On the validation of crystallographic symmetry and the quality of structures

Jimin Wang*

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520

Received 4 September 2014; Revised 20 October 2014; Accepted 24 October 2014
DOI: 10.1002/pro.2595
Published online 29 October 2014 protein.science.org

Abstract: In 2008, Zwart and colleagues observed that the fraction of the structures deposited in the PDB alleged to have “pseudosymmetry” or “special noncrystallographic symmetry” (NCS) was about 6%, and that this percentage was rising annually. A few years later, Poon and colleagues found that 2% of all the crystal structures in the PDB belonged to higher symmetry space groups than those assigned to them. Here, I report an analysis of the X-ray diffraction data deposited for this class of structures, which shows that most of the “pseudosymmetry” and “special NCS” that has been reported is in fact true crystallographic symmetry (CS). This distinction is important because the credibility of crystal structures depends heavily on quality control statistics such as $R_{\text{free}}$ that are unreliable when they are computed incorrectly, which they often are when CS is misidentified as “special NCS” or “pseudosymmetry”. When mistakes of this kind are made, artificially low values of $R_{\text{free}}$ can give unjustified confidence in the accuracy of the reported structures.

Keywords: pseudosymmetry; special noncrystallographic symmetry; $R_{\text{free}}$ values; reliability of reliability index; symmetry R-factor; merging R-factor; measurement R-factor; crystallographic R-factor; symmetry downshifting; local minimum; refinement statistics

Introduction

Phase retrieval is a classic inverse problem because the phases associated with the Bragg reflections in a diffraction pattern, which are not directly measured, are fully encoded in their amplitudes. Phases can be retrieved from the diffraction patterns of the crystals of any continuous, three-dimensional object by computation as long as: (i) diffraction amplitudes are accurately measured with a signal-to-noise ratio of over 20, and (ii) there is 2.6-fold oversampling in reciprocal space (for review and additional information, see Palatinus and online supporting materials, OSM, Supporting Information Fig. S1). This level of oversampling occurs naturally in macromolecular crystals that have an unordered solvent content of 65%, or greater. A survey of the protein data bank (PDB) carried out in April 2014 demonstrated that the average solvent content for all macromolecular crystals is about 50%, which is not much less than 65%, and also showed that about 40% of all entries are reported to have at least twofold noncrystallographic symmetry (NCS), which should also contribute to oversampling. Thus, it is possible that the majority of the structures in the PDB could have been solved starting with measured amplitudes only. Even though none of the structures in the database was in fact determined this way, complete a priori phase retrieval using NCS averaging has been successfully carried out for a number of test cases. Thus, it would be interesting to know if NCS really is as prevalent in macromolecular crystals as the data in the PDB might lead one to believe, and this is why the work described below was undertaken.

In 2008, using coordinates available in the PDB repository, Zwart et al. calculated the root-mean-square deviations of the Cα coordinates of molecules...
supposedly related by NCS from average structures that were obtained from the same data assuming that the symmetry operations relating them are actually CS. They concluded that 6% of all structures in the PDB ought to be considered as having “special NCS” or “pseudosymmetry,” and noted that the percentage of such structures reported each year was increasing.\textsuperscript{8} In addition, 15 years earlier, Wang and Janin had shown that the NCS axes being reported in the PDB tended to be either parallel or orthogonal to a unit cell axis.\textsuperscript{9} Taken together, these observations suggested that at least some of the claims of NCS that have been made might be spurious.

Shortly after the study of Zwart and colleagues was published, another report appeared indicating that there is indeed a widespread problem with NCS and CS in the PDB. Poon and colleagues found that up to 2% of all the crystal structures in the PDB belonged to higher symmetry space groups than the ones to which they were assigned by their authors.\textsuperscript{10} Their study focused on whether one could detect “missed symmetry” by analyzing the coordinates of already determined structures, trace any problems so identified back to the X-ray diffraction data, and then correct them. However, given the large differences that exist between the calculated and observed amplitudes of structure factors for crystal structures in the PDB,\textsuperscript{11} this kind of coordinate-based analysis is likely to underestimate the extent of the special NCS problem because when CS is treated as NCS, experimental errors are certain to make coordinates diverge that should be identical by reason of symmetry.

Here, I report the result of an analysis of the X-ray diffraction data in the PDB,\textsuperscript{12} which confirms that there is reason for concern about the way crystal symmetry is being treated by macromolecular structural biologists. Unlike the earlier studies of Zwart \textit{et al.} and Poon \textit{et al.},\textsuperscript{8,10} the analysis discussed below focused on how widespread the phenomenon of missed symmetry might be, and the deleterious effects it can have on structure quality. It is clear that at least in some cases, the symmetry in question was not missed. It was instead treated as NCS for the purposes of refinement in the belief that it would lead to an improvement in the quality of the structures that emerged, which is not the case. It is also clear that the problems created by missed symmetry cannot be addressed using techniques based on quality control statistics such as $R_{\text{free}}$, the crossvalidation (CV) statistic introduced in 1992 on which so much reliance is placed today.\textsuperscript{13}

Results

**Special noncrystallographic symmetry is often crystallographic symmetry**

NCS $R$-factors ($R_{\text{symm}}$) were computed from observed intensities for all low-symmetry space group entries in the PDB database using XPREP from the Shelx suite, and the orientations of NCS axes were determined using MolRep from CCP4.\textsuperscript{12,14–16} It was found that in most cases where NCS has been reported, and $R_{\text{symm}}$ between NCS-related reflections is (less than 15% and) often smaller than expected given the value quoted for \(<I/\sigma(I)>\) (Table I, Fig. 1, OSM, Supporting Information Table S1–S3). Thus, the data do not support the hypothesis that the NCS reported for most of these crystals is anything other than CS.

