

## 2.2. QUALITY INDICATORS

indicators here. For a thorough treatment of translation-function applications, the reader is referred to Chapter 2.3 of *International Tables for Crystallography* Volume B and Chapter 13.3 of the present volume.

**Log-likelihood gain, LLG.** In likelihood-based molecular replacement (McCoy *et al.*, 2007), potential molecular-replacement solutions are evaluated using likelihood, which is defined as the probability  $P$  that the observed diffraction data would have been measured if the orientation (and, usually, position) of the model were correct (Read, 2001). The score is reported in terms of the log-likelihood gain (LLG), which is defined as the logarithm of the likelihood score for the model  $p(F_{\text{obs}}; \text{model})$  minus the logarithm of the likelihood score for a random-atom Wilson distribution  $p_{\text{Wilson}}(F_{\text{obs}})$ . The LLG measures how much better the data can be predicted from the molecular-replacement model than from a collection of random atoms.

$$\text{LLG} = \sum_{hkl} \ln[p(F_{\text{obs}}; \text{model})] - \sum_{hkl} \ln[p_{\text{Wilson}}(F_{\text{obs}})]. \quad (2.2.7.2)$$

**LLG-Z score.** It is important to note that the LLG depends on the quality of the model and the number of reflections, so the absolute values cannot be compared between different molecular-replacement applications. Instead, the quality of a molecular-replacement solution can be judged by the LLG-Z score, which is defined as the number of standard deviations a score is above the mean score in a particular rotation or translation search.

$$\text{LLG-Z} = \text{LLG} - \langle \text{LLG} \rangle / [(\text{LLG} - \langle \text{LLG} \rangle)^2]^{1/2}. \quad (2.2.7.3)$$

The translation function  $Z$  score (TFZ) for the last component placed in a molecular-replacement search is often a good indicator of the confidence that can be placed in the solution. If TFZ is greater than 8 and there is no translational pseudo-symmetry, the solution is almost always correct.

Detailed descriptions of the background and proper application of molecular-replacement approaches are presented in Chapter 2.3 of *International Tables for Crystallography* Volume B and Chapters 13.2 and 13.3 of the present volume.

## 2.2.8. Quality indicators for refinement

The last step of a structure determination is the refinement of the model against the observed data. Refinement is in principle a mathematical operation that is applied in order to minimize the discrepancy between the observed structure-factor amplitudes  $|F_{\text{obs}}|$  and the calculated ones  $|F_{\text{calc}}|$ .

**Crystallographic  $R$  factor,  $R$ .** The crystallographic  $R$  factor  $R$  is defined as the fractional disagreement between the set of observed structure-factor amplitudes and amplitudes calculated from the structural model. Of course, observed and calculated reflection sets need to be on the same scale.

$$R = \sum_{hkl} \frac{||F_{\text{obs}}| - |F_{\text{calc}}||}{\sum_{hkl} |F_{\text{obs}}|}. \quad (2.2.8.1)$$

**Free  $R$  factor,  $R_{\text{free}}$ .** The free  $R$  factor  $R_{\text{free}}$  is defined in the same way as the crystallographic  $R$  factor, but it is based on a set of reflections that have been excluded from the refinement (Brünger, 1992). The excluded set of reflections is called the *test set*, while the set of reflections used for refinement is called the *working set*. The test set can be chosen randomly or systematically, either in thin resolution shells or to account for the presence of noncrystallographic symmetry, respectively. In order to minimize the impact on the final model, the test set should be as small as possible. Typically, it contains about 5–10% of the

reflections, or at least enough reflections to keep the standard deviation of  $R_{\text{free}}$  below 1%, but there is no need to use more than 2000 reflections (Kleywegt & Brünger, 1996; Brünger, 1997). The standard deviation of  $R_{\text{free}}$  has been empirically estimated to be  $R_{\text{free}}/N^{1/2}$ , where  $N$  is the number of reflections in the test set (Brünger, 1997). Of course, there may be concerns about the impact of excluding 5–10% of reflections on the final model, but a few final cycles of refinement against the recombined full data set should allay them.

**Correlation coefficients  $\text{CC}(F_{\text{obs}}, F_{\text{calc}})$  and  $\text{CC}(I_{\text{obs}}, I_{\text{calc}})$ .**  $\text{CC}(F_{\text{obs}}, F_{\text{calc}})$  and  $\text{CC}(I_{\text{obs}}, I_{\text{calc}})$  are Pearson linear correlation coefficients [see equation (2.2.2.13)] between observed and model-based calculated structure-factor amplitudes or intensities, respectively, that find use from time to time. One advantage of the use of a correlation coefficient instead of an  $R$  factor is that it avoids the problem of scaling the two sets of numbers relative to each other.

## 2.2.9. Quality indicators for the refined model

In MX, the observable-to-parameter ratio is mostly unfavourable. Therefore, structure refinements are carried out with boundary conditions, constraints and restraints. Constraints reduce the number of parameters which need to be refined, while restraints provide additional information to the refinement procedure that increases the number of observables. A refined model, therefore, has to fulfil not only the criterion that the crystallographic  $R$  factor [see equation (2.2.8.1)] is good and that the model fits well to the electron density, but also that it fits well to the restraints used in the refinement procedure.

**Real-space residual, RSR.** The real-space residual, RSR (Jones *et al.*, 1991), quantifies the discrepancies between the electron-density maps  $\rho_1$ , calculated directly from a structural model, and  $\rho_2$ , calculated from experimental data. RSR can take the form of a real-space  $R$  factor RSRF and of a real-space correlation coefficient RSCC.

$$\text{RSRF} = 2 \sum_{xyz} |\rho_1 - \rho_2| / \sum_{xyz} (\rho_1 + \rho_2). \quad (2.2.9.1)$$

The sum  $\sum_{xyz}$  runs over all grid points of the electron-density maps that are close to the model. A big advantage of RSRF is that it can be calculated on a residue-by-residue basis. It therefore gives a local picture of structure quality. It can also be used throughout model building and refinement in order to follow the improvement of the model locally on a per-residue basis.

RSCC is defined as the Pearson linear correlation coefficient [see equation (2.2.2.13)] between  $\rho_1$  and  $\rho_2$ . Everything said about RSRF above applies to RSCC as well.

**R.m.s. deviation from ideal of geometric parameter  $x$ .** The root-mean-square deviation of a set of geometric parameters  $x$  from their ideal values is defined as

$$\text{r.m.s.d}(x) = \left\{ \sum_i [x_i(\text{ideal}) - x_i(\text{observed})]^2 / N \right\}^{1/2}. \quad (2.2.9.2)$$

The sum runs over all  $N$  instances of the geometric parameter occurring in a structure. The geometric parameters  $x$  that are typically considered are bond lengths, bond angles, dihedral angles, chiral volumes, planar groups *etc.* The ideal values for proteins are typically taken from the study of Engh & Huber (1991) and for nucleic acids from Parkinson *et al.* (1996).

**Z score.** A measure of the likelihood that an individual geometric parameter is correct is given by its  $Z$  score. The  $Z$  score is defined as the distance of an individual data point of a distribution from the mean of the distribution expressed in standard