Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Bias in cross-validated free *R* factors: mitigation of the effects of non-crystallographic symmetry

Felcy Fabiola,[a] Andrei Korostelev[a] and Michael S. Chapman[a,b]*

[a]Kasha Laboratory of Molecular Biophysics, Florida State University, Tallahassee, FL 32306-4381, USA, and [b]Department of Chemistry and Biochemistry, Florida State University, Tallahassee, FL 32306-4381, USA

Correspondence e-mail: chapman@sb.fsu.edu

Received 23 August 2005
Accepted 5 December 2005

Current methods of free *R* factor cross-validation assume that the structure factors of the test and working sets are independent of one another. This assumption is only an approximation when the modeled structure occupies anything less than the full asymmetric unit. Through progressive elimination of reflections from the working set, starting with those expected to be most correlated to the test set, small biases in free *R* can be measured, presumably because of oversampling of the Fourier transform owing to bulk solvent in the crystal. This level of bias may be of little practical importance, but it rises to significant levels with increasing non-crystallographic symmetry owing to wider correlations between structure factors than hitherto appreciated. In the presence of 15-fold non-crystallographic symmetry, with resolutions commonly attainable in macromolecular crystallography, it may not be possible to calculate an unbiased free *R* factor. Methods are developed for the calculation of reduced-bias free *R* factors through elimination of the strongest correlations between test and working sets. With 180-fold non-crystallographic symmetry they may not be an accurate indicator of absolute quality, but they do yield the correct optimal weighting for stereochemical restraints.

1. Introduction

Protein structure determination involves the fitting of atomic models to the observed diffraction data. The best possible approximation to reality is sought by fitting the data as closely as possible without overfitting to the experimental errors of the data. This is a challenge because most macromolecular structure determinations are not highly overdetermined, *i.e.* there is not an excess of data points over model parameters/degrees of freedom. Refinement is made possible by the addition of subsidiary restraints (or equivalent energy terms) based on an *a priori* understanding of stereochemical geometry (Waser, 1963; Hendrickson, 1985; Brünger *et al.*, 1987). Optimization of a cost function similar to that shown leads to a stereochemically reasonable solution,

$$E_{\rm cost} = w_a \sum_h [|F_{\rm o}(h)| - |F_{\rm c}(h)|]^2 + E_{\rm chemical}, \qquad (1)$$

where it is assumed that the calculated structure amplitudes ($F_{\rm c}$) for each reflection (h) are already scaled to the observed amplitudes ($F_{\rm o}$). The first term is the crystallographic residual weighted by the parameter w_a and the second term ($E_{\rm chemical}$) approximates a molecular-mechanics energy function (Brünger *et al.*, 1987). The challenge now becomes choice of weighting (w_a) between diffraction and stereochemical restraints to yield a well fitted but not biased or overfitted structure.

The quality of the fit of the model to the diffraction data is usually assessed with R factors,

$$R = \sum_h |F_o(h)| - |F_c(h)| / \sum_h |F_o(h)|. \quad (2)$$

Conventional R factors calculated with the same reflections used for refinement are of little use in assessing weighting, because an increase in w_a or de-emphasis of $E_{chemical}$ improves the agreement between F_o and F_c whether the model is really improved or merely overfitted. Within reason, arbitrarily low conventional R factors can be obtained when there are too many model parameters, stereochemical restraints are too loose or there are insufficient experimental data. Thus, conventional R factors need not be objective indicators of model quality (Brünger, 1992, 1993).

The concept of an unbiased indicator of model quality, R^{free}, was introduced by Brünger (1992, 1993). It has had substantial impact both in the setting of refinement weights and in objective assessment of model quality. The diffraction data is (randomly) divided into a large 'working' set (90–95%) and a small 'test' set (5–10%). The working set is used for refinement, from which the test data are excluded so that they can be used for an unbiased cross-validation of the model. R^{free} is defined in the same way as the conventional R, but is calculated from just the test set. R^{free} is therefore free of the artifactual lowering caused by overfitting, so the difference ΔR between R^{free} and R is used as an indicator of overfitting. High values of R^{free} can reveal errors in the structure determination (Brünger, 1992, 1993). R^{free} can be used to evaluate refinement strategies and set optimal weighting for restraints (Brünger, 1992, 1993). It is now the most ubiquitous overall indicator of the progress of refinement and a model's final quality (Kleywegt & Brünger, 1996).

A critical assumption for cross-validation is that the reflections of the test set are not correlated to those of the working set. If they were, optimization of the model to working-set reflections would inherently improve the agreement with test-set reflections whether the model was being improved or overfitted and R^{free} would not be unbiased. There have been widely held suspicions that reflections may be interdependent in ways that affect R^{free} calculation, especially in the presence of non-crystallographic symmetry (NCS; Kleywegt & Brünger, 1996). Here, we demonstrate that no macromolecular structure is completely immune from these effects and show that they can have an important impact with medium- to high-order NCS. The interdependencies of test- and working-set reflections can be mostly eliminated with medium-order NCS, but can only be partially reduced with high-order NCS. The proposed remedies will be sufficient to calculate an essentially unbiased R^{free} with up to medium-order NCS. They will allow weighting strategies to be determined for high-order NCS, but not unbiased cross-validation. To understand the proposed remedies and their limitations, it is helpful to consider the origins of the interdependencies of reflections.

The issues can be understood in terms of sampling and information content of the Fourier transform (Bricogne, 1996). Should the atomic model precisely fill the crystallographic asymmetric unit, then there is exactly the right amount of information in the amplitudes and phases of the Fourier transform to reconstruct its image to a given resolution. In practice, this is never true. Macromolecules are surrounded by disordered solvent which do not contribute as much to the Fourier coefficients of the diffraction pattern. The asymmetric unit is of larger volume than the ordered model, so there are more structure factors (vectors) than needed to specify the molecular image. In fact this overdeterminacy is exploited in phase refinement by solvent flattening (Wang, 1985), because the redundancy in information allows us to recalculate improved estimates of the phases. The overdeterminacy is greater in the presence of non-crystallographic symmetry (NCS), which can also be exploited in phase refinement (Rossmann, 1990; Tong & Rossmann, 1995; Chapman, 1998; Chapman *et al.*, 1998). A larger molecular copy number means a larger unit cell and therefore a proportionately larger number of structure factors, but the information content of the Fourier transform does not change: when the structure of one equivalent is defined, so are all of its symmetry copies. Both disordered bulk solvent and NCS lead to a crystallographic asymmetric unit that is larger than the unique part of the structure and more Fourier coefficients in the diffraction than are needed to image the structure. The structure factors must be interdependent, with the prospect of correlated test and working reflections.

What do we know of the magnitude and nature of these potential correlations? The power of phase-refinement methods (Arnold & Rossmann, 1986) suggests that twofold NCS and typical bulk solvent would have effects of similar magnitude and that NCS could have greater impact in proportion to the NCS redundancy. It is well known that real-space (molecular) symmetries lead to corresponding rotational symmetries and therefore interdependencies in the diffraction pattern (Rossmann & Blow, 1963; Rossmann, 1964; Tong & Rossmann, 1995). It is also well known that in oversampled transforms, Fourier coefficients are most correlated to their immediate neighbors in reciprocal space (Rossmann & Blow, 1962; Vellieux & Read, 1997; Chapman, 1998; Chapman *et al.*, 1998). Once a structure is determined, it is possible to calculate the extent of the interdependencies exactly: it is given by the Fourier transform of the molecular envelope. However, we are more interested in what can be done before the structure is known. Rossmann & Blow (1962) derived an interference (autocorrelation) function G, which describes the magnitude of interdependence between a pair of reflections using the crude assumption that the molecular envelope is spherical,

$$G(H) = \{3[\sin(2\pi Hr) - 2\pi Hr\cos(2\pi Hr)]\}/(2\pi Hr)^3, \quad (3)$$

where H is the length of the reciprocal-lattice vector between the two reflections and r is the radius of the spherical molecular envelope. The G function is the three-dimensional Fourier transform of a solid sphere and is qualitatively similar to the one-dimensional transform of a step function (the familiar 'sinch' function). It is centrosymmetric with a

maximum at $H = 0$ with damped oscillations as $|H|$ increases. In the context of phase refinement, it has been argued that beyond the first node correlations become smaller and those that are correlated and anticorrelated will approximately cancel, and that to a first approximation one need only consider pairs of reflections within the first node, *i.e.* for $|Hr| < 0.72$ (Rossmann & Blow, 1962; Rossmann *et al.*, 1992).

