Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Identifying non-crystallographic symmetry in protein electron-density maps: a feature-based approach

Reetal Pai,[a]* James Sacchettini[b] and Thomas Ioerger[a]

[a]Department of Computer Science, Texas A&M University, USA, and [b]Department of Biochemistry and Biophysics, Texas A&M University, USA

Correspondence e-mail: reetalp@cs.tamu.edu

Received 18 February 2006
Accepted 17 June 2006

Non-crystallographic symmetry (NCS) averaging is a well known method for improving the quality of an electron-density map and thus aiding structure determination. Prior methods of NCS-operator determination based on estimated heavy-atom positions are prone to errors arising from inaccuracies in these coordinates or differences in the relative orientations of domains between molecules. In this paper, two real-space methods to determine NCS relationships from initial electron-density maps are presented. A brute-force method identifies matching regions in a map by local density correlation. A feature-based algorithm uses rotation-invariant features to reduce the computational time taken by the brute-force algorithm by filtering out regions that are likely to have dissimilar density patterns. This makes the feature-based algorithm faster and as accurate as the brute-force approach. Neither method requires the positions of heavy atoms or any information regarding the protein sequence. Both methods have been tested on a diverse range of experimentally phased maps and the correct NCS relationships were accurately identified for almost all of the test cases. The NCS operators obtained by the feature-based algorithm were used to perform NCS averaging and an improvement in map correlation was observed for some cases.

1. Introduction

Knowledge of the non-crystallographic symmetry (NCS) operators in an asymmetric unit (ASU) can greatly benefit the structure-determination process (Rossmann, 1972; Bailey *et al.*, 1988; Bricogne, 1974). Real-space redundancies can be exploited using density-modification techniques to average out noise arising from phase error and to increase the signal-to-noise ratio (Muirhead *et al.*, 1967; Cowan *et al.*, 1993; Cowtan & Main, 1993; Terwilliger, 2000, 2002*b*). The density-modification techniques result in higher quality electron-density maps, which aids structure determination. However, as NCS is not always exact and some regions of the subunits can be more similar than others, it is often difficult to describe the precise NCS relationships in poor/medium-quality electron-density maps. The determination of NCS operators involves the determination of one or more rotation matrices, the translation vectors and the definitions of masks around regions of the electron-density map related by the specified rotation matrices and translation vectors.

Identification of NCS operators was first used to improve estimates of phases by Rossmann & Blow (1962). Molecular-replacement techniques have been used to identify symmetry relationships (Main & Rossmann, 1966; Crowther, 1967; Kleywegt & Jones, 1994; Vellieux & Read, 1997). These techniques are capable of finding the rotational component of

the NCS operators and subsequent search methods are required to find the translational component to generate the complete transformation (McCoy *et al.*, 2005; Navaza, 1994). The presence of heavy atoms (obtained by MAD, SAD or MIR techniques) in the crystal has often been used to identify NCS. Since heavy atoms often bind to the same locations in each protein subunit, they can be used as fiducial points and the configuration of these atoms can be used to identify the symmetry operators. The initial implementation of this methodology (Lu, 1999) was a slow process and required N^5 comparisons (N being the number of heavy-atom sites). A more efficient algorithm for heavy-atom matching was implemented in *SOLVE* (Terwilliger, 2002*a*). This implementation considered interatomic distances as constraints in matching heavy atoms, leading to a reduction in the overall run time. This technique suffers from several drawbacks. A minimum of three heavy-atom positions in each subunit is required to accurately identify the NCS operators. The process of heavy-atom matching can also be sensitive to errors in the heavy-atom coordinates, which can introduce errors into the NCS operators. This is especially true in the early stages of structure determination when phases are not very accurate.