Included among the many entries that appear to have been incorrectly identified as having NCS are 219 of the 3003 P1 entries in the database (Table I). The mean value of the averaged \(<I/\sigma(I)>\) for these entries is 12.1 ± 6.2, suggesting that the data for most of them was quite weak. However, the mean value of $R_{\text{symm}}$ for the NCS-related reflections of these entries is 0.071 ± 0.032, which is smaller than that expected, given the average value of \(<R(\sigma(I))>\) for these data, which is 0.083 \(R(\sigma(I))\) is \(<\sigma(I)>/\langle I\rangle\) and can sometimes be approximated as \(<\sigma(I)>/\langle I\rangle\). Thus poor data quality does not explain why so many of these crystals were assigned to the space group P1.

A strong hint about what has been going on is to be found in the fact that the data associated with a noticeable fraction of the P1 entries in the PDB, plus all other low-symmetry entries, and even some P2\textsubscript{1}2\textsubscript{1}2\textsubscript{1} entries (and beyond, see below), have $R_{\text{symm}}$ values of zero when the NCS it is claimed they

| Table I. Number of Structures with Misassigned Space Groups$^a$ |
|------------------|-----------------|---------|
| Space group      | Entries$^b$    | Total   | Percentage (%) |
| P1               | 219            | 3003    | 7.3            |
| P2               | 14             | 152     | 9.2            |
| P2\text{1}      | 563$^c$        | 12,345  | 4.6            |
| C2               | 187$^d$        | 7705    | 2.4            |
| P3               | 10             | 547     | 1.8            |
| P3\text{1}      | 15             | 521     | 2.9            |
| P3\text{2}      | 29             | 494     | 5.9            |
| P2\text{1},2,2\text{1} | 133$^e$        | 17,965  | 2.2            |
| P2,2,2\text{1} from P2\text{1},2,2\text{1} | 106          | 1823    | 5.8            |
| Tetragonal or Cubic$^f$ | 7          | 2034    | 0.3            |

\(\text{a An analysis was carried out with entries released prior to mid-May 2014. Whereas this analysis focused on low-symmetry space groups, higher symmetry space groups are not immune to the problems identified here.}

\(\text{b Entries with } R_{\text{symm}} \text{ less than 0.15 in higher symmetry space groups are counted.}

\(\text{c Break-down is: 530 orthorhombic, 23 tetragonal, 9 hexagonal, and 1 cubic.}

\(\text{d Break-down is: 101 orthorhombic, 33 rhombohedral, 24 hexagonal, 15 tetragonal, and 1 cubic, and a few other choices.}

\(\text{e Break-down is: 128 tetragonal and 5 cubic.}

\(\text{f Crystals having unit cell lengths within 2.5% of those expected either for tetragonal or cubic were selected for analysis.}

\(\text{g Crystals having unit cell lengths within 2.5% of those expected for hexagonal were selected for analysis.}
display is treated as CS for purposes of data processing (Fig. 1). The only way this could happen is if those data sets were processed initially in a higher symmetry space group, and then deliberately expanded into P1 without reprocessing, which is totally unacceptable.

**The special noncrystallographic symmetry problem is widespread**

The criterion used above to assess whether special NCS is actually CS, namely that $R_{\text{symm}}$ is smaller than the value expected on the basis of the $<|I|/\sigma_I>$ value of the data, probably underestimates the extent of the special NCS problem in the database.

It is instructive in this regard to analyze the orientations of NCS axes (Supporting Information Table S1, Fig. 2). In P1, the $c^* \text{ axis}$ is the shortest unit cell vector in the reciprocal space, and it corresponds to the longest real-space unit cell vector, $c$. When an axis of NCS is almost perfectly parallel or orthogonal to the $c^* \text{ axis}$, and the peak in the self-rotation function (SRF) corresponding to this NCS is sufficiently high relative to the origin peak, it is highly probable that the axis of the alleged NCS will be perfectly aligned from one unit cell to the next throughout the entire crystal lattice. If this is so, then the NCS will certainly be CS. The NCS axes of about one-third (34.6%) of all P1 entries in the PDB have this property.