The Fourier theory indicates that there are interdependencies between all reflections in the oversampled crystallographic transforms. The interdependencies are strongest between reflections that are close in reciprocal space or close after rotation according to the NCS. In reducing the bias in $R^{\rm free}$, a compromise must be struck between eliminating reflections from the working set that are significantly correlated to test-set reflections and retention of sufficient reflections in the working set for a robust refinement. The extent of interdependencies depends on the NCS redundancy and the molecular volume relative to that of the crystallographic asymmetric unit. The minimal fractions of data that must be retained in the working set for refinement and in the test set for reliable statistics depend on the total number of reflections and therefore upon the resolution. Thus, the appropriate balance between eliminating bias and retaining refinement overdeterminacy depends on the extent of NCS symmetry and the resolution of data available.

Others have been aware of the potential complications of NCS for cross-validation (Kleywegt & Brünger, 1996) and there have been informal discussions at scientific meetings since the mid-1990s. Proposed remedies were embodied in two software packages. *DATAMAN* (Kleywegt & Jones, 1996) and *SFTOOLS* (Collaborative Computational Project, Number 4, 1994) allow for test reflections to be selected in thin resolution shells rather than randomly, so that NCS-equivalent reflections are either all part of the test or working sets. *DATAMAN* also has an option to exclude from the working set *G*-function neighbors of randomly chosen test reflections. The results presented here indicate that the potential for bias has been underestimated and that greater efforts should be made to mitigate the impact of reflection interdependence where there is NCS.

Here, representative structures are used to quantify the bias in cross-validation arising from interdependent structure factors. It is shown that the previously suggested strategies are insufficient to ensure unbiased $R^{\rm free}$ with even moderate NCS. New approaches for the complete elimination of bias are successful for low-order NCS, but are impractical for high-order NCS (as in viral structures). However, partial elimination of bias by the proposed methods is sufficient to obtain optimal refinement weights even in pernicious cases.

2. Methods

2.1. Basic algorithms for partitioning into working and test sets

Test-set reflections were either selected randomly or by thin resolution shells (see below). Reflections that neighbored test reflections in reciprocal space were eliminated from the working set on the basis that they could be strongly correlated. The criterion for rejection was $|Hr| < X_{\rm cut}$. Use of $X_{\rm cut} = 0.72$ would correspond to elimination of all reflections within the first node of the *G* function (3). *G* functions have previously been used in several contexts and have been defined variously in terms of subunit (Rossmann & Blow, 1962) or assembly (Tong & Rossmann, 1995) radius. Here, we are interested in the degree to which the reciprocal lattice oversamples the ordered part of the structure and so choose a radius corresponding to the contents of the primitive reduced unit cell by multiplying the protomer radius of gyration by the cube root of the number of NCS and crystallographic equivalents. (NCS redundancy will be considered later.) Our radius corresponds to an 'assembly' expanded by non-crystallographic and crystallographic symmetry when relevant. (For viruses, it includes the protein capsid but not disordered nucleic acid.) The use of the *G* functions embodies other crude approximations, including that the molecular structure is spherical. Thus, $X_{\rm cut}$ was treated as an empirically adjustable parameter rather than an *a priori* constant.

Reflections related by non-crystallographic symmetry were treated in two ways. In the first, for each randomly chosen reflection the rotationally symmetric parts of reciprocal space were identified and reflections from adjacent lattice points were eliminated from the working set. (Note that a rotation in Cartesian space of a lattice vector generally gives off-lattice symmetry equivalents.) For high-order NCS, this strategy can eliminate most of the working set. Thus, our second treatment was to select test reflections in resolution shells as in *DATAMAN* (Kleywegt & Jones, 1996) or *SFTOOLS* (Collaborative Computational Project, Number 4, 1994), except that different methods were used for determining the width and spacing of shells (see below).

It proved to be critical that neighbor exclusion could be combined with the treatment of NCS. In our first treatment of NCS, this meant that for each test reflection that for *N*-fold symmetry there would be *N* spheres in reciprocal space from which all other reflections would be excluded from both test and working sets. For the second thin-shell treatment, it meant that each test shell was flanked either side by margin shells within which reflections were excluded from test and working sets. The width of the margins was set according to the same $|Hr| < X_{\rm cut}$ criterion that would eliminate the strongest correlations. Note that prior thin-shell implementations have not excluded neighboring reflections.

Prior thin-shell implementations have used constant-volume shell widths and spacings that are evenly distributed in $1/d^2$ space. This was impractical with neighbor exclusion, because the shells of excluded correlated reflections are constant thickness in $1/d$ space. The combination leads to a working set with completeness that decreases, sometimes sharply, at high resolution and therefore to inferior refinements. The solution is to space the shells evenly in $1/d$ and to have them of thickness proportional to $1/d$ such that the test and working sets maintain similar (representative) population distributions as a function of resolution.

Table 1
Test structures.

PDB code	Structure	Space group	NCS	Solvent + disordered DNA/RNA content (%)	Resolution (Å)	Reported R^{free}	Reported R	Reference
1aky	Adenylate kinase	$P1$	N/A	46	1.63	N/A	19.4	Abele & Schulz (1995)
4sli	Trans-sialidase	$P1$	N/A	40	1.8	22.9	18.9	Luo *et al.* (1999)
1m15	Arginine kinase	$P2_12_12_1$	N/A	39	1.2	12.25	10.82	Yousef *et al.* (2002)
1brm	Aspartate β-semialdehyde dehydrogenase	$P2_12_12$	Threefold	50	2.5	29.4	22.5	Hadfield *et al.* (1999)
1k5m	Rhinovirus	$I222$	15-fold	67	2.7	N/A	21.6	Ding *et al.* (2002)
1lp3	Adeno-associated virus-2	$P1$	180-fold	62	3.0	34.2	33.8	Xie *et al.* (2002)

Table 2
Bias in R^{free} from correlated structure factors, as revealed by test refinements with experimental data.

$R^{free}_{T_1}$ and $R^{free}_{T_2}$ were calculated for many non-overlapping test sets, each with its own refinement. Tabulated are the mean and standard deviations of these multiple refinements. Z scores show the significance of the differences between free R factors calculated with (T_2) and without (T_1) removal of NCS-related and neighboring reflections. Test-set reflections were selected randomly. T_1 and T_2 were set aside simultaneously for each refinement batch so that the statistics could be compared for identical models (hence single values of R^{work} and $R^{free}_{T_3}$). The estimated errors for $R^{free}_{T_1}$ and $R^{free}_{T_2}$ are large because they reflect the sampling variation between small test sets that were different for each run. They are therefore a measure of the total error in full cross-validation and not the precision of an individual R^{free} calculation by conventional methods. Full cross-validation was needed here to establish the statistical significance of differences between $R^{free}_{T_1}$ and $R^{free}_{T_2}$, but would not be required for routine structure determination.

PDB code	Resolution (Å)	Conventional free R $R^{free}_{T_1}$	Neighbors/ NCS omitted $R^{free}_{T_2}$	R^{work}	No. of test sets	Approximate test-set size	Approximate working-set size	Z score
1aky	2.0	21.1 ± 1.4	21.5 ± 1.3	16.8 ± 0.1	68	200	11000	1.9
1m15	2.0	30.3 ± 2.2	31.0 ± 2.4	24.9 ± 0.2	100	200	18000	2.0
1lp3	3.0	33.9 ± 1.2	37.0 ± 2.4	33.6 ± 0.4	10	150	200000	2.1

These algorithms were implemented through modifications to and accessory programs for the macromolecular refinement program *CNS* (Brünger *et al.*, 1998).

2.2. Modifications to test the proposed methods of cross-validation

The following methods were used to test the performance of modified R^{free} statistics and would not be used in future applications. A total of three test sets were selected, all of which were excluded from the refinement working set. T_3 contained a fixed set of reflections chosen randomly by the conventional algorithms (Brünger, 1992) right at the start and served as a 'control'. The corresponding $R^{free}_{T_3}$ was used to ensure that test refinements were of commensurate quality when experiments with different working sets could have affected refinement. T_2 was the test set selected by the new algorithms. NCS equivalents and ($|Hr| < X_{cut}$) neighbors of each T_2 reflection were excluded. T_1 was matched in selection criteria to T_2, but its neighbors were not excluded from the working set. Thus, comparisons of $R^{free}_{T_2}$ and $R^{free}_{T_1}$ were used to evaluate the modified methods of cross-validation.