An alternative approach to NCS-operator determination involves looking for similarities between regions in the electron-density map. Since NCS-related regions have more similar density patterns in their local neighborhoods compared with regions not related by NCS, a distinction can be made between NCS-related regions and those that are not based on a local density-correlation metric. These similarities in density correlation can be recognized even at moderate or low resolution. In this paper, two methods to compute NCS operators based on analysis of patterns in the electron-density maps are presented. A brute-force method was preliminarily developed to identify NCS, which involved an all-against-all computation of density correlations between regions of the electron-density map. While this provides an accurate identification of the NCS operators and the masks defining the region boundaries, it is inefficient. Thus, a heuristic feature-based pattern-recognition approach was developed to improve the time-performance of the algorithm.

The two methods start by constructing a rough initial approximation of Cα chains and use this to define centers of regions to compare. They can be implemented early on in the structure-determination process, thereby aiding rapid structure solution. Neither requires any information regarding heavy-atom positions or any sequence information. The operators determined by these algorithms are output in the canonical format as required by *DM* (in the *CCP*4 program suite) and are then automatically used to perform NCS averaging. The input to the methods is a set of initial solvent-flattened structure factors and the output is a set of improved structure factors along with the NCS operators and masks.

2. Methods

The algorithms presented here attempt to imitate the process intuitively used by a crystallographer to recognize regions of electron density related by NCS, *i.e.* visually examining the map for similar patterns of density in local neighborhoods. Local neighborhoods refer to spherical regions surrounding a particular point of interest. In this paper, regions are centered on putative Cα atoms obtained from a rough initial approximation of Cα chains determined by *CAPRA* (Ioerger & Sacchettini, 2002). *CAPRA* uses a neural network to reason about rotation-invariant features extracted from the electron-density maps and determine preliminary Cα chains. It is capable of identifying (approximately) the trace of the backbone even in noisy density and at medium resolutions. While these chains are not necessarily as accurate as a refined model, the algorithms described here are tolerant to the minor variations in Cα coordinates.

The feature-based algorithm then calculates rotation-invariant features (further described in §2.3) that characterize the density patterns in a region. The features are computed using a local neighborhood of 5 Å radius (the motivation for the choice of this parameter is further described in §2.6). Similar rotation-invariant features were used successfully in model building as implemented in *TEXTAL* (Ioerger & Sacchettini, 2003; Gopal *et al.*, 2003). These features are designed so that similarities in the electron density can be captured irrespective of the three-dimensional orientations of the corresponding regions.

In order to compare the densities between any two regions, they are first optimally superposed and the similarity is then computed using a local density-correlation metric. In *TEXTAL*, the optimal superposition was found by a method similar to sampling Euler angles, *i.e.* testing a subset of rotations to find that which superposes the two regions with the highest density correlation (Gopal *et al.*, 2004). The density correlation between any two regions given a rotation $\mathbf{R}$ is computed as

$$\mathrm{cc}(\mathbf{R}) = \int_{x,y,z=-\infty}^{\infty} \int_{x,y,z=-\infty}^{\infty} \int_{x,y,z=-\infty}^{\infty} \rho_1(\mathbf{R}\rho_2) \cdot w(\langle x, y, z\rangle)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z, \quad (1)$$

where ρ_1 and ρ_2 are the densities in the two regions at the point described by Cartesian coordinates $\langle x, y, z\rangle$ (each assumed to be translated to the origin) and $w(\langle x, y, z\rangle)$ is a weighting function characterizing the boundaries of the region. In this paper, a spherical region of radius 5 Å surrounding each Cα atom is used as the region of integration, which is accomplished by setting the weighting function to be

$$w(\langle x, y, z\rangle) = \begin{cases} 1 & \text{if } ||\langle x, y, z\rangle|| < 5 \text{ Å} \\ 0 & \text{otherwise} \end{cases}.$$

The rotation matrix that optimally superposes two regions is found by determining the $\mathbf{R}$ that maximizes the density correlation between the two regions. Hence, the optimal rotation matrix is

$$\mathbf{R}^* = \arg\max_{\mathbf{R}} \mathrm{cc}(\mathbf{R}). \quad (2)$$

Fig. 1 provides the pseudocode for the two algorithms described here. In both methods, the local density-correlation calculations output a best-matching region (region with

highest density correlation) elsewhere in the map for each region surrounding a Cα atom. This procedure also yields the rotation matrix $\mathbf{R}^*_{UV}$ that optimally superposes a region U (centered on $\overline{u}$) and its match V (centered on $\overline{v}$).