It is not easy to decide how high the SRF peak attributed to special NCS must be to make it reasonable to conclude that it is, in fact, CS. This is because CS peaks in SRF maps are often substantially weaker than origin peaks when real CS is ignored during data processing, especially when the data are weak. The extent of that reduction in strength is a function of multiplicity number of the NCS. For example, an SRF peak associated with threefold NCS is likely to be weaker than a twofold NCS peak. Nevertheless, a crude estimate of the extent of the special NCS problem can be obtained using the criterion that the height of the SRF peak (relative to the origin peak) assigned to NCS should be at least one standard deviation above the mean value within a given class. For example, the mean SRF peak height for all P1 entries is $0.45 \pm 0.25 \text{ (Fig. 2)}$, including all possible kinds of NCS. Thus, if a special NCS SRF peak is higher than 0.70, there is a high probability that it really represents a CS operation. This class of structures includes 17.5% of all P1 entries.
When this analysis was carried out on the subgroup of P1 structures for which cyclic twofold rotational NCS was reported (i.e., excluding all other types of NCS), the mean SRF peak height proved to be $0.64 \pm 0.18$. Thus, 11.3% of the data sets in this class have an SRF peak height greater than 0.82, which could be indicative of an invalid P2 to P1 conversion. When this kind SRF peak analysis is done on all the P1 entries (including ones without any NCS), the mean SRF peak height drops to $0.35 \pm 0.25$, and 18.6% of them are found to have special NCS SRF peak heights higher than 0.60. These estimates of the extent of the special NCS problems in the PDB are broadly consistent with the two studies discussed earlier. It is a significant problem (Supporting Information Table S1).

Systematic absences are as useful for determining the true crystal symmetry as $R_{\text{symm}}$ and SRF analysis even when data are very weak. Systematic absences are a property of the entire crystal lattice whereas NCS is a local property of individual asymmetric units. For example, when a structure is refined in P1 that has special P2$_1$2$_1$2$_1$ NCS, it is highly unlikely that the refined structure that emerges will predict the systematic absences characteristic of crystals that belong to the space group P2$_1$2$_1$2$_1$. Although translational NCS may exist, it can at most generate pseudosystematic absences along a single axis, which often break down rapidly with increasing resolution. The probability of encountering a crystal that has three orthogonal sets of translational NCS characterized by perfect systematic absences along all three major axes is zero for all intents and purposes. Examination of the data of the P1 space group entries containing so-called pseudo P2$_1$2$_1$2$_1$ NCS reveals that the systematic absences called for by P2$_1$2$_1$2$_1$ CS are often unmistakably clear in the data. It follows that most of the structures solved in P1 that are alleged to involve special P2$_1$2$_1$2$_1$ NCS actually belong to the space group P2$_1$2$_1$2$_1$.

On the intensity differences that distinguish crystallographic symmetry from noncrystallographic symmetry

To address whether $R_{\text{symm}}$ statistics are sensitive enough to distinguish between special NCS and CS, a simulation was carried out in which CS was converted computationally into NCS by tilting a true CS rotation axis in the unit cell of a crystal with respect to the axes of that unit cell. The homodimeric structure 3R5G, which was solved at 1.5 Å resolution in the space group P2 (a = 51.36 Å, b = 36.31 Å, c = 94.76 Å, $\beta$ = 98.76°), was used for this simulation (for this structure, the SRF peak for the two nondyad-related subunits in the original asymmetric unit is only 11% of the origin peak). The contents of the P1 unit cell content were generated from the published structure using the CS twofold operation at (1/2, 0, 0), and the Bragg reflection intensities were computed for the crystal that resulted when this new local dyad was tilted by $1^\circ$, $2^\circ$, $3^\circ$, $4^\circ$, $5^\circ$, and $10^\circ$ toward either the $a^*$ or $c^*$ axis of the unit cell. This tilting operation, of course, converted what was previously a CS dyad into an NCS dyad. The calculated intensities in P1 for ($hkl$) reflections were then compared with those of ($-h,-k,-l$) reflections, which would be identical if the NCS-so-generated was actually CS. Figure 3 shows the results. A $1^\circ$ difference between the orientation of the twofold NCS axis and a unit cell axis can result in an intensity $R_{\text{symm}}$ of over 18% in the 5-Å resolution shell, which is far greater than the merging R-factor statistics $R_{\text{merge}}$ are likely to be at such low resolution ($R_{\text{merge}} \leq 5\%$). Thus, even small deviations in axis of orientation from true rotational CS are readily detected using $R_{\text{symm}}$.  

Figure 2. Distribution of the orientations of SRF peaks in $\theta$ value in all P1 entries with only one top peak per entry. (a) The majority of orientation with $\theta$ value is parallel (180°) or orthogonal (90°) to c* axis are overlaps in the plot so that single dots may represent as many as 100 entries as the range of high SRF peaks. (b) Histogram of the distribution. Black line is the counts for $\theta$ value being parallel (180°) or orthogonal (90°) to c* axis and red line is the counts for $\theta$ value elsewhere.

PROTEINSCIENCE.ORG
On the effects of expanding data using crystallographic symmetry

The practice of expanding high symmetry data to lower symmetry for the purpose of structure refinement has become widespread in the last decade (Supporting Information Table S1). The justifications most commonly given for doing so are: (i) that $R_{\text{free}}$ values fell as a result, and (ii) that composite-omit, simulated-annealing procedures produced better maps in $P1$ or some other reduced symmetry space group.18–22 Neither is valid for reasons that will shortly be explained.