2.3. Test-set sizes and selection criteria

Brünger (1992) recommended that test sets contain 5–10% of the reflections. This is not possible here as exclusion of NCS equivalents and neighbors can decimate the working set. T_1 and T_2 were typically 1%. To compare small differences in free R factors, the refinements were repeated with different T_1/T_2 sets, so that more precise mean and variance values could be calculated, often completing 'full' cross-validation (Brünger, 1992). (Now that the new statistics have been validated, full cross-validation will not be needed in future applications.)

Experiments to determine the appropriate X_{cut} could affect $R^{free}_{T_2}$ by changing not only the amount of bias from correlated working-set reflections, but also indirectly by changing the number of reflections remaining for the working set and hence the quality of the refinement. In evaluating bias in R^{free}, such changes to the working-set size had to be avoided. This was accomplished by determining the number of working reflections at the largest X^{max}_{cut}, then for refinements with smaller X_{cut}, randomly eliminating reflections until the same working-set size was achieved.

2.4. Test structures and refinement protocols

Tests were performed with structures and diffraction data available from the Protein Data Bank (Berman *et al.*, 2000). Structures were chosen to exemplify a range of resolutions, crystallographic and non-crystallographic symmetries, including one-, three-, 15- and 180-fold (Table 1). The database structures were perturbed to obtain a 'crude' model to serve as a starting point for test refinements. The perturbation was accomplished using simulated-annealing torsion-angle dynamics refinement at 500 K with stereochemical restraints only. Refinements used simulated-annealing torsion-angle dynamics. Unless otherwise stated, weights w_a for the diffraction terms (*c.f.* stereochemical terms) were determined using the automatic method of *CNS* (Brünger *et al.*, 1998) that balance the overall contributions to the gradients. Refine-

Table 3
Bias in R^{free} from correlated reflections as revealed by test refinements with simulated data.

Statistics were averaged over 100 refinements with different non-overlapping T_1 and T_2 test sets, but the same random seed for initial dynamics trajectories.

		Conventionally calculated R factors				NCS and neighboring reflections omitted				
PDB code	Resolution (Å)	$R^{free}_{T_1}$	R	$R^{free}_{T_3}$	R.m.s.d. (Å)	$R^{free}_{T_2}$	R	$R^{free}_{T_3}$	R.m.s.d. (Å)	Z score
4sli	1.8	7.6 ± 1.4	6.3 ± 0.5	7.5 ± 0.7	0.408	7.9 ± 1.1	6.3 ± 0.5	7.5 ± 0.4	0.408	1.6
1m15	2.0	7.2 ± 0.6	6.4 ± 0.3	7.7 ± 0.5	0.125	8.3 ± 1.4	6.4 ± 0.3	7.7 ± 0.5	0.125	7.2

ments were performed both against real diffraction data and simulated data calculated from the database structures. The latter enabled comparisons of free R factors with coordinate deviations from known target structures. For the simulated diffraction data sets, a Gaussian distributed set of errors was applied randomly to the simulated structure amplitudes, averaging 5% for protein structures and 10% for virus structures, roughly equivalent to 10 and 20% errors on intensity, respectively.

3. Results

3.1. Correlated structure factors bias R^{free}

The lowering of R^{free} through correlations between test- and working-set reflections is demonstrated by the results in Table 2 with real data and Table 3 with simulated data. With real data, the effect is modest (∼0.5%) when, without NCS, oversampling arises only from solvent content and an X_{cut} of 0.72 is used. It is larger (∼3%) with high-order NCS. Even if the lowering of R^{free} is sometimes numerically small, one-tailed Student t-tests (Spiegel, 1975) show them always to be statistically significant at the 95% level ($Z > 1.7$). The statistics (R, $R^{free}_{T_3}$) indicate that the lowering of R^{free} is not an artifact of differing refinements, as confirmed with the refinements using simulated data (Table 3), where the deviations from the known target structure indicate that the refinements have been effectively identical.

For a protein structure with typical solvent content, elimination of neighbors within the first node of the G function excludes about 20 neighbors of each test reflection. Even with a minimally sized test set (Chen *et al.*, 1999), this is tractable only if there is little or no NCS. In a worst-case scenario, one might imagine that fivefold NCS could eliminate the working set completely! Actually, some working-set reflections remain, because the neighboring regions of randomly selected test reflections overlap, but working sets would be too small for refinement. It is more practical to use thin resolution shells of test reflections. It is still necessary to eliminate neighboring reflections from the working set, but larger working sets remain. This can be understood by thinking of reciprocal space in spherical polar coordinates. With randomly chosen test-set reflections, a (three-dimensional) volume of reciprocal space is excluded around each test reflection. With test-set shells, the angular dimensions follow the surface of the test shell and we encounter only other test reflections. In terms of reducing the working set, we need only be concerned with the radial direction: the number of eliminated neighbors is linearly proportional to the number of test reflections rather than dependent on the cube. The bottom line is that if neighboring reflections are to be excluded, random selection of test-set reflections is practical only with low-order NCS and that with high-order NCS test reflections can only be drawn from thin shells.

The origins and extent of R^{free} bias were explored by examining the dependence of $R^{free}_{T_2}$ on X_{cut}, the parameter that controls the extent of the neighborhood of reflections excluded. Of interest was the impact of X_{cut} upon free R-factor

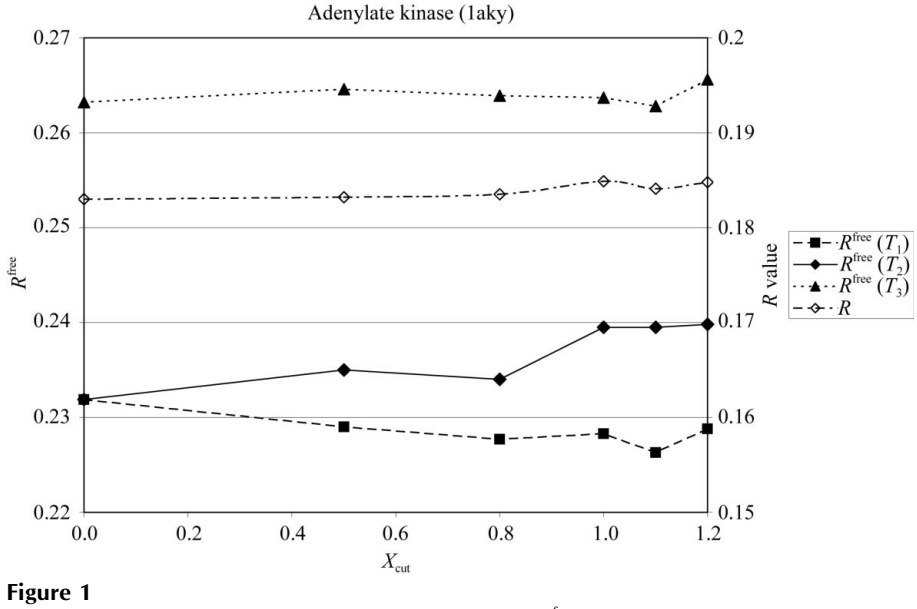

Figure 1
Extent of structure-factor correlations leading to biased R^{free} in the absence of NCS. Reflections surrounding each randomly selected test-set reflection were eliminated if $|Hr| < X_{cut}$, and $R^{free}_{T_2}$ and R are plotted for separate refinements with different X_{cut}. Conventionally calculated $R^{free}_{T_1}$ and $R^{free}_{T_2}$ are also shown. $R^{free}_{T_2}$ reaches an asymptote when (presumably) X_{cut} is large enough to eliminate most correlations between test and working sets. For low-order NCS, the point at which this is achieved is not too dissimilar from the first node of the G interference function at 0.73 (Rossmann & Blow, 1962). The test-set and working-set sizes were ∼250 and ∼14 000 reflections, respectively.

statistics and not the indirect effects through the dependence of the working-set size and hence quality of refinement upon $X_{\rm cut}$. To keep refinement quality substantially the same for different $X_{\rm cut}$, reflections from the working set were eliminated at random until the working set was of the same size as that of the highest $X_{\rm cut}$ examined. Thus, as $X_{\rm cut}$ was varied, the distribution but not the number of working-set reflections changed. All refinements in the $X_{\rm cut}$ surveys were equally degraded by the artificial restriction of the working set to a small proportion of the observed reflections and their statistics cannot be compared with unencumbered refinements. Furthermore, with the small working sets, the variance between refinements with different random test-set selections was higher, the errors of R factors higher and the differences between R factors less significant. However, important trends can still be seen.