Next, rotation matrices are grouped into clusters based on similarity. Similar rotation matrices are defined based on whether they can map coordinates to the same locations (or close enough, within a tolerance of 2 Å).

Definition 1: similar rotation matrices. Given $\mathbf{R}^*_{UV}$ and $\mathbf{R}^*_{PQ}$ as rotation matrices that optimally superpose regions U and V and regions P and Q, respectively, and $\overline{u}$, $\overline{v}$, $\overline{p}$ and $\overline{q}$ as the coordinates of the centers of regions U, V, P and Q, respectively, then $\mathbf{R}^*_{UV}$ is similar to $\mathbf{R}^*_{PQ}$ if $\overline{q} - \mathbf{R}^*_{UV}\,\overline{p} \leq 2$ Å and $\overline{u} - \mathbf{R}^*_{\mathbf{PQ}}\,\overline{v} \leq 2$ Å.

This definition is used to collate similar rotations, resulting in clusters of rotations $\{\mathbf{R}^*_{U_iV_i}\}$ that relate multiple pairs of regions U_i and V_i. These pairs of regions can then be used to construct a common rotation matrix $\mathbf{R}^\dagger$ by simultaneously superposing the pairs of coordinates $\{\overline{u}_i\}$ and $\{\overline{v}_i\}$ over all matched pairs of regions in the cluster to minimize the r.m.s.d. (Kabsch, 1978; Mackay, 1984; Coutsias *et al.*, 2004).

To superpose two sets of points in the cluster, the centers of masses must first be translated to the origin. The relationship between the two sets of regions in the cluster can then be described as

$$(\overline{u}_i - \overline{c}_u) = \mathbf{R}^\dagger(\overline{v}_i - \overline{c}_v), \tag{3}$$

where $\overline{c}_u$ and $\overline{c}_v$ are the centroids of the two sets of regions $\{U_i\}$ and $\{V_i\}$, respectively. (3) can be rewritten as

$$\overline{u}_i = \mathbf{R}^\dagger\overline{v}_i + \mathbf{T}^\dagger. \tag{4}$$

Program flow

Step 1: Run *FINDMOL* on the input MTZ file to find a contiguous region of protein and create an electron-density map surrounding this contigous protein region.

Step 2: Run *CAPRA* and find approximate Cα positions for the input protein.

Step 3: if method = feature-based

Compute density-based features for regions surrounding each Cα atom.

For each Cα atom, select a subset of regions which have very similar feature vectors, *i.e.* normalized Euclidean distance < 0.04.

For each Cα atom, compute local density correlations with the selected subset of regions.

else if method = brute-force

For each Cα atom, compute local density correlations between all the regions in the electron-density map.

Step 4: For each Cα atom find its top match (region with the highest density correlation).

Step 5: Cluster all pairs of Cα atoms that are related by similar rotation matrices.

Step 6: Superpose matched pairs of Cα atoms in the largest cluster using the Kabsch algorithm to find the optimal rotation matrix $\mathbf{R}^\dagger$ and optimal translation vector $\mathbf{T}^\dagger$.

Step 7: Extend and refine the cluster transforms by appending other atoms that transform similarly.

Step 8: Repeat from Step 5 to find $N - 1$ Rs and Ts.

Figure 1
Flowchart showing the algorithm flow for the two methods presented in this paper.

where $\mathbf{T}^\dagger$ is the translational component between the two sets of points and is defined as

$$\mathbf{T}^\dagger = \overline{c}_v - \mathbf{R}^\dagger\overline{c}_u. \tag{5}$$

$\langle\mathbf{R}^\dagger, \mathbf{T}^\dagger\rangle$ will be henceforth referred to as cluster transforms.