Many of the problems that arise when the true symmetry of crystals is not respected stem from the effects this practice has on data statistics. If the redundancy that results from high symmetry is not taken into account, the merging statistics obtained for a data set will not only underestimate its true quality (in which case $R_{\text{pim}}$ is more appropriate, which includes both redundancy and multiplicity) but will also lead to increases in the level of noise in the electron density maps that ultimately emerge. For example, the $\langle I/\sigma I \rangle$ value for the data associated with the 4HYO structure, which is one of the many characterized by missed symmetry, increases by ~2.8-fold when the data originally processed in $P1$ are eightfold averaged in the correct space group, $P4212$.20 Problems caused by symmetry downshifting were first addressed over a decade ago,23 and since then many other examples of this practice have appeared in the literature. Several have been provided by a series of structures that have been reported for crystals of the Thermus thermophilus RNA polymerase ($TTHRN$) complex.24–34 One of the crystal forms of this molecule has data-scaling statistics, systematic absences, and molecular-replacement solutions that are consistent with space group $P6_5$.24–26 Nevertheless, the authors of all the structures obtained with this crystal form (e.g. 1SMY) expanded data sets that had been processed in $P6_5$ into $P3_2$ for the purposes of structure refinement, and attributed the additional symmetry apparent in the data to perfect merohedral twinning with the twinning operation parallel with the 3$\bar{2}$ axis.24–26 The rationale given was that, after molecular replacement, rigid-body refinement in $P6_5$ led to a reduction in $R_{\text{free}}$ from 50.0% to only 45.3%, whereas rigid-body refinement done in $P3_2$, assuming perfect merohedral twinning, reduced the $R_{\text{free}}$ value to 38.7%, and therefore the latter refinement had to be correct.26 Similar arguments were made recently by a different group to justify the correctness of molecular replacement solutions in the presence of NCS even though such solutions they obtained thereby could not be further refined.35 In fact, the mistake made by these authors, like many others, is to give more credence to differences in $R_{\text{free}}$ values that they deserve. Evans and Murshudov have shown that in the presence of perfect merohedral twinning, a structure with an $R_{\text{free}}$ value of 29.1% can be completely wrong (OSM).36

There are problems with the 1SMY structure that emerged from the refinements mentioned above that should have been enough to call it into question.24–26 For one thing, it does not explain the pattern of systematic absences along the screw axis that are clearly evident in data (Supporting Information Fig. S2). In addition, perfect merohedral twinning of a $P3_2$ crystals along its threefold axis can only result in apparent symmetry of $P6_2$ or $P6_4$, but never $P6_5$.1 Furthermore, there are so many divalent cations included in the structure that at neutral pH, the charge per enzyme molecule would be +1,856.

Given the relatively modest resolution of this structure (2.71 Å), the addition of 9031 geometrically unconstrained water molecules (20% of all the atoms

Figure 3. Intensity change caused by tilting a CS axis to convert it an NCS axis. (A) Dyad tilts toward x axis by $1^\circ$, $2^\circ$, $3^\circ$, $4^\circ$, $5^\circ$, and $10^\circ$ (black, red, green, blue, dark brown, and brown, respectively). (B) Dyad tilts toward z axis. Vertical axis is fractional intensity change and horizontal axis is reciprocal resolution ($\AA^{-1}$).
in the model!), together with 1897 unconstrained Mg\(^{2+}\) ions is likely to have resulted in severe overfitting of the data. However, it did not produce the increase in \(R_{\text{free}}\) values it should have because the selection of reflections for the CV set was done incorrectly. The correct space group for this TTHRNP crystal form is P6\(_5\) as initially identified by the original authors (Supporting Information Table S4).\(^{20-26}\) The data from all closely crystal structures display varying degrees of mild twinning (e.g., about 4.4% in 3DXJ) with the twin axis orthogonal to the P6\(_5\) axis in such a way that the selection of reflections for CV has to be made assuming that the higher symmetry space group is P6\(_{22}\), and then expanded back to the correct, lower symmetry space group of P6\(_5\) to obtain the correct structure (Supporting Information Table S4). This was not done.\(^{24-26}\)

To further illustrate why the methods used for both CV reflection selection and refinement of the TTHRNP structures in question (i.e., data-expansion with added twinning) are invalid, a computational experiment was done in which the original CV set was kept, and a molecular replacement solution for the TTHRNP data was sought using as a starting model the structure of a smaller, completely unrelated polymerase (RB69 DNA polymerase). The molecular centers of the two models were aligned to generate a plausible model for the packing of the RB69 DNA polymerase structure in the TTHRNP unit cell, and the resulting model was then refined. During rigid-body refinement, \(R_{\text{free}}\) values fluctuated more than 10% from one refinement step to the next, and it decreased additionally by more than 7% (to about 49% for the data between 40.0 and 4.0 Å resolution) during atomic refinement. These reductions in \(R_{\text{free}}\) had nothing to do with the validity of the molecular model being used, which was unrelated to the crystals in question, but spoke only to the fact that after refinement, the symmetry of the calculated amplitudes now approximately reproduced the symmetry of the data. Thus, reductions in \(R_{\text{free}}\) values obtained by refining molecular replacement models should not be interpreted as proof that a correct solution has been found, especially when that solution cannot be refined further (or when \(R_{\text{free}}\) values are above 42.3%, see OSM). Specific examples of other problems that have been caused by the incorrect assignment of space groups may be found elsewhere.\(^{8,27}\)

### On the effects of treating crystallographic symmetry as noncrystallographic symmetry: A test case

The practical consequences of refining a structure of high symmetry in a lower symmetry space group were further explored by refining the same crystal structure under strictly identical conditions using data for that structure that had been processed in the correct space group, and the same data processed in a reduced symmetry space group. 4HYO was chosen for this computational experiment,\(^{20}\) for the two following reasons. First, the noise in the data set was relatively small because the resolution was quite high, ~1.65 Å. Second, the data clearly displays P42_2 symmetry, which means that the difference in the quality of the data processed between P1 and P42_2 is magnified because of the eightfold averaging that occurs when they are processed in P42_2 (Table II).