Consider first adenylate kinase, an example without NCS. As $X_{\rm cut}$ increases, $R^{\rm work}$ and the conventionally calculated $R^{\rm free}_{T_1}$ and $R^{\rm free}_{T_3}$ do not change significantly, indicating similar quality refinements, despite which $R^{\rm free}_{T_2}$ increases to a new asymptote at $X_{\rm cut}$ = 1.0 that is about 1% higher (Fig. 1). Thus, the change in $R^{\rm free}_{T_2}$ reflects a change in the statistic, not in the quality of the atomic model. The new $R^{\rm free}_{T_2}$ appears to be reached when a large enough neighborhood of test reflections has been eliminated from the working set, so that there is effectively no more bias. The $X_{\rm cut}$ of 1.0 approximates the first node in the G interference function (Rossmann & Blow, 1962). The higher asymptote should be the more truly cross-validated statistic. Adenylate kinase is typical of examples without NCS in that the effects are very modest, *i.e.* the bias in a conventionally calculated $R^{\rm free}$ is small.

Examples with NCS show qualitatively similar effects, but over a larger range. To explore wide ranges in $X_{\rm cut}$, test-set reflections must now be selected from thin resolution shells rather than randomly. Aspartate semialdehyde dehydrogenase provided an example of threefold NCS. The range of $X_{\rm cut}$ explored allowed a constant-size working set of 70% of all observations. The qualities of refinements at different $X_{\rm cut}$ are substantially the same, as indicated by a conventional $R^{\rm work}$ that is near constant for refinements with both real and simulated data and by coordinate deviations from the target structure that show no consistent trend (Fig. 2). Of particular interest, $R^{T_2}_{\rm free}$ rises with $X_{\rm cut}$, completing a transition to a higher level by an $X_{\rm cut}$ of $\sim$1.0 (Fig. 2). Again, the change in $R^{\rm free}_{T_2}$ is a result of a change in the statistic, not the quality of the underlying model. $R_{\rm free}$ appears to be biased slightly until the new asymptote is reached when a large enough margin of test-set neighbors has been eliminated from the working set.

The dependence of $^{\rm free}_{T_2}$ upon $X_{\rm cut}$ in the presence of high-order NCS is more difficult to interpret owing to complications discussed later, but several clear observations can be made. The first is that with real experimental data at resolutions of 2.7 to 3.0 Å, $R^{\rm free}_{T_2}$ increases sharply as soon as neighboring reflections are omitted (Figs. 3 and 4). In the $X_{\rm cut}$ = 0 limit (no excluded neighbors), $R^{\rm free}$ and $R^{\rm work}$ nearly converge upon one another. In this limiting case, $R^{\rm free}_{T_2}$ calculated by the modified methods is no different from the usual $R^{\rm free}$ and the convergence of $R^{\rm free}$ with $R^{\rm work}$ indicates that $R^{\rm free}$ values calculated by conventional methods are nearly as biased as $R^{\rm work}$ (Ding *et al.*, 2002). The message here is that with high-order NCS, the web of reflection interdependencies is such that an $R^{\rm free}$ calculated by conventional methods is not a cross-validated metric.

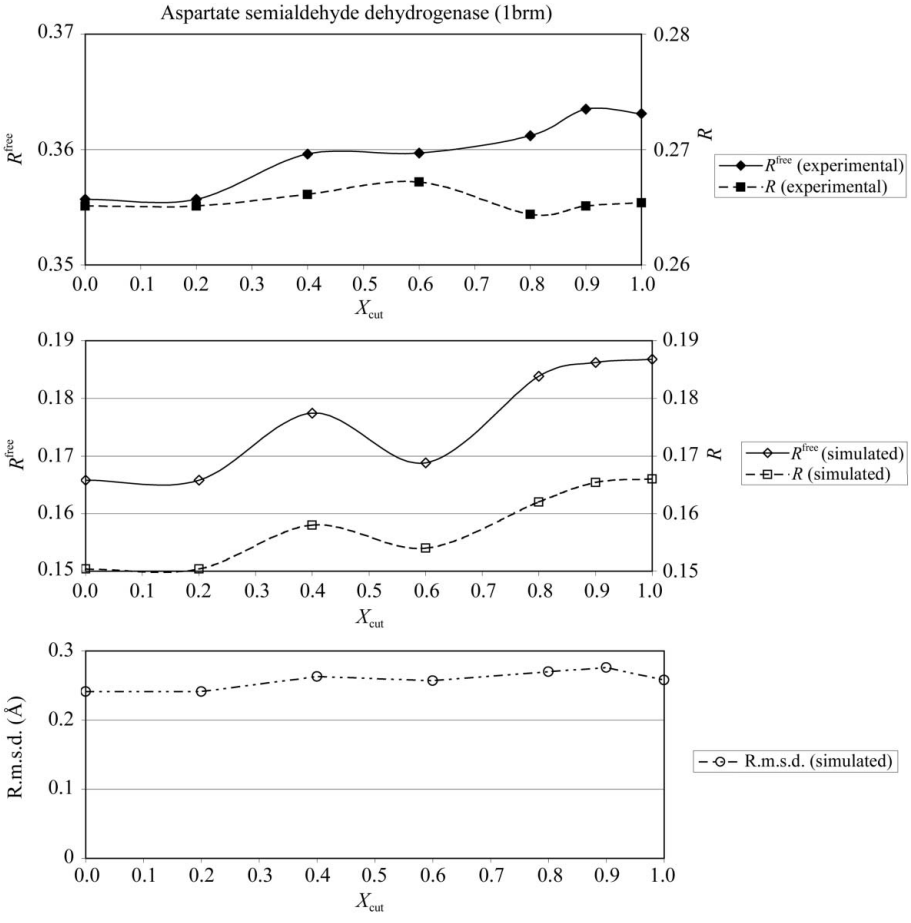

Figure 2
Extent of structure-factor correlations leading to biased $R^{\rm free}$ with low-order NCS. Reflections surrounding each thin-shell test-set reflection were eliminated if $|Hr| < X_{\rm cut}$; $R^{\rm free}_{T_2}$ and R are plotted for separate refinements with different $X_{\rm cut}$. They are shown for both experimental and simulated data for a 2.5 Å resolution structure of aspartate semialdehyde dehydrogenase that has threefold NCS (Hadfield *et al.*, 1999). For the experimental data, test and working sets sizes were 1058 and 25 310, while for the simulations they were 1321 and 27 551, respectively.

The difference between $R^{\rm free}_{T_2}$ calculated in the low and high $X_{\rm cut}$ limits provides some indication of possible bias/overfitting in the conventional statistic. The ∼6% difference is an upper bound: the bias is likely to be lower in real refinements. To allow exploration of $X_{\rm cut} > 2.0$, the working sets were restricted to just 22 and 16% of the observations for rhinovirus and AAV, respectively. Improved refinements and less overfitting would be expected with fuller use of the data sets. Furthermore, in these high-order NCS examples, there is evidence of progressively degraded refinements at high $X_{\rm cut}$. Tests with simulated data provide clues to these problems.

Within simulated data, overfitting can be essentially eliminated under favorable conditions: high resolution (2.0 Å), 100% complete data and errors in structure amplitudes of 10%. Under these conditions $R^{\rm free}$ tracks $R^{\rm work}$ (Figs. 3 and 4) with neighbor elimination at modest $X_{\rm cut}$. At high values of $X_{\rm cut}$ (>2 for rhinovirus; >3 for AAV) the refinements break down, as indicated by increasing R, $R^{\rm free}_{T_2}$ and deviation from the target coordinates (which are known for simulations). To support exploration of high $X_{\rm cut}$, working sets contained only 15% of the rhinovirus reflections or 10% for AAV reflections. The overall number of reflections remains unchanged with $X_{\rm cut}$, but changes in their distribution are unavoidable, leading to dire consequences with high-order NCS. As $X_{\rm cut}$ increases, wider margins about each test-set shell are excluded and reflections remaining in the working set are more confined to narrow shells. They become progressively more highly correlated with each other by NCS and G-function interdependencies.

The problems are compounded for the real (experimental) data sets for large unit cells with high NCS. Data sets are often lower resolution and often less complete than for more 'typical' protein structures. To allow testing of $X_{\rm cut} = 2.0$, the rhinovirus working set was 115 000 reflections (22% measured values). At low $X_{\rm cut}$ the working-set reflections are a near-random selection (excluded only from thin test-set resolution shells). At high $X_{\rm cut}$, thick shells are excluded from the working set, the combination of test shells and excluded neighbors. Working-set reflections sample only thin shells between. At $X_{\rm cut} = 2$, each working-set shell falls within a single G-function node (equivalent to an $X_{\rm cut}$ of 0.6), *i.e.* working-set reflections are significantly correlated with each other. Estimating an average twofold interference-function redundancy and 15-fold NCS, the 115 000 reflections are equivalent to 3800 independent data points, far fewer than required to robustly refine 6500 non-solvent atoms. With 180-fold NCS, the effects in AAV are even more severe (compare Fig. 3 and Fig. 4).