2.1. Density-map preparation and generation of Cα chains

The input to the algorithms is a set of structure factors, which are then used to generate an electron-density map. This map should contain representatives from all N subunits of the protein in the asymmetric unit, but none extra. To facilitate the backbone tracing used by the algorithm, it is helpful for the map to cover a complete molecule. To accomplish this, *FINDMOL* (McKee *et al.*, 2005) is used, which identifies a contiguous cluster of atoms representing the complete protein molecule using symmetry operations. A map, centered on these atoms, with borders around it and the excess density is masked to zero, is created.

CAPRA is then used to analyze the map built over the multiple subunits of the protein and build a set of Cα chains that roughly approximates the protein backbone. Owing to occasional breaks in the electron density, the *CAPRA* output defining a single subunit of the protein typically consists of multiple Cα chains, with lengths ranging from 10 to 100 Cα atoms.

2.2. Local density-correlation calculations

In the case of the brute-force algorithm, the local density-correlation metric is used to compute an all-against-all comparison between the regions surrounding each Cα atom in the map. In the case of the feature-based algorithm, the correlation is only computed between pairs of regions obtained after filtering out regions with very large feature-vector differences. In both cases, for each Cα atom (C_i), its top match [$M(C_i)$, the region with highest density correlation] and the rotation matrix, $\mathbf{R}^*_{C_iM(C_i)}$, optimally superposing these two regions is found.

When the number of NCS operators is greater than two, a single top match for each Cα atom is still used. However, different atoms from one subunit may be mapped to different symmetry copies. This ensures that if some region of one of the protein subunits is less ordered, the algorithms can still accurately identify the relationship between other NCS copies of the same region.

2.3. Feature-based region matching

For the feature-based algorithm, numeric feature vectors based on the electron-density patterns are calculated for each spherical region centered on a Cα atom in the protein. The features used are listed.

(i) The number of neighbors (Cα atoms) within a 5 Å radius.

(ii) The average value of density at all the neighboring Cα atoms within a 5 Å radius.

(iii) The distance between the central Cα atom and the center of mass of neighbors.

(iv) The eigenvalues, sorted by magnitude, for the three mutually perpendicular moments of inertia calculated based on the inertia matrix.

(v) The three ratios of the eigenvalues: λ_0/λ_1, λ_0/λ_2 and λ_1/λ_2, where λ_i is the ith eigenvalue

(vi) The standard deviation of the densities at each Cα atom within the 5 Å radius.

(vii) The variance of the densities at each Cα atom within the 5 Å radius.

(viii) The skew of the densities at each Cα atom within the 5 Å radius, where skew is defined as

$$\text{skew} = \frac{\sum_{i=1}^{n}(x_i - \mu)^3}{N * (\sigma)^3}. \quad (6)$$

(ix) The kurtosis of the densities at each Cα atom within the 5 Å radius, where kurtosis is defined as

$$\text{kurtosis} = \frac{\sum_{i=1}^{n}(x_i - \mu)^4}{N * (\sigma)^4}. \quad (7)$$

Each region has a unique feature vector $\langle g_i \ldots g_K \rangle$ (where K is the number of features) that is a function of its local density pattern. Since each of the features has a different range, it is necessary to normalize the feature vectors to ensure that this difference in ranges does not bias the feature-vector comparison (Duda & Hart, 1973). Given a set of M feature vectors $\{\langle f_{11} \ldots f_{1K} \rangle, \ldots, \langle f_{M1} \ldots f_{MK} \rangle\}$, the normalized feature vector $\langle g'_1 \ldots g'_K \rangle$ is computed as

$$g'_j = \frac{g_j - a_j}{b_j - a_j}, \quad (8)$$

where $a_j = \min_i\{f_{ij}\}$ and $b_j = \max_i\{f_{ij}\}$ are the minimum and the maximum values of the jth feature over all i feature vectors.

