_{11}, U_{22}, U_{33}, U_{12}, U_{13} , U_{23}. The listed values are scaled by a factor of 10^4[3].

where "obs" and "calc" indicate the observed and modeled values respectively. Our task is to vary $\{C_{kk'}\}$ to minimize $\mathcal{O}$.

Let $A_{kk'}(\mathbf{q}) \equiv f_k f_{k'} \exp\{\, i\mathbf{q}(\mathbf{r_k} - \mathbf{r_{k'}}) - \frac{1}{2}q^2(B_k + B_{k'})\,\}$ and $\alpha_{kk'} \equiv (B_k B_{k'})^{\frac{1}{2}}$. Then we may write

$$I^{\text{calc}}_{\text{esb}}(\mathbf{q}; \{C_{kk'}\}) = \sum_{kk'} A_{kk'}(\mathbf{q})\, e^{\alpha_{kk'}C_{kk'}}$$

and the derivatives of the objective function take on a simple form

$$\begin{aligned}\frac{\partial\mathcal{O}}{\partial C_{kk'}} &= \sum_{\{\mathbf{q}\}} 2\alpha_{kk'} A_{kk'} e^{\alpha_{kk'}C_{kk'}} \left(A_{kk'} e^{\alpha_{kk'}C_{kk'}} - I^{\text{obs}}_{\text{esb}}\right)\\ &= \sum_{\{\mathbf{q}\}} 2\alpha_{kk'} I^{\text{calc}}_{\text{esb}} \left[I^{\text{calc}}_{\text{esb}} - I^{\text{obs}}_{\text{esb}}\right]\end{aligned}$$

this shows the derivative is zero and the objective is at an extreme when $C_{kk'} = \hat{C}_{kk'}$ with

$$\hat{C}_{kk'} = \alpha^{-1}_{kk'} \log \frac{\sum_{\{\mathbf{q}\}} I^{\text{obs}}_{\text{esb}}(\mathbf{q}) A_{kk'}(\mathbf{q})}{\sum_{\{\mathbf{q}\}} A^2_{kk'}(\mathbf{q})}$$

computing the second derivative

$$\frac{\partial^2\mathcal{O}}{\partial C^2_{kk'}} = \sum_{\{\mathbf{q}\}} 2\alpha^2_{kk'} A_{kk'} e^{\alpha_{kk'}C_{kk'}} \left(2A_{kk'} e^{\alpha_{kk'}C_{kk'}} - I^{\text{obs}}_{\text{esb}}\right)$$

and substituting $C = \hat{C}_{kk'}$

$$\frac{\partial^2\mathcal{O}}{\partial C^2_{kk'}}(\hat{C}_{kk'}) = 2\alpha^2_{kk'} \frac{\sum_{\{\mathbf{q}\}} A^2_{kk'}(\mathbf{q}) {I^{\text{obs}}_{\text{esb}}}^2(\mathbf{q})}{\sum_{\{\mathbf{q}\}} A^2_{kk'}(\mathbf{q})}$$

proves that this extremum is indeed the minimum and $\mathcal{O}$ is convex.

References

[1] P. Moore.

[2] K. Ayyer, O. M. Yefanov, D. Oberthur, S. Roy-Chowdhury, L. Galli, V. Mariani, S. Basu, J. Coe, C. E. Conrad, R. Fromme, A. Schaffer, K. Dorner, D. James, C. Kupitz, M. Metz, G. Nelson, P. L. Xavier, K. R. Beyerlein, M. Schmidt, I. Sarrou, J. C. Spence, U. Weierstall, T. A. White, J. H. Yang, Y. Zhao, M. Liang, A. Aquila, M. S. Hunter, J. S. Robinson, J. E. Koglin, S. Boutet, P. Fromme, A. Barty, and H. N. Chapman. Macromolecular diffractive imaging using imperfect crystals. Nature, 530(7589):202–206, Feb 2016.

[3] L. Zhou and Q. Liu. Aligning experimental and theoretical anisotropic B-factors: water models, normal-mode analysis methods, and metrics. J Phys Chem B, 118(15):4069–4079, Apr 2014.