Summary of lecture at biophysics and structural biology at synchrotrons workshop Crystallography 101: diffraction theory and space groups

Elsbeth F. Garman¹

Published online: 9 July 2019

© International Union for Pure and Applied Biophysics (IUPAB) and Springer-Verlag GmbH Germany, part of Springer Nature 2019

Commentary

To determine the three-dimensional structures of proteins, we need to ‘look’ at them using electromagnetic radiation which has a similar wavelength to the interatomic distances in the molecule. These are typically around 1.5×10^{-10} m (= 0.15 nm = 1.5 Å) but visible light has wavelengths between 400 and 700 nm which is far too long to enable us to visualise the detailed structures of proteins and viruses. We thus usually use X-rays, neutrons or electrons of much shorter wavelength: for X-rays between 0.7 and 1.7 Å which are ideal for this. However, unlike in a light microscope, we do not have any lenses that can form an image from X-rays scattering from such small objects. We have to rely on the interference patterns (so-called ‘diffraction’) formed between the X-rays as they scatter elastically (no energy loss in the sample) from crystals of the macromolecule. We then deconvolute these patterns using the tools of Mathematics (Fourier transforms) and some extra information (the ‘phases’) that we can obtain either experimentally or from a known macromolecular structure.

This crystallographic method allows us to find the structures of biological molecules over an enormous range of scales from small proteins to large macromolecular complexes and whole viruses. For instance, the structure of human insulin, a 5.8 kDa hormone vital for our metabolism of carbohydrates, fats and proteins, is ~20 Å in diameter, and using X-ray crystallography, a structure to 1 Å resolution has been determined. At the other end of the size range, the 70S

ribosome, a 2.7 MDa complex of proteins and RNA which is about 200 Å across, and the ~700 Å diameter and ~70 MDa core of the Bluetongue virus have both been solved to a resolution of 3.5 Å, allowing elucidation of biological function and mechanism.

Protein crystals are typically >30% solvent, and less than 30% of a protein surface is involved in contacts between the ordered molecules, so they can thus be thought of as very high concentration solutions (typically ~10 mM). Imagine many rows of synchronised swimmers lined up in a swimming pool in an ordered arrangement, with more layers of swimmers going underwater too, with each swimmer representing one protein molecule. Although overall the swimmers are dressed the same, they are not quite identical in size and demeanour. So it is with protein molecules: the structures we obtain by macromolecular crystallography (MX) are time and space averages of the many millions of molecules in the crystal, rather than being the structure of a single molecule. Thus, dynamic regions (e.g. flexible loops, the arms of the swimmers) will be less well defined in the final electron density maps than the more static parts of the protein. Functionally, it has been shown that many enzymes are active in the crystalline state, and crystal structures generally agree well with measurements from spectroscopic techniques (e.g. NMR, Fourier transform IR, fluorescence). Proteins are often observed in more than one ‘packing’ arrangement in different crystal forms, but generally, the differences in their structures are either slight (~0.1 Å) or conversely are potentially biologically relevant (e.g. multiple conformational states).

To form a crystal, copies of a simple component (the unit cell) are packed together by translation to fill space, and this packing must be infinitely extensible in all directions. A plane lattice is an infinitely extensible construct of intersecting parallel and equidistant lines, forming identical unit lattices. In 3-D, this is a basic parallelepiped (or hexagonal or trigonal sided cylinder) shaped block from which the whole volume of the crystal may be built by repetition in 3 dimensions. Its axes

This article is part of a Special Issue on ‘Biophysics & Structural Biology at Synchrotrons’ edited by Trevor Sewell.

✉ Elspeth F. Garman
elsbeth.garman@bioch.ox.ac.uk

¹ Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

form a set of vectors labelled $\underline{a}$, $\underline{b}$ and $\underline{c}$, with the angle enclosed by any two of them being the remaining complementary Greek letter (e.g. angle between $\underline{a}$ and $\underline{c}$ is β). Any point in the repeating volume can be specified with respect to the origin by the fractional coordinates x , y , z measured parallel to the unit cell axes.

X-ray scattering The electrons round the atoms interact with the electromagnetic wave (the X-ray), and this makes them oscillate. These oscillating electrons (i.e. accelerating charges) act as new X-ray sources. Where there are multiple atoms giving rise to a charge distribution, the scattered beams interfere, and the pattern is therefore no longer spherically symmetrical. The scattered intensity in any direction depends on the charge distribution throughout the illuminated sample (i.e. on its atomic structure).