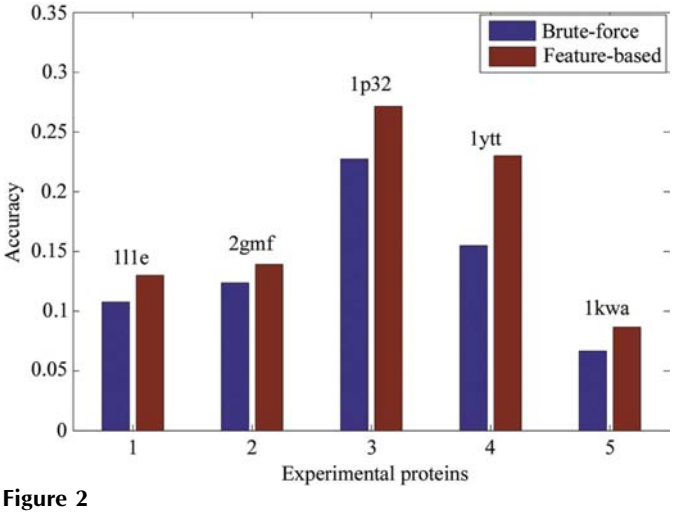

Figure 2
Comparison of accuracy of identifying the NCS matches using the brute-force method and the feature-based method presented in this paper on a subset of six maps.

Given two normalized feature vectors $\langle g'_1 \ldots g'_K \rangle$ and $\langle h'_1 \ldots h'_K \rangle$, the similarity between them is evaluated using the Euclidean distance metric (Duda & Hart, 1973) as

$$d = \left[\frac{1}{K}\sum_{j=1}^{K}(g'_j - h'_j)^2\right]. \quad (9)$$

The mean normalized feature differences between all possible pairs of feature vectors (representing regions centered on Cα atoms) are then used in a selection step. For each feature vector, only a subset of regions with relatively small feature difference scores are considered for future density-correlation calculations. This selection step is based on the reasoning that regions related by NCS are expected to have similar density patterns in the local neighborhood and therefore similar feature vectors. Hence, pairs of regions with highly dissimilar density patterns are filtered out.

The filtering procedure requires the determination of a feature-difference threshold. The feature-vector difference for most of the NCS-related regions should be below this threshold, which was empirically determined to be 0.04 (further details regarding threshold selection are given in §2.7). Hence, a subset of regions with $d \leq 0.04$ is chosen as candidate-matching pairs for future computations of local density correlations.

The feature-based algorithm was developed mainly to improve the time-performance of the brute-force algorithm (which scales up as N_{ca}^2, where N_{ca} is the number of predicted Cα atoms). The number of correlation calculations required is significantly reduced by filtering out a subset of regions with $d \geq 0.04$. The time-performance of the algorithm can be further improved by setting a threshold for the maximum number of correlations computed for each region. This threshold was determined experimentally to be $N_{ca}/5$.

The filtering steps drastically reduced the time taken for the local density-correlation calculations. For example, on a map with 810 Cα atoms, the time taken for the algorithm reduced from 51 to 7 min owing to the filtering. However, the improvements in time-performance arising from the aforementioned heuristics do not significantly degrade the accuracy of the algorithm (see Fig. 2).

2.4. Extending and refining cluster transforms

The pairs of coordinate centers related by local rotation matrices in a cluster are superposed to find the common cluster transformation which best superposes these sets of points. The confidence in a cluster transformation is proportional to the number of center pairs in the cluster. Therefore, in order to find the initial estimates of the NCS operators relating N protein subunits, $N - 1$ cluster transforms are chosen in order of largest first in terms of the number of matched region pairs in the cluster.

Since the initial estimates for each cluster transform are based on a small set of coordinate pairs (region centers), it is necessary to refine these operators. The refinement is achieved by extending the number of matched pairs of regions in each cluster. This extension is based on the assumption that when

an appropriate transformation is applied to an additional atom, there should be a nearby atom in its implied symmetry position. Hence, the initial $\mathbf{R}^{\dagger}$ and $\mathbf{T}^{\dagger}$ are applied to all the Cα atoms in the map. If a transformed Cα atom is close (within 2 Å) to any other Cα atom, then both Cα atoms are added to the initial center pairs and $\mathbf{R}^{\dagger}$ and $\mathbf{T}^{\dagger}$ are recomputed. Each of the $N-1$ cluster transforms are similarly extended and these $N-1$ refined transforms then yield the final NCS operators between the N protein subunits.