Three parallel structure refinement runs were carried out (see Materials and Methods). First, the structure was refined against the 4HYO data after the P1 data had been eightfold averaged to create a P42_2 data set, and reflections had been selected for the CV set in a manner that respected the P42_2 symmetry of the data (the “P42_2” run). Second, it was also refined against the original 4HYO P1 data using reflections for the CV set chosen by the original authors in P1 using the thin slice method of constant resolution shells (the “P1-P1” run).\(^{20}\) Third, it was refined against the 4HYO P1 data using the CV reflections that were used for the P42_2 run, followed by expansion into P1 (the “P1-P42_2” run). The same protein-only 4HYO model was used as the starting model for all three runs. After symmetry downshifting, the number of atoms in the asymmetric unit increased about eightfold but the number of observations increases only about sevenfold because about 12% data in P42_2 are centric but none are centric in P1 (Table III). The reduced observation-to-atom ratio in the symmetry-downshifted model will presumably lower both free and working R-factor values. As expected, after the first two passes of refinement (a total of 40 cycles, see Material and Methods), the P42_2 run had the highest \(R_{\text{free}}\) value of the three by more than 0.5% (Table III). Thus, if \(R_{\text{free}}\) values had been used at this stage to decide which refinement strategy to continue pursuing, one might have stopped the P42_2 run. However, after all three refinements converged fully, the \(R_{\text{free}}\) value for the P42_2 run was 16.4%, but it was 17.5% for the P1-P42_2 run and 16.8% for the P1-P1 run (Table III).

### Table II. Data Reprocessing Statistics of 4HYO in P42_2 Space Group

<table>
<thead>
<tr>
<th>Resolution (Å)</th>
<th>(&lt;f&gt;\sigma_{f})</th>
<th>(R_{\text{symm}})</th>
<th>(R_{\text{free}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>99–3.55</td>
<td>49.0</td>
<td>0.016</td>
<td>0.018</td>
</tr>
<tr>
<td>3.55–2.82</td>
<td>60.0</td>
<td>0.019</td>
<td>0.020</td>
</tr>
<tr>
<td>2.82–2.46</td>
<td>60.4</td>
<td>0.023</td>
<td>0.025</td>
</tr>
<tr>
<td>2.46–2.24</td>
<td>54.9</td>
<td>0.031</td>
<td>0.033</td>
</tr>
<tr>
<td>2.24–2.08</td>
<td>46.6</td>
<td>0.041</td>
<td>0.044</td>
</tr>
<tr>
<td>2.08–1.96</td>
<td>36.8</td>
<td>0.053</td>
<td>0.057</td>
</tr>
<tr>
<td>1.96–1.86</td>
<td>22.6</td>
<td>0.082</td>
<td>0.089</td>
</tr>
<tr>
<td>1.86–1.78</td>
<td>13.6</td>
<td>0.132</td>
<td>0.143</td>
</tr>
<tr>
<td>1.78–1.71</td>
<td>10.4</td>
<td>0.163</td>
<td>0.176</td>
</tr>
<tr>
<td>1.71–1.65</td>
<td>6.1</td>
<td>0.256</td>
<td>0.277</td>
</tr>
<tr>
<td>Overall</td>
<td>46.1</td>
<td>0.025</td>
<td>0.027</td>
</tr>
</tbody>
</table>
Because CS averaging reduced the levels of noise in the data used for the P42\(\times 2\) run, the quality of the electron density map that emerged from this run was superior to those obtained in the other two runs, particularly in the vicinity of both the N and C-termini of the protein. For the same reason, the Ramachandran statistics of the product of the P42\(\times 2\) refinement are superior to those obtained in the two P1 refinements, even though the starting model for all three was identical and had perfect geometry, and even though same manual editing was done to correct geometrical errors introduced by refinement (Table III).

If P42\(\times 2\) is, in fact, the right space group to use for this structure, why did the P1-P1 run produce such a “good” result as measured by \(R\)\(_{\text{free}}\)? This question can be answered by comparing the results of the P1-P1 run and the P1-P42\(\times 2\) run. Even though the converged \(R\)\(_{\text{free}}\) value for the P1-P1 run is significantly better than that of the P1-P42\(\times 2\) run, their converged working R-factor distributions are nearly identical (Fig. 4). This observation suggests that the reason the P1-P1 refinement has a lower \(R\)\(_{\text{free}}\) value is that reflections in its CV set are not fully independent of the reflections in the working set, which is not the case for the CV set used for the P1-P42\(\times 2\) run (OSM).