With high-order NCS the web of reflection interdependencies is stronger and larger $X_{\rm cut}$ might be required to effectively eliminate bias. AAV, with 180-fold NCS, required an $X_{\rm cut}$ of 1.8, compared with 1.0–1.2 for lower NCS examples. Establishing the bias-free asymptote requires exploration of higher $X_{\rm cut}$, with exponentially increasing numbers of excluded neighbor reflections. With high-order NCS, this leaves too few independent experimental observations at ∼3 Å resolution to avoid overfitting. It is unlikely actual

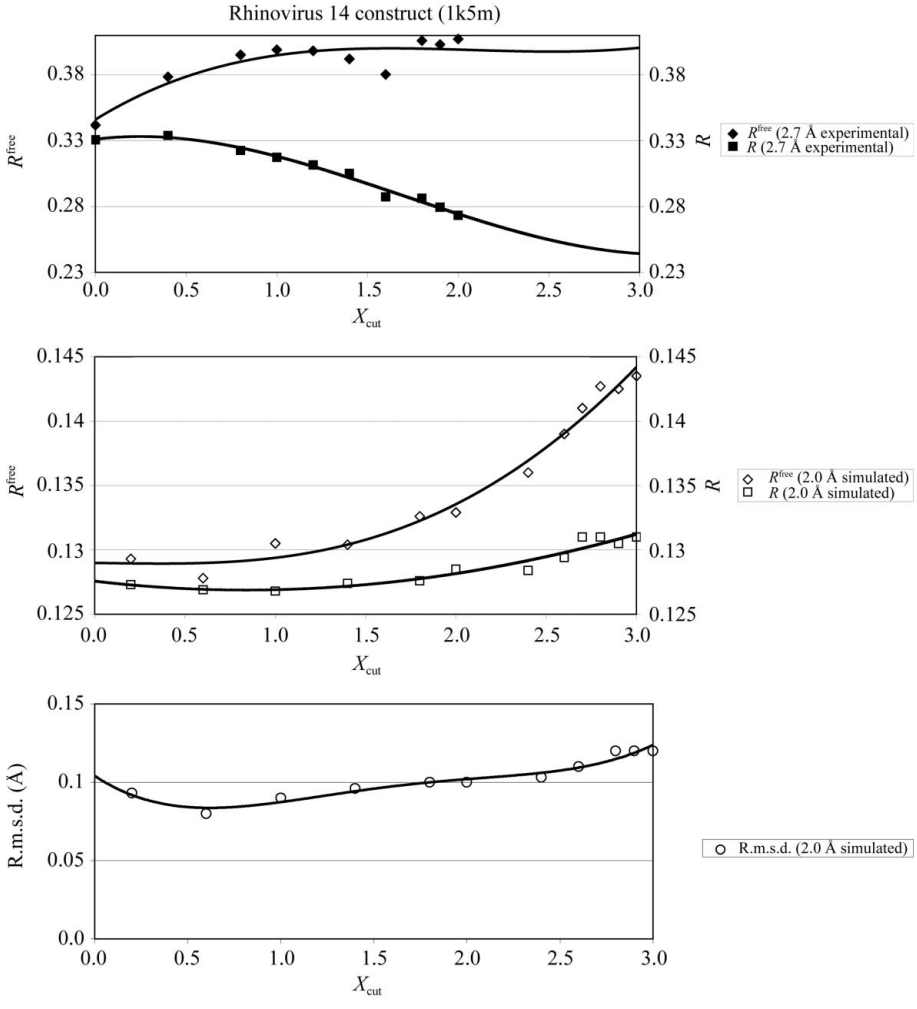

Figure 3
Extent of structure-factor bias for $R^{\rm free}$ calculation with moderately high NCS. As in Fig. 1, $R^{\rm free}_{T_2}$ is shown as a function of the $X_{\rm cut}$ used to eliminate test-shell neighbors from the working set prior to refinement. $R^{\rm free}_{T_2}$ for the 2.7 Å data reaches a plateau at the first G-function node (0.73). Comparison of this $R^{\rm free}_{T_2}$ with that at $X_{\rm cut} = 0$ (a conventional $R^{\rm free}$) gives an upper bound to the biasing of about 6%. (These tests are performed with smaller working sets than usual, with more potential for overfitting.) At higher $X_{\rm cut}$, $R^{\rm work}$ drops as a result of greater NCS correlations between remaining working-set reflections leading to refinement underdeterminacy (see text). With higher completeness and resolution, these effects can be delayed to higher $X_{\rm cut}$ for simulated data, but also the potential for overfitting and hence bias in $R^{\rm free}_{T_2}$ is lower. For the experimental data, test-set and working-set sizes were 1259 and 113 687, respectively, while for the simulations they were 1571 and 189 422, respectively.

refinements will be performed, like our diagnostic tests, using only a small proportion of the experimental data.

$R^{\rm free}_{T_2}$ calculated with a smaller (arbitrarily chosen) $X_{\rm cut}$ may be partially biased, but would support the use of larger working sets in refinement. The question is whether for pernicious high-NCS cases a partially biased $R^{T_2}_{\rm free}$ is a useful statistic. An important application of $R^{\rm free}$ is the assessment of refinement strategies and appropriate model freedom (strength of restraints, parameterization of solvent and disorder *etc.*) by searching for the lowest $R^{\rm free}$ as a function of w_a (Brünger, 1992, 1997). If, as indicated above, conventionally calculated $R^{\rm free}$ is as biased as $R^{\rm work}$ with high NCS, it would not be a good metric for w_a optimization, as confirmed below. It is also demonstrated below that the modified $R^{\rm free}_{T_2}$, calculated with modest $X_{\rm cut}$, is a reduced-bias indicator that can reveal overfitting and that can be used for w_a optimization.

Consider first the behavior of a conventionally calculated $R^{\rm free}$ as a function of w_a in the presence of high NCS. For both rhinovirus (Fig. 5) and adeno-associated virus (Fig. 6), $R^{\rm work}$ decreases monotonically with w_a, as expected with increasing weight on diffraction data relative to stereochemical terms. The expected behavior for an unbiased $R^{\rm free}$ is that it should decrease similarly until the stereochemical weight is insufficient to stop overfitting, the minimum occurring at the optimal w_a (Brünger, 1992, 1997). However, with the 15-fold NCS of rhinovirus (or 180-fold NCS of AAV), $R^{\rm free}$ decreases monotonically. This is true whether test sets are selected randomly or in the thin resolution shells previously suggested as a remedy. With $w_a = 10^8$, the r.m.s. deviations from ideal are 0.8 Å for bond lengths in rhinovirus and 30° for bond angles for rhinovirus, and 0.4 Å and 17°, respectively, for AAV. Clearly, these would be unacceptable, but the previously proposed methods of $R^{\rm free}$ calculation fail to indicate the appropriate value of w_a to use with high-order NCS.

Now consider $R^{\rm free}_{T_2}$. In contrast to the monotonic decrease in conventional $R^{\rm free}$, $R^{\rm free}_{T_2}$ shows a minimum with respect to w_a providing that neighbors of the eight test-set shells are also excluded from the working set. $X_{\rm cut}$ values of 2.0 and 2.6 for rhinovirus and AAV, respectively, are in the asymptotic region of $R^{\rm free}$ *versus* $X_{\rm cut}$, indicating unbiased $R^{\rm free}$ (Figs. 3 and 4). These $X_{\rm cut}$ values give clear minima in $R^{\rm free}_{T_2}$, with an optimal w_a of about 10^5 for both refinements. For rhinovirus, the optimal w_a is $\sim 1.6 \times 10^5$ (Fig. 5), yielding r.m.s. deviations of $\sim$0.003 Å and 0.9° from ideal bond lengths and angles, respectively. Such a high $X_{\rm cut}$ of 2.0 leaves only 24% of the reflections for the working set. An $X_{\rm cut}$ of 1.0, just short of the asymptote, yields an optimal w_a of $\sim 6.3 \times 10^5$ (Fig. 5). An $X_{\rm cut}$ of 0.3, far before the asymptote, but with a measurable $\Delta R = R^{\rm free}_{T_2} - R$, yields an optimal w_a of $\sim 1.0 \times 10^6$ (Fig. 5). Refinements with these lower $X_{\rm cut}$ values yield weights within one log unit and similar stereochemical statistics to $X_{\rm cut} = 2.0$. $X_{\rm cut}$ values of 1.0 and 0.3 leave 61 and 90% of the observations, respectively, for refinement, although for these comparative trials reflections were deleted randomly to the 24% level needed for $X_{\rm cut} = 2.0$.