2.5. Output manipulations

The output consists of a set of $N-1$ transformations, including the refined cluster transformations and the corresponding center pairs. While these transformations describe the relationships between all of the protein subunits, they need not necessarily be based on a common protein subunit. For example, the transformations for a map with four protein subunits $M_1 \ldots M_4$ could be of the form $M_1 \rightarrow M_2$, $M_2 \rightarrow M_4$, $M_3 \rightarrow M_4$. In order to input directly into *DM* they have to be made relative to a single protein subunit, *i.e.* they have to be in the form of $M_1 \rightarrow M_2$, $M_1 \rightarrow M_3$ and $M_1 \rightarrow M_4$. This conversion into the *DM* canonical format can be achieved with a few matrix manipulations. For example, considering the two transformations relating $M_1 \rightarrow M_2$ and $M_2 \rightarrow M_4$, they can be written as

$$M_2 = \mathbf{R}_1 \cdot M_1 + \mathbf{T}_1, \quad M_4 = \mathbf{R}_2 \cdot M_2 + \mathbf{T}_2. \tag{10}$$

The second equation can be rewritten to provide a relationship between M_1 and M_4 as follows,

$$M_4 = (\mathbf{R}_2 \cdot \mathbf{R}_1) \cdot M_1 + (\mathbf{R}_2 \cdot \mathbf{T}_1 + \mathbf{T}_2). \tag{11}$$

The resulting transformation between $M_1 \rightarrow M_4$ consists of a rotational component given by $\mathbf{R}_2 \cdot \mathbf{R}_1$ and a translational component given by $\mathbf{R}_2 \cdot \mathbf{T}_1 + \mathbf{T}_2$.

In order to represent the transformations in canonical form, it is necessary to define the protein subunit boundaries. When the map is skeletonized by an automated algorithm such as *CAPRA*, the extent of each subunit is not always obvious, since the protein backbone is often broken into multiple chains and these chains have to be partitioned into the various subunits. Our approaches to extending clusters grow regions to maximum boundaries that preserve symmetry, can detect and exclude local regions of non-isomorphism and are not sensitive to breaks in backbone connectivity.

This method of determining boundaries works best in cases of improper symmetry, when the NCS operators do not form a closed group (*e.g.* the operator that transforms A to B does not transform B to A). In such cases, different operators are required, so it can be safely assumed that any given transformation will not hold outside of the two subunits related by that transformation. This allows a clear distinction between the various subunits and each symmetric unit is defined as a set of Cα atoms that, when transformed using one of the $N-1$ transformations, are within 2 Å of their NCS matches. This grouping of atoms can be used to define a mask in the traditional sense. In the case of proper NCS symmetry, there can be ambiguity in the boundaries between some protein subunits. However, molecule boundaries could be identified by other automated masking methods (Vellieux & Read, 1997).

2.6. Parameter selection: radius for density correlations

To find an NCS match for a Cα atom, it is necessary to compare density patterns in its local neighborhood to the patterns surrounding other Cα atoms. In this work, the definition of a local neighborhood is a spherical region centered on a Cα atom. The radius must be chosen with care. An ideal radius is one that is large enough to ensure uniqueness while at the same time being general enough to recognize the similarities between the density patterns in a region and its NCS matches. The optimal radius was determined empirically by comparing regions in a subset of maps. The density patterns between pairs of these regions, centered on Cα atoms, were compared using local density correlation at various radii.

In order to evaluate the efficacy of the algorithm over the various radii, two metrics, distinguishability and accuracy, were used. These two metrics define the suitability of the density correlation as a distinguisher between regions related by NCS and those that are not. Distinguishability is defined as

$$D = \frac{\text{cc(true match)}}{\text{cc(top match)}}, \tag{12}$$

where cc(true match) refers to the density correlation between region U and its true NCS match (any Cα within 2 Å of one of its symmetry positions, based on the actual operators from the refined model) and cc(top match) refers to the highest density correlation between region U and any other region in the electron-density map. The value of distinguishability is equal to 1 when the top match region is the same as the true match region. A value close to 1 suggests that the density correlation between the region U and its true match is very close to the density correlation between region U and its top match.