Symmetry downshifting can have consequences for the resulting structures that are biochemically significant. The structure 4HYO discussed above is closely related to 4HZ3.\(^{20}\) The channel blocker tetra-ammonium antimony was included in the crystallizations that led to 4HZ3, but not in the crystallizations that resulted in 4HYO. The authors reported that bound tetra-ammonium antimony could be visualized in the 4HZ3 structure, but only in the crystallizations that resulted in 4HYO. Hence, the channel blocker in question is bound in the 4HZ3 structure (Fig. 4, OSM). Phases for these maps were calculated from the final complete 4HYO model that emerged from the P42\(\times 2\) run described above (which has a \(R\)\(_{\text{work}}\) of 9.1\% at 1.65-Å resolution, Table III). Experimental errors in the data for other structures of this kind that have resolutions lower than that of 4HYO are likely to be much higher that those in 4HYO, and thus the errors introduced into the resulting structures by symmetry downshifting are likely to be greater, and every bit as likely to lead to misleading conclusions about biochemically significant issues.

**Better structures can often be obtained when CS is properly taken into account**

Two of the three P1 structures deposited prior to 2005 that are listed in Supporting Information Table III. Refinement Statistics of 4HYO in P42\(\times 2\) and in P1

<table>
<thead>
<tr>
<th></th>
<th>“P42(\times 2)”</th>
<th>“P1-P1”</th>
<th>“P1-P42(\times 2)”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution range (Å)</td>
<td>45–1.65(1.69–1.65)</td>
<td>45–1.65(1.69–1.65)</td>
<td>45–1.65(1.69–1.65)</td>
</tr>
<tr>
<td>Number of observations</td>
<td>10,760</td>
<td>75,462</td>
<td>75,584</td>
</tr>
<tr>
<td>Final refinement statistics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of atoms</td>
<td>917</td>
<td>6,483</td>
<td>6,597</td>
</tr>
<tr>
<td>Observation/atom Ratio, (R)(_{\text{O2A}})</td>
<td>11.7</td>
<td>11.6</td>
<td>11.5</td>
</tr>
<tr>
<td>(R)(_{\text{work}}) (%)</td>
<td>9.1(10.2)</td>
<td>11.4(16.7)</td>
<td>11.2(17.6)</td>
</tr>
<tr>
<td>(R)(_{\text{free}}) (%)</td>
<td>16.4(16.2)</td>
<td>16.8(21.1)</td>
<td>17.5(23.6)</td>
</tr>
<tr>
<td>(R)(<em>{\text{O2A}})/(R)(</em>{\text{work}})</td>
<td>128.6</td>
<td>101.7</td>
<td>102.3</td>
</tr>
<tr>
<td>rmsd Bond length (Å)</td>
<td>0.0087</td>
<td>0.0082</td>
<td>0.0091</td>
</tr>
<tr>
<td>Ramachandran Plotc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preferred</td>
<td>76 residues (96.2%)</td>
<td>627(95.6%)</td>
<td>602 (94.2%)</td>
</tr>
<tr>
<td>Generously allowed</td>
<td>1 residue (1.3%)</td>
<td>12 (1.8%)</td>
<td>14 (2.2%)</td>
</tr>
<tr>
<td>Disallowed</td>
<td>2 residues (2.5%)</td>
<td>17 (2.6%)</td>
<td>23 (3.6%)</td>
</tr>
<tr>
<td>Initial R-factor statistics of protein-only models with bulk solvent correctiond</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(R)(_{\text{work}}) (%)</td>
<td>20.98</td>
<td>20.53</td>
<td>20.57</td>
</tr>
<tr>
<td>(R)(_{\text{free}}) (%)</td>
<td>22.64</td>
<td>22.50</td>
<td>21.45</td>
</tr>
<tr>
<td>R-factor statistics after two passes of refinement without manual editings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(R)(_{\text{work}}) (%)</td>
<td>20.1(19.1)</td>
<td>19.7(22.7)</td>
<td>19.6(23.3)</td>
</tr>
<tr>
<td>(R)(_{\text{free}}) (%)</td>
<td>24.2(24.1)</td>
<td>23.5(25.8)</td>
<td>23.3(25.8)</td>
</tr>
</tbody>
</table>

\(a\) The highest resolution shell and their statistics are in parenthesis.

\(b\) When \(R\)\(_{\text{O2A}}\)/\(R\)\(_{\text{work}}\) is used as a measure for global goodness of fit, the model refined in the correct P42\(\times 2\) space group is about 30\% better than those in the reduced P1.

\(c\) Increased outliers in Ramachandran plot in P1 relative to P42\(\times 2\) were clearly due to reduced quality of resulting electron densities, particularly for residues at both N and C termini where all outliers are located.

\(d\) R-factor statistics for the same protein-only model with symmetry expansion (for P1) and with bulk solvent correction at the start of refinement and after the second pass of refinement each with 20 cycles. The first pass was refined without NCS restraints and the second pass with NCS restraints in P1. The lowered R-factor values for the “P1-P42\(\times 2\)” and “P1-P1” runs (particularly the “P1-P42\(\times 2\)” run) before any refinement than those for the P42\(\times 2\) run was due to reduced multiplicity weights in R-factor calculations in P1 for the pseudocentric reflections whose reflections typically have higher R-factors. Once refinement was carried out in P1, all pseudocentric reflections became acentric, when NCS axes were not constrained.
Table S1 as likely to belong to a higher symmetry space group (1JXO and 1EGW) were rerefined in that higher symmetry space group to further assess the impact that CS-to-NCS conversions might have on structural quality (all other entries in Supporting Information Table S1 were deposited after 2005). Re-examination of the data on file for 1JXO revealed that the crystal on which it is based belongs to P2$_1$; the systematic absences along the screw axis are unmistakably clear. When the P1 data were remerged into P2$_1$, the quality of the data for rerefinement in P2$_1$ improved, and with the
CV set chosen appropriately, all of the 42 residues (14% of the total residues) missing from the original 1JXO structure, which was refined in P1, plus 39 missing side chains (13%), were clearly evident in $F_{\text{obs}} - F_{\text{calc}}$ difference Fourier maps (Fig. 5, Supporting Information Tables S5 and S6). The refined structure that emerged has working and free R-factors of 17.3 and 23.6%, respectively, as compared to the original P1 structure for which the corresponding values are 20.6 and 22.9% (Supporting Information Figs. S4 and Supporting Information Tables S7 and S8). In that case, the CV set used for the refinement of the original structure in P21 was correctly selected, and thus the same set was kept for its rerefinement.