Diffraction and Bragg's law The diffraction phenomenon itself can be modelled as the reflection of X-rays, represented by sine waves, from ordered planes of molecules in the crystal. In any particular direction, the intensity of the wave emerging from the crystal will depend on the molecular charge distribution as described above. Because the molecules are regularly arranged in the crystal, the emerging rays will only be visible in those directions in which the peaks and troughs are aligned so that they constructively interfere to produce a diffraction signal. However, if the peaks and troughs of the emerging waves are not aligned, they destructively interfere and there is no X-ray intensity reflected in that direction. This diffraction condition is formulated by Bragg's law:

$$n\lambda = 2d \sin \theta$$

where λ is the wavelength of the incident radiation, d is the distance between the reflecting planes, θ is the angle of incidence and reflection, and n is an integer (see Fig. 1). Of course, there are many sets of planes for a given spacing, d , as we go down through the unit cell, and the waves scattered from these families of planes all sum together to give a wave with an amplitude, F , and an overall phase, ϕ . The pair of numbers F and ϕ is called the structure factor. We see this reflection as increased X-ray intensity, $I = |F|^2$, on our position sensitive X-ray detectors, but we do not have any time resolution to monitor the phase (for $\lambda = 1 \text{ \AA}$ X-ray, one sine wave period corresponds to $\sim 3 \times 10^{-19} \text{ s}$, so it is impossible to detect it). Thus, it transpires that the intensities that we can measure are only half the information we need to reconstruct the electron density of the molecule from the diffraction pattern. This is the so-called 'phase problem' in MX.

From Bragg's law, we can see that for a particular incident X-ray λ there is an inverse relationship between the scattering

Fig. 1 A cake illustrating Bragg's law: $n\lambda = 2d \sin \theta$ where λ is the wavelength of the incident radiation, d is the distance between the reflecting planes, θ is the angle of incidence and reflection and n is an integer

angle θ and the plane spacing: for large θ , closer planes are being sampled and we will obtain higher detail, i.e. better resolution.

Miller indices We can define sets of 2-D planes which intersect the three unit cell axes ($\underline{a}, \underline{b}, \underline{c}$) of the crystal an integer number of times and at a range of angles. Then, we can number the reflection that arises from a set of planes a distance d apart as (h, k, l) where h , k and l are the number of times per unit cell that a set of planes intersects the $\underline{a}$, $\underline{b}$ and $\underline{c}$ axes, respectively, and are called the Miller indices. For example, a diagonal plane through a unit cell that intersects $\underline{a}$ 6 times, $\underline{b}$ 2 times and $\underline{c}$ 4 times would produce a reflection with indices $(6, 2, 4)$. The higher the Miller index, the smaller will be the distance between the planes and the larger will be the scattering angle. Thus, the high-resolution data will give reflections far from the main beam position at $(0, 0)$ (note the reflections are at 2θ , not θ to the main beam position).

The position of the reflections in the diffraction pattern depends only on the shape and size of the unit cell, but the intensity and phase of the spot depend on the structure of the molecule. This means that for crystals of 2 different proteins which had identical unit cells, the only difference in their diffraction patterns would be the intensity of the spots: their spacing would be the same.

Reciprocal space We can rearrange Bragg's law to make a direct relationship between θ and $1/d$: $2 \sin \theta = n \lambda / d$, i.e. $\theta \propto 1/d$. Thus, for large d , θ is small and vice versa. We can construct a reciprocal lattice, based on $1/d$ varying directly with $\sin \theta$. Since the sets of planes diffract *only* if they cut each of the unit cell edges an *integral* number of times, we can imagine a conceptual grid of reflections related to the geometry of the repeating unit of the crystal. This grid is called the

‘reciprocal lattice’ and is a useful tool for considering the family of diffracted reflections that result from a crystal. The larger the unit cell, the smaller the difference between Bragg angles for neighbouring sets of planes in this lattice, i.e. spot separation $\propto 1/\text{unit cell}$.

Resolution The resolution of the data pivotally affects the detail we can see in the structure. In general, at 5 Å, we can see the shape of the molecule and helices stand out as rods; at 4 Å, even β -sheets are traceable; at 3 Å, we can usually get the main chain fold right, and see the positions of the side chains; at around 2.5 Å, the main chain is decorated with carbonyl bulges and there is a clear choice for the side-chain rotamer; and at 1.5 Å the structure is essentially at atomic resolution with most non-hydrogen atoms in the correct place.