Accuracy is defined as the ratio of the number of times the top match is a true NCS match,

$$A = \frac{\text{count(top match} = \text{true match)}}{\text{total number of CA atoms}}. \tag{13}$$

In order to find the ideal radius based on these two metrics, distinguishability and accuracy, a subset of regions in several test maps were considered and the variation of both metrics over different radii (ranging from 4 to 10 Å) was computed for pairs of regions in each map. For each Cα atom in the test maps, a match atom was found (as described previously in §2). The values of distinguishability and accuracy were then computed based on (12) and (13).

Fig. 3 shows the variation of distinguishability over the radius range for all the different maps in the subset and Fig. 4 shows the variation of accuracy over the same range for the maps. Both graphs show that there is a marked degradation in distinguishability and accuracy when the radius for local correlation is greater than 6 Å. The two values are maximal in most cases at 5 Å. Hence, in subsequent experiments a radius of 5 Å was used to limit the size of the local neighborhood for each Cα atom for density-correlation calculations.

2.7. Parameter selection: feature-difference threshold

Regions related by NCS are expected to have similar density patterns and hence very similar feature vectors. Thus, the feature-vector difference can be used to filter out regions that have dissimilar feature vectors and hence are not likely to have similar density patterns. A feature-difference threshold that culls out dissimilar pairs of regions and retains a subset of regions that have similar density patterns was experimentally determined based on a subset of maps. For each of these maps, the threshold was varied from 0 to 0.2 in steps of 0.01. For each threshold, the percentage of the candidate number of region pairs eliminated owing to the threshold and the percentage of remaining regions with the correct NCS match were computed.

Fig. 5 shows the variation of these two percentages with the threshold value. The figure shows that the feature difference between a region and its true NCS match is almost always less than 0.1. However, this threshold does not effectively cull out dissimilar regions. A threshold of 0.04 eliminates the largest percentage of dissimilar regions, while at the same time allowing the true match to be retained in the subset of regions considered for future calculations. Hence, the threshold was set at 0.04. Fig. 6 shows the drastic reduction in time, typically by over 60%, achieved by applying this threshold for feature-based filtering of candidate pairs of matching regions.

3. Results and discussion

The algorithms were evaluated using 11 experimental structure-factor data sets. The number of NCS-related subunits in these experimental electron-density maps varied from two to eight. *FINDMOL* was used to generate the electron-density maps for the proteins within a contiguous envelope surrounding the proteins at 2.8 Å resolution. The resolution of the map is not a limiting factor for the algorithm itself. However, *CAPRA* performs best at medium resolutions and since the backbone prediction is the main basis for the feature-matching component, the maps were calculated at 2.8 Å.

The experimental data sets used in this study were phased by various methods (MAD, MIR, SAD) and were all density-modified [solvent-flattened in *CNS* (Brünger *et al.*, 1998), although none were NCS-averaged]. Solvent flattening is important for *CAPRA*, but since this is a straightforward pre-processing step, this requirement does not reduce the generality of the proposed methods. The experimental data sets were for cyclopropane synthase (Huang *et al.*, 2002; PDB code 1l1e), granulocyte macrophage colony-stimulating factor (Rozwarski *et al.*, 1996; PDB code 2gmf), isocitrate lyase (Sharma *et al.*, 2000; PDB code 1f61), flavin reductase (Tanner *et al.*, 1996; PDB code 1bkj), mannose-binding protein (Burling *et al.*, 1996; PDB code 1ytt), osmotically induced protein C from *Escherichia coli* (Shin *et al.*, 2004; PDB code 1nye), P32 (Jiang *et al.*, 1999; PDB code 1p32), PDZ domain (Doyle *et al.*, 1996; PDB code 1kwa), *S*-adenosylhomocysteine hydrolase (Turner *et al.*, 1998; PDB code 1a7a), phosphatase from *Thermotoga maritima* (Shin *et al.*, 2003; PDB code 1nf2) and Lyme disease variable surface antigen (Eicken *et al.*, 2002; PDB code 1l8w). In this work, the experimental phases before symmetry averaging were used as inputs to the algorithms. Table 1 further describes the input data sets. The quality of the input phases for each of the experimental data sets was measured using the normal correlation coefficient between the map densities based on true phases ($2F_o - F_c$ map calculated from refined model) and experimental phases. Table 1 shows that the map correlation for the data sets ranges from 0.31 to 0.84. Lunin & Woolfson (1993) suggest that a correlation coefficient greater than 0.4 or 0.5 indicates a promising starting point for map interpretation.