A similar picture emerges from AAV. $X_{\rm cut} = 2.6$ appears excessive, as $X_{\rm cut} = 1.4$ generates a minimum that is nearly identical at w_a of $\sim 1.0 \times 10^5$ (Fig. 6). A minimum in $R^{\rm free}_{T_2}$ is just observable with $X_{\rm cut} = 0.4$ at the same w_a. Refinement with this weight yields r.m.s. deviations of $\sim$0.003 Å and 0.8° from ideal bond lengths and angles, respectively. All of these refinements

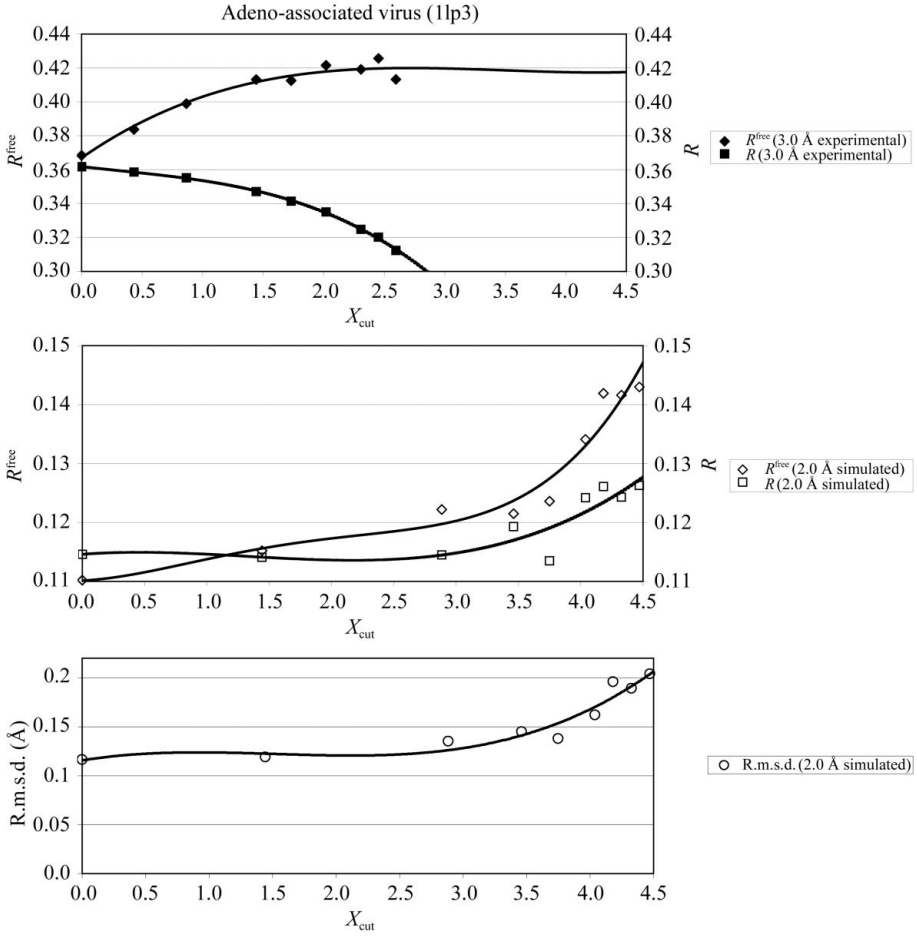

Figure 4
Challenges of $R^{\rm free}$ calculation with high-order NCS. Adeno-associated virus-2 is a pernicious example because of its 180-fold NCS (and moderate resolution). The correlations between working and test reflections appear to be strong, leading to substantial lowering of $R^{\rm free}$ at $X_{\rm cut} = 0$. There is also a decrease in $R^{\rm work}$ with $X_{\rm cut}$ owing to refinement underdeterminacy (Fig. 4). With simulated data, it can be seen that deviation from the target also worsens, but with more complete data the effects occur at higher $X_{\rm cut}$. These graphs show the challenge with high NCS of finding an $X_{\rm cut}$ where it is possible to calculate a completely unbiased $R^{\rm free}$. For the experimental data, test-set and working-set sizes were 1616 and 238 610, respectively, while for the simulations they were 1507 and 186 391, respectively.

were performed with 18% of the reflections in the working set, although $X_{\rm cut}$s of 1.4 and 0.4 could support working sets of 54 and 87%, respectively.

Better results are achieved when the full working set is used at low $X_{\rm cut}$. For rhinovirus with $X_{\rm cut} = 0.3$, comparing refinement with a 90% working set to the prior 24%, $R^{\rm free}_{T_2}$ is ∼1% lower and R is ∼1% higher (*i.e.* less overfitting). The optimum in w_a is more distinct and in even better agreement with that obtained at higher $X_{\rm cut} = 2.0$ (Fig. 5).

The messages from both rhinoviral and AAV trials are that (i) elimination of test-set neighbors allows calculation of optimal w_a, (ii) $X_{\rm cut}$ values do not have to be high enough to completely eliminate bias to obtain an adequate approximation to w_a and (iii) with appropriately chosen small test sets, it is possible to optimize w_a while refining with working sets that are nearly as large as those in common use.

4. Discussion

The potential for bias in $R^{\rm free}$ has not been widely appreciated. For the many structures with at most low-order NCS, the small bias of $R^{\rm free}$ might not be considered to be of much practical significance: the free R factor as originally implemented is a good nearly unbiased indicator of overall model quality. However, the impact on derived indicators can be greater. Thus, a 0.5% bias in $R^{\rm free}$ with a typical $\Delta R = 2.5\%$ results in an underestimate of the overfitting by 20%. The changes are certainly not large enough to consider re-refining solved structures. However, as the remedy of eliminating the closest neighbors of test reflections is simple, it should be performed in future structure determinations.

Our studies of low- and high-order NCS cases indicates qualitatively similar biases that differ in degree. For structures with high-order NCS, a potential for bias had previously been recognized (Kleywegt & Brünger, 1996), although the magnitude was not known and the responses were varied. Symmetrical virus capsid structures are being published with roughly equal frequency (i) without $R^{\rm free}$, (ii) with $R^{\rm free}$ calculated using a randomly selected test set and (iii) with $R^{\rm free}$ calculated from thin resolution shells of test reflections. For most virus structures $R^{\rm free}$ is closer to R than for protein structures at similar resolution. Many (including the corresponding author!) saw the similarity between $R^{\rm free}$ and R as an indicator of minimal overfitting with the favorable data-to-parameter ratio available with high-order NCS. Our work shows that none of the prior responses were sufficient to remove the strong biases with high-order NCS. Thus, previously calculated viral capsid '$R^{\rm free}$' are essentially self-validated not cross-validated indicators of quality, similar to a conventional R, but calculated from a subset of the data. It is unlikely that these past refinements will be repeated. Therefore, quality of the prior viral capsid refinements needs to be judged carefully by old

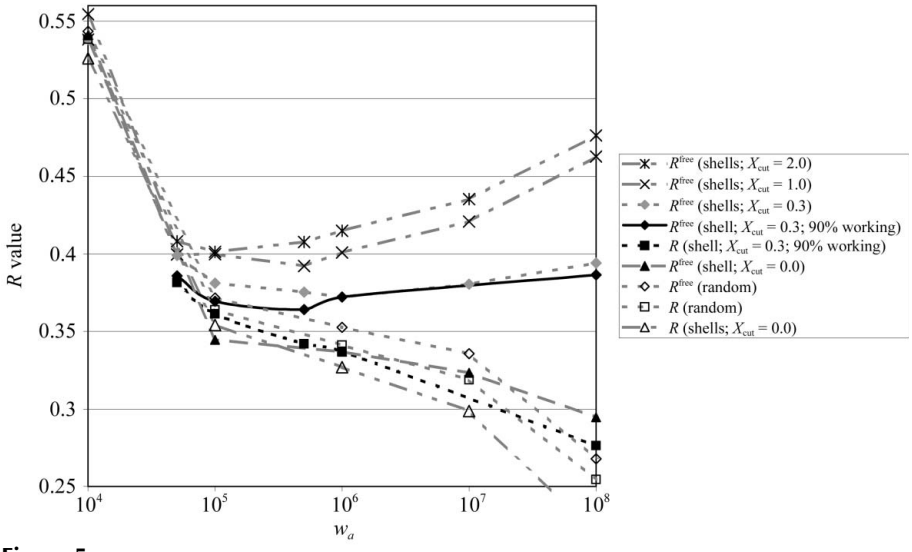

Figure 5
Optimization of w_a for refinements of the rhinovirus construct. $R^{\rm free}$ was calculated in several ways: through random selection of test reflections and with test-set resolution shells, excluding varying numbers of neighboring reflections from the working set. Refinements were performed with constant-sized working sets (here 26% of the data) that supported the largest $X_{\rm cut}$ values tested, except for those at $X_{\rm cut} = 0.3$, noted in the key as refined with 90% working sets.