Fourier transforms The physical phenomenon of X-ray scattering from molecules maps naturally onto the mathematical construct of the Fourier transform (FT), and thus enables an established formalism to be used to analyse crystallographic data. To understand and become wiser (Fig. 2) about Fourier theory through studying some excellent illustrative figures, it is worth visiting Kevin Cowtan’s WWW page at York University: <http://www.ysbl.york.ac.uk/~cowtan/fourier/fttheory.html>.

Fig. 2 A wiser crystallographer after attending the 101 lecture. Dr. Alice Brink and a perching owl at Eagle Encounters, Spier Wine Farm, Stellenbosch, Republic of South Africa

Consider the convolution of the FT of a macromolecule (which is a smooth continuous function) and the FT of the crystal lattice (an infinite three-dimensional grid which is discontinuous and just has histogram peaks at each lattice position with nothing in between). The Fourier transform of the crystal lattice (i.e. the infinite grid) is the reciprocal lattice (i.e. another infinite grid). The product of these two FTs is the diffraction pattern of a protein crystal. It is a smooth function sampled at reciprocal lattice points. The intensities of the reflections are thus modulated by the molecular shape, and the spacing of the reflections is determined by the unit cell, as already stated above.

FTs obey an inversion theorem which states that if function F is the Fourier transform of function G , then function G can be obtained from the Fourier transform of function F . Thus, in the way as we can use Fourier transforms to calculate the diffraction pattern of a protein crystal, we can use Fourier transforms to calculate the electron density distribution in a crystal from an observed diffraction pattern. When both the amplitudes and phases of the structure factors are known, an image can be reconstructed from diffraction data. However, again as we saw above, a diffraction image contains information on the amplitudes of structure factors but not their phases: hence the ‘phase problem’ in MX.

Note that every atom in the crystal contributes to every reflection, and each pixel of the reconstructed image of the electron density of our macromolecule depends on all of the diffracted waves.

Space groups If we apply a symmetry operation to the objects in the unit cell, we are not allowed to generate any changes in the motif. Thus, for chiral molecules such as proteins, all inversions or mirror operations are forbidden and all symmetry operations must also be compatible with the translational requirements of a given lattice. Any rotations can only involve operations by $(360/n)^\circ$ where n is an integer. Thus, only 2-, 3-, 4- and 6-fold rotations are allowed and are indicated in the International Tables for Crystallography by dyad, triad, tetrad and hexad symbols. Rotational symmetry must be compatible with lattice symmetry, e.g. a square plane lattice is not compatible with a 3-fold rotation. There are a finite number of crystal lattices and a finite number of ways of combining the crystal lattice with other symmetry elements. The combination of crystallographic symmetry elements present defines the ‘space group’ (65 possible for proteins, 230 for non-chiral molecules). The symmetry in the crystal lattice is paralleled by double the symmetry in the diffraction pattern.

Structure factors and electron density The relationship between the scattering and atomic (or electron density) distribution is an FT. The prediction and interpretation of scattering patterns can be understood in terms of the convolution and inversion theorems of FTs. For FT convolution, a protein

crystal is described as a mathematical function by a structural motif (the protein or proteins) repeated at each position of an infinite lattice. By the FT inversion process by which the inverse of a FT is another FT, an image of a crystal can be regenerated from an FT of the diffraction pattern. However, we do not have the phases for the FT that would give us the electron density. We have to find them experimentally by multiple- or single-isomorphous replacement (MIR, SIR), by multi- or single-wavelength anomalous dispersion (MAD, SAD) methods, or combine one of these modalities with anomalous scattering signal. Alternatively, we can use ‘molecular replacement’ where we use the phases from a homologous structure which is already known. Given crystals of the target protein, determining the phases can then be the main challenge!

Acknowledgements This is a very short summary of a 2.5-h lecture which has evolved over many years of teaching crystallography. It

contains input from Martin Noble and Airlie McCoy, and I am very grateful to both of them for continued discussions on improving our teaching of these topics, which are conceptually challenging.

Funding The funding for the Workshop came from the GCRF START (Synchrotron Techniques for African Research and Technology), IUPAB and the SA National Research Foundation (NRF).

Compliance with ethical standards

Conflict of interest Elspeth F. Garman declares that she has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by the author.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.