

# Crystallographic studies of biological macromolecules: from atomic resolution to molecular giants

Mariusz Jaskolski

Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University and Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

When we look at an X-ray diffraction pattern of a protein crystal, we usually see myriads of reflections. But two other things are also striking: (1) that the reflections are very closely spaced and (2) that their intensity falls off rather quickly as we move away from the center, i.e. from the direction of the primary beam. Usually there is also another curious feature, namely (3) a very strong background visible as a diffuse dark ring at a certain angle with the primary beam. This latter feature is due to the scattering of X-rays by disordered water molecules which always accompany protein molecules in their crystals. Protein-to-water volume ratio is typically 1:1. This property of protein crystals is on the one hand a blessing for the protein crystallographer, because it guarantees that protein molecules even in crystalline form are in their native aqueous environment and thus have native structure. But it can be a curse on the other hand, because with weak, water-shielded direct protein-protein contacts the degree of molecular order can be less than perfect, which in consequence leads to poor diffraction and to poor structure determination.

Bulk water in protein crystals is "structured" as liquid water, i.e. we have an endless network of tetrahedrally arranged water molecules connected by hydrogen bonds, whose donor/acceptor properties fluctuate throughout the entire network. Typical O...O distances in those hydrogen bonds are 2.7 Å. This leads to a great number of 1-3 O...O vectors, whose directions, but not the length, can change. Those repeated  $d=4.4$  Å distances lead to scattering of the X-rays at a certain angle  $\theta$ , easily calculated from **Bragg's Law**:  $\lambda=2d.\sin\theta$ .

With tuneable X-ray radiation, such as obtained in a synchrotron, the "water ring" recorded on a flat detector perpendicular to the primary beam can be bigger or smaller depending on the wavelength  $\lambda$  of the radiation (typically about 1 Å), but even at a constant  $\lambda$  we can change its radius by moving the detector in and out. It would appear that it would be advantageous to move the detector far from the crystal, because then the closely spaced diffraction spots would become well separated. While we indeed do want to achieve a physical resolution of the diffraction spots on the detector, there is a different consideration, connected with another meaning of the term "**resolution**" which prompts us to move the detector as close to the crystal as possible. Why?

To understand this we must first realize that X-rays are scattered by electrons, primarily by those in atomic cores but also by bonding electrons. The diffraction image is a **Fourier Transform** of the scattering object, i.e. of the electron distribution (also called electron density) in the crystal. To be able to calculate back the information about this electron density is of great importance to us. In chemistry – everything is explained by electrons: the nature of different atoms and the bonds between them. It is thus very fortunate that we have a mathematical apparatus, the Inverse Fourier Transform, allowing us to calculate **electron density maps** in the crystallographic unit cell. They will tell us all we want to know about the chemical molecules that build up our crystal.

However, there is a serious obstacle on our way from the diffraction pattern to the electron density map,  $\rho(xyz)$ , or to the crystal structure. It is known as the **phase problem** because in the simple Fourier formula  $\rho(xyz)=\Sigma F(hkl)\exp[-2\pi i(hx+ky+lz)]$ , the so-called **structure factors**  $F(hkl)$  are known from the diffraction experiment with respect to their amplitudes, calculated

simply as  $|F(hkl)| = \sqrt{I(hkl)}$  (where  $I(hkl)$  is the intensity of reflection  $hkl$ ), but not with respect to their phases.

The formidable task of estimating the phases of many thousands of individual reflections can be solved on three ways, each of which relies to some extent on a peculiar Inverse Fourier Transform,  $P(uvw) = \sum |F(hkl)|^2 \exp[-2\pi i(hu + kv + lw)]$ , known as the **Patterson Function**. While the desirable function  $\rho(xyz)$  represents the distribution of atoms in the crystal unit cell,  $P(uvw)$  (which mathematically represents an autocorrelation function or the convolution of the atomic structure with its centrosymmetric image) represents the distribution of all interatomic vectors. It is clear that for large structures, such as protein structures, that contain thousands of atoms, the Patterson Function is astronomically complex, containing millions of vectors. But it's easily calculated and, with judicious use, is of great help.

The Patterson Function finds the most straightforward application in the method of **Molecular Replacement**. Here, we have an approximate atomic model of our macromolecule from which we can generate all the interatomic vectors. The problem of solving an unknown crystal structure is then reduced to finding the correct rotation and translation of this set of known vectors in the unit cell of our unknown structure.

The other two methods solve the phase problem by first locating in the unit cell of a small number of special atoms (special because they must scatter the X-rays in a special way), which become the starting point for deciphering the complete structural puzzle. The classic method of **Isomorphous Replacement**, developed by the pioneer of protein crystallography, Max Perutz, uses very heavy, electron-rich atoms which are attached to the protein molecules in an isomorphous way, i.e. without altering the crystal structure. If we are lucky, the differences in the diffraction pattern of the derivative and native crystals can reveal (via a Patterson Function) the locations of the heavy atoms, which are the first, very crude, approximation of the complete structure.

The third method is based on a somewhat similar principle but it uses, as phasing marker, atoms that do not need to be very heavy but must scatter the X-rays in an anomalous way. **Anomalous scattering** occurs when the energy of the X-ray quanta is tuned to (in resonance with) the electronic energy levels of the scattering atom. To exploit the method of anomalous scattering, we must be able to tune the wavelength of the X-ray beam, something that is possible with **synchrotron radiation**, and have a special atom type in our crystal structure. The normal protein atoms (C, N, O, H, and S) are not good for anomalous scattering. Therefore, a trick is used to introduce into the protein molecule several selenium (Se) atoms, which can be excellent anomalous scatterers. Typically, the experiment will be conducted at several carefully adjusted wavelengths, which gives the method its name, **Multiwavelength Anomalous Diffraction** or **MAD**. The trick with the Se atoms is to replace in the make-up of our protein the natural sulfur-containing amino acid methionine with its close chemical cousin containing selenium. The replacement is possible if we harness bacteria to manufacture our protein and supply them with selenomethionine instead of methionine. Although it sounds very bizarre, this method is commonplace practice of genetic engineering. It is easy to introduce into a bacterial cell the DNA coding our desired protein, and turn the cell into a protein-producing factory.