Eight other structures were rerefinied in higher symmetry space groups using properly selected CV reflection sets, and with only one exception, the structures obtained for all of them were statistically better than those on deposit in the PDB, which firmly supports the conclusion that better structures result when data are processed in the correct highest-possible symmetry space group (Supporting Information Table S2). As most of these structures were solved recently using the same or similar versions of the refinement programs used here to rerefine them, the improvements in structure quality obtained are likely to be ascribable both to the improvement in data quality that resulted from merging them in the correct, higher symmetry space group, and to the use of a properly selected CV set. This conclusion is strongly supported with the results of three parallel refinement runs for 4HYO described above.

It should be noted that comparison of the $R_{\text{free}}$ values of these rerefinied structures with those originally reported may be unfair given the fact that the reflections in the CV sets used were incorrectly selected in P1 or reduced symmetry space groups (OSM). In such cases, comparison of corresponding $R_{\text{work}}$ values might be more appropriate. For example, the newly rerefined P4212 4HYO structure has a working R-factor of 9.1% (Table III), whereas the working R-factor for the original P1 4HYO structure was 16.5% (Supporting Information Table S2). Additionally, it is noted that the 1JXO data set has $I/\sigma_I$ of 12.9 in the highest resolution shell and a value of $R_{\text{free}}$ in that shell that is less than the overall $R_{\text{free}}$ value for the structure (22.7 vs. 23.6% (Fig. 5, Supporting Information Tables S5 and S6). Similarly, the 1EGW data set also has very high $I/\sigma_I$ value (9.31) and a very low $R_{\text{free}}$ value (18.4%) in the highest resolution shell (Fig. 5, Supporting Information Tables S7 and S8).

**Figure 5.** Rerefined 1JXO structure and refinement statistics. (A) The complete structure from N to C in rainbow colors. (B) Superposition with the original 1JXO structure (gray) shows locations of four gaps in the original structures and other differences in termini. (C) R-factor distributions for both 1EGW (black) and 1JXO (red) rerefin ed structures. It should be noted that R-factors for high-resolution data were actually smaller than low-resolution data due to an excessive truncation of high-resolution data.

Examination of the data for 1EGW shows that its crystal too belongs to P21. However, while the systematic absences along the twofold screw axis are obvious upon inspection of the reported intensities, they are much less clear from $I/\sigma_I$ ratios for unknown reasons (Supporting Information Fig. S3).

Discussion

Treating CS as NCS for the purpose of refinement is invalid for two reasons. First, when higher symmetry
data are expanded into lower symmetry space groups, the coordinates of “NCS-related” molecules that ought to be identical are allowed to vary independently during refinement, thereby increasing the number of parameters that can be adjusted but not the number of independent observations. This will result in a decrease in $R_{\text{free}}$, even when the set of reflections used for CV is properly selected, which brings us to the second problem. For most of these structures, the reflections for CV were selected using invalid procedures (OSM).

In addition, symmetry downshifting can increase the noise in the electron density maps, and thereby degrade the quality of the structures obtained from them even though this (as well as multicopy refinement, see OSM) always leads to a decrease in $R_{\text{free}}$ values, as demonstrated in this study. Symmetry downshifting for structure refinement should be discouraged, or even banned since there is no valid basis for it. Moreover, the invocation of special NCS or unidentified CS not only complicates the selection of reflections for CV, it also reduces the effectiveness of composite-omit map calculations. For example, when a crystal structure that belongs to the space group $P_2_1_2_1_2_1$ is refined in $P_1$, its four identical asymmetric units are treated as though they are independent. For every omit block, therefore, there will be three identical unomit blocks that are related to the omitted block by the “special $P_2_1_2_1_2_1$ NCS.” Model-bias due to the three nonomitted blocks will completely invalidate the resulting omit map. Because this is so, composite-omit maps always look better in a symmetry-downshifted space group than in the correct space group when the maps from these two procedures are compared side-by-side.

Historical perspective
In crystallography, the term “crystal symmetry” is well defined. Operationally, a crystal must be assigned to the highest symmetry its data permit, given the level of error evident in the experimental data, that is, provided $R_{\text{symm}} < R(\sigma)$. The term “pseudosymmetry” is occasionally used to describe the symmetry evident in data sets that show peaks in their SRFs that are aligned well enough with crystal axes to be true CS, but fail the conventional $R_{\text{symm}} < R(\sigma)$ test nonetheless. Not only is it wrong to treat CS as NCS, it can have serious consequences for the quality of inferred crystal structures, as is the case for the TTHRNp structure discussed above.