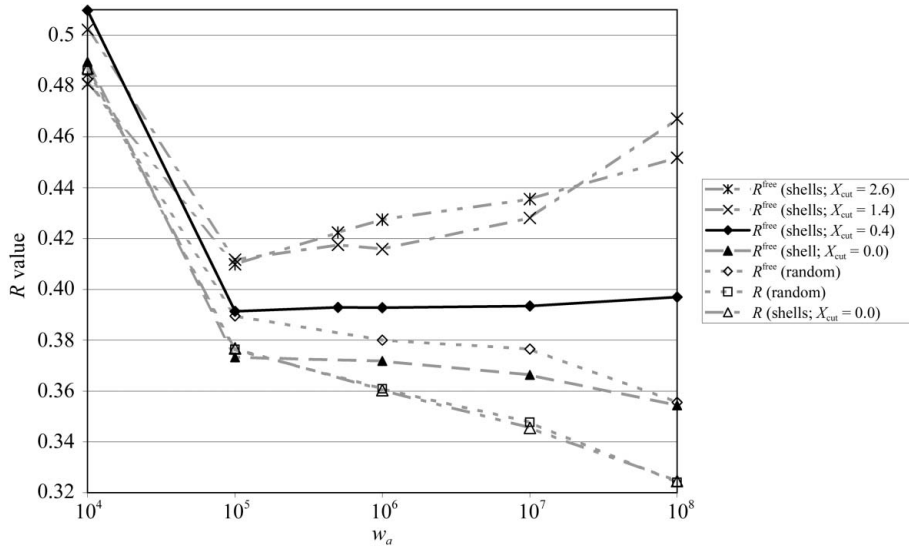

Figure 6
Optimization of w_a for refinements of adeno-associated virus. $R^{\rm free}$ was calculated in several ways: through random selection of test reflections and with test-set resolution shells, excluding varying numbers of neighboring reflections from the working set. Refinements were performed with constant-sized working sets (here 18% of the data) that supported the largest $X_{\rm cut}$ values tested.

metrics that can also indicate overfitting, such as deviations from ideal geometry (Hendrickson, 1985). It is especially important that the methods developed here of reducing bias by eliminating the strongest interdependencies be used in the future, so that even with high NCS one can be sure that the weighting schemes used do not promote overfitting.

Crude estimates of the 'typical' amount of overfitting can be made from our high NCS examples. $\Delta R = (R_{T_2}^{\rm free} - R)$ can be measured at an $X_{\rm cut}$ close to the $R_{T_2}^{\rm free}$ asymptote, but before the degradation of R at high $X_{\rm cut}$ that has been discussed above. For both the rhinovirus construct and AAV, ΔR is of the order of 4%. Several caveats have already been noted including the use of only partial working sets in these $X_{\rm cut}$ surveys and that our tests used no manual rebuilding to improve convergence, so real structures are likely to be better. However, we note that the ΔR values are similar to those expected of less symmetrical protein structures at a similar stage. Furthermore, when w_a weights were optimized with reference to $R_{T_2}^{\rm free}$, stereochemical parameters were similar to those optimized in protein structure refinements.

These observations indicate that virus structure refinements are more typical than sometimes thought. Many of the perceived advantages of NCS and an apparently favorable data-to-model ratio in refinement appear to be offset by correspondingly strong interdependencies between reflections. Before it became tractable to refine capsid structures against entire data sets (Chapman, 1995), it was usual to refine against alternating subsets of the strongest reflections (Silva & Rossmann, 1985; Arnold & Rossmann, 1988). Low R factors were sometimes obtained with high r.m.s. deviations from standard stereochemistry (0.038 Å and 4.2°, respectively, for HRV14 bond lengths and angles). These were considered justified with respect to the high map quality, low resolution and data-to-model ratios that, even with 4–25% of the reflections, appeared similar to protein structure refinements. Our results indicate that with NCS reflection interdependencies, the information content of a subset of data is much less than that of an equal number of independent reflections. Our optimizations of w_a also indicate that virus structures are not an exception to the stereochemical quality that should be expected. In retrospect, some of the early viral capsid refinements appear overfitted and the benefits of high NCS in refinement are not as great as once thought. Computational tractability is no longer a limitation. Complete data can be used for refinement and the methods presented here can be used to determine the optimal weighting. It is likely that ligand-bound and mutant capsid structures will continue to be solved with fractional data sets (Badger *et al.*, 1988). If elimination of test-set interdependencies leaves too small a working set, weights should probably be chosen for the variant structure that reproduces the r.m.s. stereochemical deviations of the parent structure where weights can be optimized by cross-validation.

We are not the first to recognize that bias in $R^{\rm free}$ might be reduced or eliminated by exclusion of test-set neighbors from the working set. This is straightforward and works well where there is at most low-order NCS. However, the proposed remedies are more complicated in the presence of high-order NCS, as revealed by the tests reported here. A larger group of test-set neighbors must be excluded than previously anticipated. In summarizing the causes and remedies, Fig. 7 might be a helpful reference. This shows refinements similar to Figs. 3 and 4, except that as $X_{\rm cut}$ is varied, the weight w_a is not automatically adjusted by *CNS* (Brünger *et al.*, 1998), but kept constant at 10^5, close to value optimized with reference to $R_{T_2}^{\rm free}$. For both rhinovirus and AAV, $R_{T_2}^{\rm free}$ starts essentially identical to R, then increases hyperbolically with $X_{\rm cut}$, reaching an asymptote when apparently enough test-set neighbors have been excluded from the working set for $R_{T_2}^{\rm free}$ to be effectively bias-free. At low $X_{\rm cut}$ R is quite stable, but at higher $X_{\rm cut}$ there is a precipitous decline. Our tests indicate that with high $X_{\rm cut}$, the reflections available for the working set come themselves from resolutions shells as thin as the test-set reflections and can be highly correlated to each other in the presence of high-order NCS. The ratio of independent data points to model parameters ceases to be sufficient for a refinement without overfitting.

It would be possible to calculate an $R_{T_1}^{\rm free}$ with an $X_{\rm cut}$ of about 1, large enough to eliminate most bias, but not large enough for the increased overfitting. With high-order NCS (even with thin resolution shell of test reflections) an $X_{\rm cut}$ of 1 requires exclusion of 40–50% of the data from the working set. Most will find this an unacceptable sacrifice of data just to be able to calculate an unbiased $R^{\rm free}$ for high-NCS assemblies.

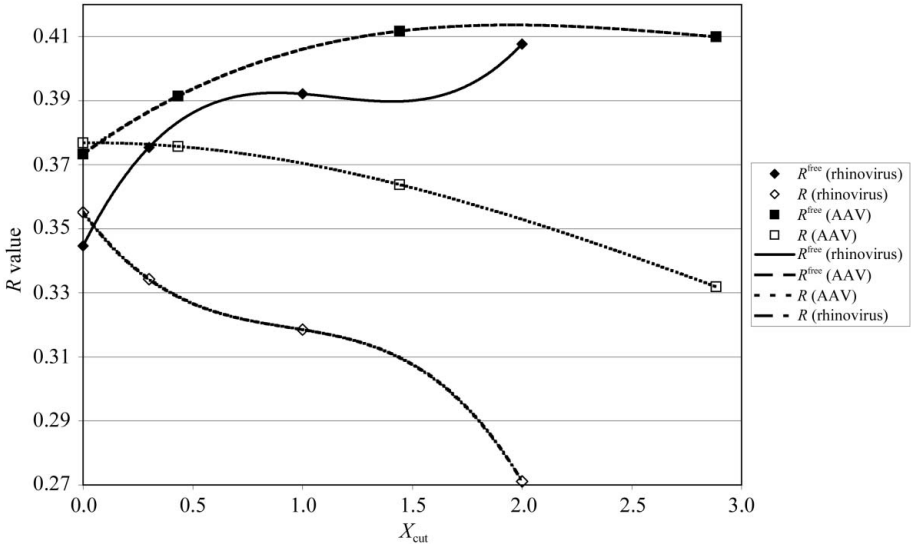

Figure 7
Refinements of the rhinovirus construct and AAV with constant weight, $w_a = 10^5$, as a function of the number of excluded neighbors surrounding each test set shell.

