Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Separating model optimization and model validation in statistical cross-validation as applied to crystallography

Gerard J. Kleywegt

Department of Cell and Molecular Biology, Uppsala University, Biomedical Centre, Box 596, SE-751 24 Uppsala, Sweden

Correspondence e-mail: gerard@xray.bmc.uu.se

Received 16 May 2007
Accepted 9 July 2007

Statistical cross-validation has become an integral part of the model-refinement process in macromolecular crystallography. However, the test set of reflections, for which the free R value is calculated, is used both to optimize the parameterization of the structure model and to validate the model itself. This practice could introduce bias and diminish the value of $R_{\rm free}$ as an independent check of model quality. It is proposed here that by introducing a dormant hold-out set of reflections, any problems with such bias can be avoided. This procedure requires only a small modification of the standard cross-validation protocol.

Following Brünger's introduction of the free R value in the early 1990s (Brünger, 1992, 1993), statistical cross-validation has rapidly become an integral part of macromolecular structure-refinement procedures (Kleywegt & Jones, 1995, 1996, 1997, 2002; Dodson *et al.*, 1996; Kleywegt & Brünger, 1996; Brünger, 1997; Tickle *et al.*, 1998, 2000; Fabiola *et al.*, 2006; Gauch, 2006). In practice, the diffraction data are divided into a large work set, which is used for refinement and calculation of the traditional R value, $R_{\rm work}$, and a small test set (typically 5–10% of the data or ~1000–2000 reflections). The test set is not used during refinement [other than, perhaps, for $\sigma_{\rm A}$ (Read, 1990) or map calculations] and is only used to calculate a more-or-less unbiased cross-validation R value, $R_{\rm free}$. Many discussions in conferences, workshops, electronic bulletin boards and, to a lesser extent, the literature have focused on the practical aspects of cross-validation (*e.g.* the size of the test set) and on the circumstances under which $R_{\rm free}$ could be biased. Such bias could arise, for instance, as a result of inappropriate test-set selection methods or of inherent relationships between reflections owing to the presence of bulk solvent, non-crystallographic symmetry (NCS; Fabiola *et al.*, 2006), twinning or pseudo-symmetry.

Another issue that has been raised but never adequately addressed relates to the fact that in crystallographic cross-validation the same test set of reflections is used both for optimization of the model parameterization (*i.e.* finding the optimal choice of model parameters to refine) and for validation of the final model (*i.e.* assessing the reliability or predictive value of the combination of parameterization and refined parameter values). For example, in the refinement of a low-resolution structure with NCS, the behaviour of the free R value can be used to decide whether a structure is most faithfully modelled by applying NCS constraints, NCS restraints (possibly of varying degrees of tightness) or without any consideration of NCS at all. In practice, this is accomplished by determining whether the introduction of additional parameters into the model (by the removal of NCS constraints or the relaxation or removal of NCS restraints) leads to significantly lower values of $R_{\rm free}$. However, it has been argued that by doing so the test set of reflections becomes biased itself and that $R_{\rm free}$ consequently has diminished validity as an independent criterion for model validation.

This issue is not unique to crystallography (Hastie *et al.*, 2001) and can be addressed by splitting the crystallographic data into three, rather than two, sets (Fig. 1). The work set is used for refinement as usual (*i.e.* to determine the best parameter values for a given model

parameterization), the test set is used to optimize the model parameterization and a third set, the hold-out set, is used only for *a posteriori* validation of the final model. This hold-out set is dormant and 'should be kept in a vault' (Hastie *et al.*, 2001): it is used neither for refinement nor for the calculation of $R_{\rm free}$. Instead, the set is used only once to calculate an additional R value, purely for model-validation purposes, at the very end of the complete model-building and refinement process. Since the terms $R_{\rm work}$ and $R_{\rm free}$ suggest a metaphor related to the diurnal activities of man, I propose accordingly to call the validation-only R value, calculated with the dormant hold-out set, $R_{\rm sleep}$.

The use of a hold-out set can easily be implemented in the framework of existing software by using standard methods for selecting a test set (*e.g.* those implemented in the program *DATAMAN*; Kleywegt & Jones, 1996) to select the hold-out set before the start of the structure-refinement process and by storing the flagged reflections in a separate file (the 'vault'). The remaining reflections are then considered as the complete data set, which can be divided into a work and a test set as usual. Once the normal model-optimization and refinement process is finished, the test set is incorporated into the work set and the model refined in one last round against the combined (work and test) reflections. Finally, the hold-out set is re-introduced (and treated as a test set) and the value of $R_{\rm sleep}$ calculated (as '$R_{\rm free}$' of the total data set) without any further refinement. Macros for *DATAMAN* that implement the separation and the rejoining steps for *CNS*-style (Brünger *et al.*, 1998) reflection files are available (http://xray.bmc.uu.se/gerard/supmat/rsleep).