Before the introduction of $R_{\text{free}}$, in a commentary on erroneous structures that shook the confidence of the structural biology community, Branden and Jones warned that (working) R-factor values of 25% or higher could be indicative of problematic crystal structures. In a test case at 1.8-Å resolution, it was shown that a structure obtained by deliberately fitting its polypeptide chain backwards through the corresponding electron density map could be refined so that its $R$-factor was less than 25%. The introduction of $R_{\text{free}}$ went a long way towards curing many of the problems Branden and Jones identified. Nevertheless, numerous cases are known where the use of $R_{\text{free}}$ did not prevent the publication of seriously defective structures.

Today, one might want to make a statement about structure quality similar to the one Branden and Jones made 2 decades ago, namely that $R_{\text{free}}$ values of 25% or higher could be indicative of structures that have serious problems.

As shown above, the $R_{\text{free}}$ values calculated for structures obtained for crystals alleged to contain NCS may have little value as measures of structure quality, especially when the CV sets used have been chosen improperly. Many other factors can also reduce the objectivity of $R_{\text{free}}$ values, a subject that will be discussed elsewhere. Nevertheless, the majority of NCS problems identified here could easily be avoided, or corrected, if necessary. All that is required is for the community to decide collectively that this problem needs to be addressed. An important point of this manuscript is to deliver a call to arms to the entire structural biology community so that the important, but entirely correctable problems identified above get resolved. Our scientific credibility is on the line.

A proposal for reform
Given the extent of the NCS problem discussed here, it would be beneficial if, along with the many other statistical tests that are already routinely done to assess structure quality, $R_{\text{symm}} < R(\sigma)$ tests were run on all the data associated with structures for which NCS is claimed at the time of structure/data deposition. Equivalent tests have been included for decades in many of the programs used for crystallographic data processing; they enable those programs to determine the symmetries of crystals automatically. Those who choose to over-ride the space group assignments provided by these programs should be warned that this is bad practice, and asked either to show cause, or to take corrective steps. If they do not, it would be appropriate that warning flags be added to the PDB entry.

Materials and Methods
All observed structure factors were retrieved in mid-April 2014 from the PDB, and converted in the scalepack format. When experimentally measured intensities ($I$) were absent and measured amplitudes ($F$) were present, the intensities were used assuming that $I = F^2$ and $\sigma I = 2F\sigma F$. Entries were rejected if the corresponding standard deviation column was missing. Intensities in each entry were then rescaled to have maximal accuracy. $R_{\text{symm}}$ was analyzed using XPREP from the Shelx suite. The
orientations of NCS axes were determined using MollRep from CCP4.16 Merging intensities to appropriate higher symmetry space groups was done using scalepack from HKL2000.51 R_{symm} obtained from prescaled P1 intensity data in this study differs slightly from that of unscaled integrated intensity data in a single-step data processing. This difference does not affect the conclusion of this study. For refinement of structures in the correct space group, the location of the special CS dyad was determined using least square superposition methods, and proper origin-shift was applied to convert special NCS back to CS whenever possible.

To minimize any differences due to different refinement procedures, different levels of investigators’ modeling skills, and any potential bias of this author, three parallel refinement runs were carried out using conditions as strictly identical as possible per suggestions of one reviewer of this manuscript. That was as follows: (i) the “P42 12” run against the 4HYO P42 12 data in the correct space group with properly selected reflections for the CV set,20 (ii) the “P1-P1” run against the P1 data with the reflections for the CV set selected in P1 using a thin slicing method of resolution shells by the original authors,20 and (iii) the “P1-P42 12” run against the P1 data with the reflections for the CV set selected in P42 12 and then expanded to P1. All three runs started with the same protein-alone model. This model was derived rerefined 4HYO model with R_{work} = 12.6% and R_{free} = 17.5% at 1.65 Å resolution at the time of original manuscript submission after all ordered water molecules and 1,6-hexanediol molecules were removed.20 An initial refinement was carried out in two passes each with 20 cycles using Refmac5.52 For the “P42 12” run, both passes had no NCS restraints applied. For the “P1-P1” and “P1-P42 12” runs, NCS restraints were applied in the second pass but not in the first pass. For the remaining 10 passes of structure refinement, including two passes of parallel Ramachandron backbone editing, exactly the same criteria were applied to locate the ordered water molecules and 1,6-hexanediol molecules or any errors in models. For example, all three runs used the same 5.0 σ or 4.5 σ, and finally 4.0 σ as cut-off criteria in peaks and holes in the residual F_{obs}-F_{calc} difference Fourier maps. The three models were then allowed to diverge only because of different numbers of peaks or holes in residual maps at a given selected cut-off. Results of this refinement are summarized in Table III and Figure 4.

Acknowledgments
The author is indebted to Dr. Peter Moore for extensively editing this manuscript to make it appropriate for the entire structural biology community rather than just those experts in crystallography. The author thanks Drs. Z. Dauter, and T. A. Jones for critical comments and useful suggestions to improve this manuscript, and Drs. W. A. Hendrickson, B. W. Matthews, T. Richmond, A. T. Brunger, W. Yang, S. Burley, H. Berman, M. G. Rossmann, A. Horwich, M. Hochstrasser, and W. Meng for discussion and comments.

References

Wang