Fortunately, the optimization of refinement weights can be achieved with a partially unbiased $R^{\rm free}$ calculated with an $X_{\rm cut}$ of 0.3– 0.4 that allows for working sets nearly as large (∼90%) as usually used.

The weights obtained at an $X_{\rm cut}$ of about 0.35 are robust over a wide range of $X_{\rm cut}$ and give the best balance of experimental and stereochemical restraints, as judged objectively. For the rhinovirus and AAV examples, these weights are sixfold to sevenfold lower than those given by an automatic *CNS* procedure that balances the experimental and stereochemical contributions to the gradients. The *CNS* procedure would have led in each case to an $R^{\rm free}$ that would have been about 0.01 higher and a working R about 0.01 lower, *i.e.* would have been overfitted by an additional 0.02.

The assembly radius r used in calculating $X_{\rm cut}$ and the number of neighbor reflections eliminated has been treated as an empirical parameter. The *G*-function approximation uses the unrealistic approximation of a spherical assembly. The appropriate radius may also depend on non-model parameters. If there is pseudosymmetry and the low-resolution unit cell is effectively smaller, then a smaller radius might be needed to eliminate the longer range structure-factor correlations at lower resolution. Thus, the setting of $X_{\rm cut}$ is often going to be inexact. It is therefore difficult to pre-define parameters for unbiased $R^{\rm free}$ calculation that will apply in all situations. However, in the optimization of w_a refinement weights, perhaps the most important application of $R^{\rm free}_{T_2}$, only partial bias-reduction is required, which is not sensitive to the exact $X_{\rm cut}$, and a value of about 0.35 should usually work.

Cross-validation serves a number of purposes in macromolecular structure determination, some of which can be seen in the new light of realising the extent of bias in the presence of high-order NCS. As discussed earlier, correlations between test-set and working-set reflections will often have negligible impact with low-order NCS, but with high-order NCS it may be practically impossible to calculate a free R factor in which all bias is eliminated and which is therefore a robust indicator of absolute quality. However, arguably a more important use of cross-validation is in choosing appropriate weighting schemes for refinement strategies that are not conducive to overfitting. It has been shown here that reduced-bias free R factors suffice for this purpose. A related use of cross-validation is in deriving the error distributions used in cross-validated maximum-likelihood refinements (Adams *et al.*, 1997). The results here indicate that the model errors will be underestimated in the presence of high NCS (Read, 1986). With high-order NCS, absolute estimates of error will either not be possible or will leave only a fraction of the data in the working set. The reduced-bias estimates proposed here for weight determination, calculated with $0.3 < X_{\rm cut} < 0.4$, would remove only about 1/3 of the bias (Figs. 3 and 4) and would not offer a significant improvement on current methods. Such high NCS refinements are probably best performed in real-space anyway (Chapman & Rossmann, 1996; Chapman & Blanc, 1997).

In summary, this study suggests some changes in current refinement practice. For structures with little or no NCS, test reflections should be selected randomly. Neighbors of test reflections (or symmetry-equivalent regions in reciprocal space) should be eliminated from the working set using an $X_{\rm cut}$ of about 1.1. (The exact value of $X_{\rm cut}$ required to eliminate bias depends on the molecular shape and could be determined empirically for each protein as in Fig. 1, but a value of 1.0 has been sufficient for all low-NCS examples tested.) At low NCS, the importance of taking these extra steps might be academic: $R^{\rm free}$ will likely be changed by <1%, but can now be considered to be unbiased. With increasing NCS, there comes a point (depending on resolution *etc.*) where random test-set selection followed by neighbor exclusion leaves too few reflections in the working set. At this point, the switch should be made to selection within resolution shells, again with exclusion of neighboring reflections. With medium-order NCS, elimination of the test-set/working-set interdependencies is important both in setting appropriate refinement weighting and in obtaining an unbiased $R^{\rm free}_{T_2}$, which might differ from the usual $R^{\rm free}$ by several percent. With higher order NCS, even with test reflections in shells, neighbor exclusion will usually leave too few reflections in the working set. In these cases the $X_{\rm cut}$ should be reduced to ∼0.35 to preserve ∼90% completeness in the working set, with the understanding that the bias in $R^{\rm free}$ is reduced enough for weight optimization, but it is no longer possible to fully eliminate bias for calculation of a truly cross-validated $R^{\rm free}$.

This work was supported by the National Institutes of Health through a subproject (MSC) of P01 GM64676, T. A. Cross, PI and in part by R01 GM66875 (MSC). Software for the selection of test and working sets is available under license from http://www.sb.fsu.edu/~chapman. Scripts have been written for *CNS* (Brünger *et al.*, 1998), but could be adapted for other refinement programs.

References

Abele, U. & Schulz, G. E. (1995). *Protein Sci.* **4**, 1262–1271.
Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1997). *Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.
Arnold, E. & Rossmann, M. G. (1986). *Proc. Natl Acad. Sci. USA*, **83**, 5489–5493.
Arnold, E. & Rossmann, M. G. (1988). *Acta Cryst.* A**44**, 270–282.
Badger, J., Minor, I., Kremer, M., Oliveira, M., Smith, T. J., Griffith, J. P., Guerin, D. M., Krishnaswarmy, S., Luo, M., Rossmann, M. G., McKinlay, M., Diana, G., Dutko, F. J., Fancher, M., Rueckert, R. & Heinz, B. A. (1988). *Proc. Natl Acad. Sci. USA*, **85**, 3304–3308.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bricogne, G. (1996). *International Tables for Crystallography*, Vol. *B*, edited by U. Shumueli, pp. 23–105. Dordrecht: Kluwer Academic Publishers.
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., Gross-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.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.
Chapman, M. S. (1995). *Acta Cryst.* A**51**, 69–80.
Chapman, M. S. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 99–108. Dordrecht: Kluwer Academic Publishers.
Chapman, M. S. & Blanc, E. (1997). *Acta Cryst.* D**53**, 203–206.
Chapman, M. S., Blanc, E., Johnson, J. E., McKenna, R., Munshi, S., Rossmann, M. G. & Tsao, J. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 433–442. Dordrecht: Kluwer Academic Publishers.
Chapman, M. S. & Rossmann, M. G. (1996). *Acta Cryst.* D**52**, 129–142.
Chen, Z., Blanc, E. & Chapman, M. S. (1999). *Acta Cryst.* D**55**, 219–224.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Ding, J., Smith, A. D., Geisler, S. C., Ma, X., Arnold, G. F. & Arnold, E. (2002). *Structure*, **10**, 999–1011.
Hadfield, A., Kryger, G., Ouyang, J., Petsko, G. A., Ringe, D. & Viola, R. (1999). *J. Mol. Biol.* **289**, 991–1002.
Hendrickson, W. H. (1985). *Methods Enzymol.* **115**, 252–270.
Kleywegt, G. J. & Brünger, A. T. (1996). *Structure*, **4**, 897–904.
Kleywegt, G. J. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 826–828.
Luo, Y., Li, S. C., Li, Y. T. & Luo, M. (1999). *J. Mol. Biol.* **285**, 323–332.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Rossmann, M. G. (1964). *Acta Cryst.* **17**, 1474–1475.
Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 73–82.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.
Rossmann, M. G. & Blow, D. M. (1963). *Acta Cryst.* **16**, 39–45.
Rossmann, M. G., McKenna, R., Tong, L., Xia, D., Dai, J.-B., Wu, H., Choi, H.-K. & Lynch, R. E. (1992). *J. Appl. Cryst.* **25**, 166–180.
Silva, A. M. & Rossmann, M. G. (1985). *Acta Cryst.* B**41**, 147–157.
Spiegel, M. R. (1975). *Probability and Statistics.* New York: McGraw–Hill.
Tong, L. & Rossmann, M. G. (1995). *Acta Cryst.* D**51**, 347–353.
Vellieux, F. M. D. & Read, R. J. (1997). *Methods Enzymol.* **277**, 18–53.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
Waser, J. (1963). *Acta Cryst.* **16**, 1091–1094.
Xie, Q., Bu, W., Bhatia, S., Hare, J., Somasundaram, T., Azzi, A. & Chapman, M. S. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 10405–10410.
Yousef, M. S., Fabiola, F., Gattis, J. L., Somasundaram, T. & Chapman, M. S. (2002). *Acta Cryst.* D**58**, 2009–2017.