The use of a hold-out set may become particularly relevant when automatic model-building and refinement methods can be used to produce a (possibly large) number of alternative models using different programs and scenarios (*e.g.* with respect to the treatment of NCS, temperature factors, restraint weights, modelling of disorder *etc.* and possibly different partitionings of work and test reflections). Large-scale computations could conceivably produce dozens or even hundreds of alternative models and $R_{\rm sleep}$ could be used to identify the best one amongst them.

Obviously, a disadvantage of using a hold-out set is a further reduction in the size of the work set, which will lead to poorer maps and a lower data-to-parameter ratio. However, this disadvantage is partly offset by the fact that the test set can be used in the final refinement round, provided that no major changes (that would require cross-validation to assess) are made to the model once the test set has been merged with the work set. To obtain reliable statistics, one would like the test and hold-out sets to contain 1000–2000 reflections each. Analysis of the holdings of the August 2006 release of the Uppsala Electron-Density Server (EDS; Kleywegt *et al.*, 2004) shows that 69% of crystal structures were refined against data sets that contain more than 20 000 reflections (and 90% have at least 10 000 reflections). This means that for most refinements test and hold-out sets of 1000 reflections each can be used that together still account for less than 10% of the total number of reflections (yielding an estimated relative error of ∼3% on both $R_{\rm free}$ and $R_{\rm sleep}$; Brünger, 1997). With smaller data sets, either smaller test and hold-out sets can be selected (at the expense of larger errors in the values of $R_{\rm free}$ and $R_{\rm sleep}$) or the model-optimization stage can be concluded earlier. This means that the test set is only used initially to assess the best way to model NCS relationships, temperature factors *etc.*; once that has been performed, the test set can be merged with the work data and the refinement and rebuilding continued (provided the NCS and B-factor models are not subsequently changed). This approach has the disadvantage that after the reincorporation of the test set, the refinement proceeds blindly and the calculation of $R_{\rm sleep}$ may then come as an unpleasant surprise. In addition, $\sigma_{\rm A}$ estimates would have to be calculated using the work data. For these reasons, it may be preferable to resort to using smaller test and hold-out sets when dealing with small data sets.

As a practical recommendation, I propose to create test and hold-out sets of the same size, namely 2000 reflections or 5% of the total number of reflections each, whichever number is smaller. Obviously, the same precautions that are taken to avoid or reduce dependencies between work-set and test-set reflections should also be applied when selecting the hold-out set.

As for the expected magnitude of the various R values, in general $R_{\rm work}$ would be less than the last recorded value of $R_{\rm free}$, which in turn would be smaller than $R_{\rm sleep}$. In general, the lower the value of $R_{\rm sleep}$ is, the more reliable one expects the model to be. Further, the smaller the differences between the three R values, the smaller the degree of over-fitting (*i.e.* inclusion of parameters whose refinement is unwarranted) or of 'under-modelling' (*i.e.* omission of parameters whose refinement is warranted) is expected to be.

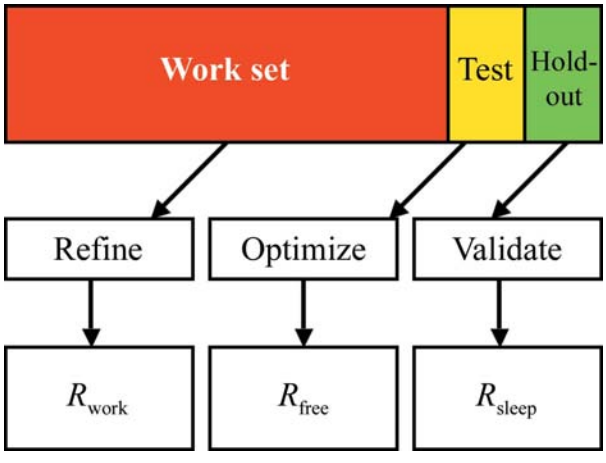

Figure 1
By splitting the crystallographic data into three sets, model refinement, model optimization and model validation can be separated.

References

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Brünger, A. T. (1993). *Acta Cryst.* D**49**, 24–36.
Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Dodson, E., Kleywegt, G. J. & Wilson, K. S. (1996). *Acta Cryst.* D**52**, 228–234.
Fabiola, F., Korostelev, A. & Chapman, M. S. (2006). *Acta Cryst.* D**62**, 227–238.
Gauch, H. G. (2006). *Am. Sci.* **94**, 133–141.
Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*, pp. 193–223. New York: Springer.
Kleywegt, G. J. & Brünger, A. T. (1996). *Structure*, **4**, 897–904.
Kleywegt, G. J., Harris, M. R., Zou, J.-Y., Taylor, T. C., Wahlby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
Kleywegt, G. J. & Jones, T. A. (1995). *Structure*, **3**, 535–540.
Kleywegt, G. J. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 826–828.
Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 208–230.
Kleywegt, G. J. & Jones, T. A. (2002). *Structure*, **10**, 465–472.
Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.
Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *Acta Cryst.* D**54**, 547–557.
Tickle, I. J., Laskowski, R. A. & Moss, D. S. (2000). *Acta Cryst.* D**56**, 442–450.