Table 2.1. Data collection statistics for a typical protein structure determination. The numbers in parentheses refer to the highest resolution bin and are indicative of how well the higher resolution data are measured. (Data adapted from the structure determination of glucosidase at pH 4.5; PDB code 2NT0)[3]

Space group	$P2_1$
Cell dimensions	110.5,91.8,152.8Å 90, 111.2, 90°
Resolution (Å)	34–2.2 (2.28–2.20)
R_{sym}	10.3 (47.3)
$I/\sigma(I)$	9.8 (2.2)
Completeness (%)	96.4 (91.2)
Redundancy	2.5 (1.8)

which data are included. A good data set is characterized by an overall R_{merge} value of about 5% or less. A value higher than ~10% suggests less than optimal data quality. At the highest resolution shell, R_{merge} can reach as high as 40% for low-symmetry crystals and 60% for high-symmetry crystals, the difference being a reflection of the level of redundancy, which is higher for high symmetry crystals.

(C) Finally, the completeness of the data is an important factor in determining data quality. Completeness is determined by comparison with the expected number of reflections for a particular space group and unit cell size, and given as a percentage. Because the ability to measure reflections decreases with resolution, completeness also decreases with resolution, so this parameter should be given for all data and for the highest resolution shell as defined for the R_{merge}.

In general, the data should have as high a resolution range as possible, with a high signal-to-noise ratio (>10), well-separated reflections, a low R_{merge} (<10), and high completeness (overall, it is acceptable to have relatively low completeness in the highest resolution shells). How well these factors interact with each other will determine the quality of the electron density map that is obtained. In practice, these measures may not be ideal, but in general, the higher the quality of the data, the greater the likelihood that they will lead to an interpretable electron density map.

However, the measures should never be used as a substitute for judgment in deciding whether to "believe" a structure or not. They are merely rough guidelines. There are many examples of acceptable structures from data of marginal quality, and, unfortunately, a few examples of wrong structures from excellent data. It is true, though, that the most important quantity is the resolution. The higher the resolution of the data the greater the likelihood that the electron density will have been interpreted correctly. Most serious mistakes in protein crystallography have resulted from overinterpretation of poor-quality electron density at relatively low resolution.

PHASING

Electromagnetic radiation can be defined in terms of waves that are defined in terms of an amplitude and a wavelength. The phase is the relative time of arrival of the crest of the wave at a reference point, compared with any other wave. Waves of identical phase will have their peaks and troughs in common and will sum accordingly. Waves with opposite phases will tend to cancel one another out, at least partially depending on their amplitudes. Both parameters are required to define a wave mathematically. To solve a crystal structure, in principle all one has to do is add up all of the diffracted waves; that is what a Fourier synthesis is. Before that can be done, however, the two parameters must be determined for every scattered wave (i.e., every reflection).

Experimentally, the amplitude manifests itself as the square root of the measured intensity of the reflection. That is easily determined with modern area detectors. However, when waves are added, they must be added with their correct phases. Consequently, to apply the Fourier transform to a reflection, both a measure of intensity and a correct phase are required. Unfortunately, in a diffraction experiment, although the intensities and positions of the diffracted waves are measurable, the phases of the reflections are not. X-rays travel at the speed of light, so as far as we are concerned, the relative time of arrival of all of the scattered waves from the crystal at the detector will appear to be the same. Consequently, the phases must be determined in some other way.

The most common method of phasing, particularly in drug design, is molecular replacement. The method relies on two factors: (1) that the structure of the protein of interest, or that of a very similar protein, has already been determined and (2) on the observation that the diffraction pattern of the object of interest is very similar to that of a related or similar object. In molecular replacement one measures the diffraction amplitudes from the crystal of the protein of interest but "replaces" their unknown phases with phases calculated from the previously determined structure of the related protein. The dominant issue that determines success with this method, and that makes it possible, is the level of similarity between the two objects. When determining the structure of a protein/ligand complex, for instance, the expectation that the binding of the ligand produces only minor changes in the structure of the protein is usually a good one, and in such cases the known structure of the apo protein or that of the protein with another ligand bound can be used as the model from which to obtain phases.

The importance of the phases in the determination of a structure cannot be overemphasized and can be demonstrated. An electron density map, calculated from the correct structure amplitudes but incorrect phases, is uninterpretable. Conversely, an electron density map determined from random structure amplitudes but correct phases is often interpretable, albeit very noisy. For the purposes of drug design, these two rules can be combined: