# 18.3. Structure quality and target parameters

By R. A. Engh and R. Huber

## 18.3.1. Purpose of restraints

> Believe statistics! Shun bias! – these, we see, are two materially different laws. (Adapted from William James.) A wise man, therefore, proportions his restraints to the evidence. (Adapted from David Hume.)

If we could adequately measure the complex diffuse scattering function of an X-ray beam from a single macromolecular structure to arbitrary resolution, we could, given an accurate model for X-ray scattering, calculate a measurement-time-averaged electron-density distribution for that structure independent of most model assumptions. Alternatively, if we knew the potential energy for the system accurately as a function of all relevant parameters, classical and non-classical, and had the computational power to analyse the function appropriately, we could model the same distribution. If both experiment and theory were accessible at these extremes, the available information would be redundant and could be used together in arbitrary ways. Of course, both experiment and theory are limited far below these extremes, and the researcher must consider what information is available in order to address the pertinent protein-structure question in the most appropriate way.

The most general theoretical assumptions used to interpret experimentally determined X-ray reflections typically include the spherical symmetry of scattering functions centred at nuclei, harmonicity (and isotropy) of deviations about average positions, infinity of a crystal lattice of identical unit cells and so on. While the majority of structures determined to date share a standard set of assumptions, significant alternatives have been proposed, such as, for example, the modelling of motion with normal-mode analyses instead of harmonic temperature factors (Kidera *et al.*, 1994). Any significant change in the set of assumptions used in structure determination must be expected to leave its mark on the statistical properties of the structure.

Besides the necessity of general model assumptions for structure determination, there is the simple necessity to supplement the measured reflection data with additional sources of information to improve the ratio of experimental observables to model parameters. For example, a crystal of a protein with some 3000 atoms that diffracts to a moderate resolution of 2.5 Å might produce some 14 000 measurable unique reflections. In this case, fitting a model that includes three Cartesian coordinates and a temperature factor for each atom (12 000 parameters) is effectively underdetermined when considering experimental error. The solution is either to reduce the number of parameters by introducing constraints to the model, or to increase the number of effective observations by using additional restraints in the structure determination.

### 18.3.1.1. *Utility of restraints: protein/special geometries*

One set of restraints arises from the expectation that the macromolecular structure should be chemically reasonable, that is, it should reflect the geometries required by chemical physics. This restricts especially values of bond lengths, bond angles, planarity and improper dihedral angles. Additional restriction comes from the expectation that these values should most closely conform to structures determined by the same experimental method and corresponding model assumptions. There are now two sources of structural information of sufficient quality for use as restraints in structure refinement: chemical fragments from the Cambridge Structural Database (CSD) (Chapter 23.4) and the very high resolution protein structures that can be refined without recourse to restraining parameters (Longhi *et al.*, 1998; Dauter *et al.*, 1997). Restraints derived from the CSD have come into widespread use for protein (Engh & Huber, 1991; Brünger, 1993; Priestle, 1994) and nucleic acid (Parkinson *et al.*, 1996) refinement; since their derivation, the database has grown from some 80 000 structures to around 200 000. The number of very high resolution protein structures solved has also increased to include dozens of structures (or thousands of chemical fragments). Chemical-fragment geometries can be studied using either or both sets of structures.

The optimal choice of restraints to use depends on the character and expected use of the refined structure. For example, the quality of comparative studies (Kleywegt & Jones, 1998) is enhanced by standardized refinement conditions, removing at least one source of systematic variation (*e.g.* Laskowski, Moss & Thornton, 1993), although effects from individual crystals and data collection and processing remain. On the other hand, a structure with an unusual chemical environment might require a tailored parameterization scheme to study the effects of this environment. Such cases might occur, for example, at enzyme active sites, where bound ligands may be distorted toward intermediate geometries (Bürgi & Dubler-Steudle, 1988).

### 18.3.1.2. *Risk of restraints: bias, lack of cross validation*

As an alternative to implementation as restraints, statistical information from known structures provides an independent check of structural integrity. However, the more information which is used for restraints, the less is available for such cross validation. For example, the use of a force field in protein refinement that strictly enforces a physical distribution of the protein backbone $\varphi$–$\psi$ angles of the Ramachandran plot would transfer error-induced strain into other degrees of freedom while eliminating a Ramachandran analysis as a tool for judging protein quality (Hooft *et al.*, 1997). There is an additional risk associated with the intentional introduction of bias into protein refinement: Restraining a model to conform to known structures is in error if the structure is unique in some way and therefore should *not* conform to these structures, or if there is erroneous bias in the set of structures from which the restraints were derived. Furthermore, the bias may be exaggerated, especially if the biased feature is so probable that deviations are considered suspect. This can be seen, for example, in the increasing frequency of *cis*-prolines as protein-structure resolution increases (Stewart *et al.*, 1990): when there is uncertainty in lower-resolution structures, the more probable *trans*-prolines are preferentially chosen in model building, exaggerating their frequency.

## 18.3.2. Formulation of refinement restraints

*A priori* information regarding protein structure can be used in two fundamental ways: as constraints by fixing parameters to target values, or as restraints by allowing limited deviation from target values. The two methods differ fundamentally when counting the total number of degrees of freedom (important, for example, when thermodynamic quantities of simulations are considered), but both improve the observation-to-parameter ratio (constraints reduce parameters, restraints increase observations). In this chapter, we focus on the use of restraints.

A common form of restraint is an energy function parameterized to represent a conformational energy of a protein, driving the refinement toward a low-energy conformation but allowing reasonable deviations from it. Strictly, however, protein-structure refinement involves fitting model parameters to data measured from the ensemble of structures in the crystal; the model parameters