Even if we solve the phase problem, the electron density map still has to be interpreted by an atomic model, and this model has to be refined. In the refinement, we calculate the structure factors using the Fourier Transform  $F_c(hkl) = \sum f_j \exp[-B_j \cdot (\sin\theta/\lambda)^2] \exp[2\pi i(hx_j + ky_j + lz_j)]$  and introduce adjustments into our model to bring the calculated values  $F_c$  to optimal agreement with

the experimental measurements. For each atom of the model, the corrections are made to its coordinates  $x,y,z$  and to a parameter, the **Atomic Displacement Parameter** (B) that describes the amplitude of its vibrations in the crystal lattice.

Although it looks very straightforward, the refinement is a difficult step, mainly because of the huge number of parameters. If our protein contains 5000 non-H atoms, then a simple isotropic model, which assumes a very inadequate approximation that the vibrations of each atom are isotropic, i.e. have the same amplitude in all directions, will require  $4 \times 5000 = 20000$  parameters. Not infrequently, this number will be close to the total number of available experimental data, making the problem barely solvable from the mathematical point of view. To improve the situation, two approaches are possible. First, we supplement the refinement with extra equations, called **restraints**, which represent our prior knowledge about the stereochemistry of the macromolecule under refinement. We can, for instance, require that the bond lengths (or angles, or other geometrical parameters) of our model have reasonable values.

The second strategy is very simple: get more data! But how can this be achieved? Experimentally, more diffraction spots can be registered by increasing the "acceptance angle"  $2\theta$  at our detector. Through Bragg's equation, an increase of  $\theta_{\max}$  is equivalent to a decrease of  $d_{\min}$ , the minimum spacing of lattice planes which still reflect the X-rays. This minimum  $d$ -value, expressed in Å, is called the **resolution limit** of our data. Optical considerations show that it is equivalent to the optical resolution of our model. In other words, if we collect diffraction data to 2.0 Å resolution, we can see in our electron density maps features that are more than 2 Å apart, but will not be able to distinguish, for instance atoms, that are more closely spaced. Fortunately, we know the basic stereochemistry of our macromolecules, so it is possible to construct an atomic interpretation of an electron density map even if it does not have true atomic resolution, but it is obvious that our main struggle should be to obtain experimental data with the highest possible resolution. Because only then will we be able to see our structure atom-for-atom, refine **Anisotropic Displacement** model (adding six variables per atom!) and see fine feature that are not visible, or blurred by the restraints, in poorly resolved maps.

As a criterion for atomic resolution, 1.2 Å has been accepted because 1.2 Å is the shortest covalent bond in proteins (C=O) not involving H atoms. One might think that it is a trivial thing to collect high-resolution data, a mere technicality. But this is not so because it is usually the crystal that "determines" the maximum resolution. In most cases there is no point in increasing  $\theta_{\max}$ , simply because there is nothing to measure beyond a certain limit. The reasons are several: the atomic scattering factors,  $f_j$ , fall-off with  $2\theta$  quite rapidly, the atomic vibrations smear out the electrons making scattering less effective. But most importantly, protein crystals have only a limited degree of crystalline order (connected with the mixed water-protein composition of their interior) which bears directly on their ability to scatter X-rays coherently.

Advancements in different areas of protein crystallography help pushing the limits towards higher and higher resolution. The progress has been especially spectacular in connection with various **structural genomics** initiatives. The first atomic-resolution protein structures appeared in the **Protein Data Bank (PDB)** in mid 1980's. Today, among the almost 60000 macromolecular structures deposited in the PDB, about 1000 are of atomic resolution. This provides us with an entirely new chemical perspective on the structure and functioning of the molecules of life.

In addition to looking into macromolecular structures with a more penetrating eye, protein crystallography is also attacking problems of ever increasing complexity. Viruses are among the

largest molecular systems whose atomic details have been deciphered by protein crystallography. The first crystal structures of **viruses** (both helical and icosahedral) were solved already in 1978, and even then the resolution was better than 3 Å. Currently, there are several hundred virus structures in the PDB.

Recently, the biggest triumph of macromolecular crystallography has been the mapping of the atomic structure of the ribosomal subunits (2.4-3.0 Å) and of the entire **ribosome** (2.8 Å), in complex with mRNA and tRNA molecules. The scale of this achievement is illustrated by the mass of this huge cellular machine, which is measured in megadaltons, corresponding to nearly two hundred thousand non-H atoms. The ribosome is composed of both proteins and ribonucleic acid molecules. One of the most unexpected secrets revealed by the structure of the ribosome was that its catalytic activity is associated with the ribonucleic acid component, and not with the proteins.

Protein crystallography has also re-defined modern approach to **drug discovery** by providing precise molecular targets for accelerated, structure-guided design of new pharmaceuticals. The best known example is illustrated by the structure of HIV protease, which very quickly after its discovery has become the most studied target for drug discovery. As a result, HIV infection has been converted, within less than a decade, from a global health threat, and for an individual patient – an irrevocable death sentence, into a disease that can be treated.

The above examples illustrate how through unveiling macromolecular structures with increasing accuracy and at increasing level of complexity, the discipline of protein crystallography helps us to better understand the secrets of biological macromolecules, and in consequence – the secret of life.