Both the algorithms were able to determine the NCS operators from the experimental phases. The accuracy of the transformations output by the two algorithms can be calculated by superposing the protein subunits related by each transformation. Given two protein subunits X and Y, one way to measure the accuracy of the transformation relating X and Y is to evaluate the superposition of transformed X (X') on Y. If the superposition is such that all the NCS-related Cα atoms in X' and Y are placed in close proximity (low r.m.s.d.) to each other, then the transformation matrix is said to be accurate.

Table 1
Information about proteins used in this study.

PDB code	Original resolution (Å)	Space group	No. of NCS subunits	Map correlation
1a7a	2.8	$C222$	2	0.845
1bkj	1.8	$P2_1$	2	0.443
1l1e	2.8	$P6_5$	2	0.505
2gmf†	2.35	$P2_12_12_1$	2	
1f61	1.8	$P6_522$	2	
1nye	3	$P2_1$	8	0.506
1kwa	1.93	$C222_1$	2	0.475
1l8w	2.3	$P2$	4	0.454
1p32	2.25	$P2_1$	3	
1nf2†	3	$C2$	3	0.313
1ytt	1.8	$P2_12_12_1$	2	0.667

† NCS averaging was used to solve the final structures of 2gmf and 1nf2, but the non-averaged phases were used as input to the algorithms described in this paper.

Table 2
Results using the brute-force and feature-based methodologies described in this paper.

		Average distance between NCS-related Cα atoms	
PDB code	No. of NCS subunits	Brute-force	Feature-based
1a7a	2	0.667	0.670
1bkj	2	0.829	0.819
1l1e	2	0.733	0.739
2gmf	2	0.858	0.857
1f61	2	0.656	0.655
1nye	8	0.758, 0.768, 0.771 0.779, 0.807, 0.813	0.713, 0.757, 0.771 0.819, 0.844, 0.917
1kwa†	2	1.06	1.43
1l8w	4	0.954, 1.039, 1.03	0.82, 0.858, 1.09
1p32	3	0.752, 0.926	0.801, 0.883
1nf2	3	0.954, 0.976	0.954, 0.979
1ytt	2	0.791	0.780

† The NCS operators for 1kwa were determined using a density correlation radius of 6 Å.

Table 2 shows that the brute-force method is able to accurately identify the NCS relationships between the various subunits and superpose the structures accurately. The r.m.s.d. of superposition based on the computed NCS operators ranges from 0.65 to 1.06 Å. All expected operators are found in each case, except for 1nye (six instead of seven, for $N = 8$).

The largest r.m.s.d. of superposition is for 1kwa (1.06 Å), which also happens to have a low map correlation (0.475 Å) with the $2F_o - F_c$ map, suggesting that there might be a link between the quality of the operators obtained and the quality of the input data. Additionally, for structure 1kwa, NCS operators could not be accurately determined when the density correlation was computed using a spherical radius of 5 Å, but the operators were accurately determined with an r.m.s.d value of 1.06 Å when the density correlation was computed using a spherical radius of 6 Å. This indicates that for some structures with low map correlation it might be necessary to increase the radius for density correlations. This increase in radius might mitigate the effects of local noise by capturing information over a larger neighborhood.

Table 2 also shows that the feature-based method is able to accurately identify the NCS relationships between the various subunits and superpose the structures with an accuracy almost equivalent to that of the brute-force method. These outputs

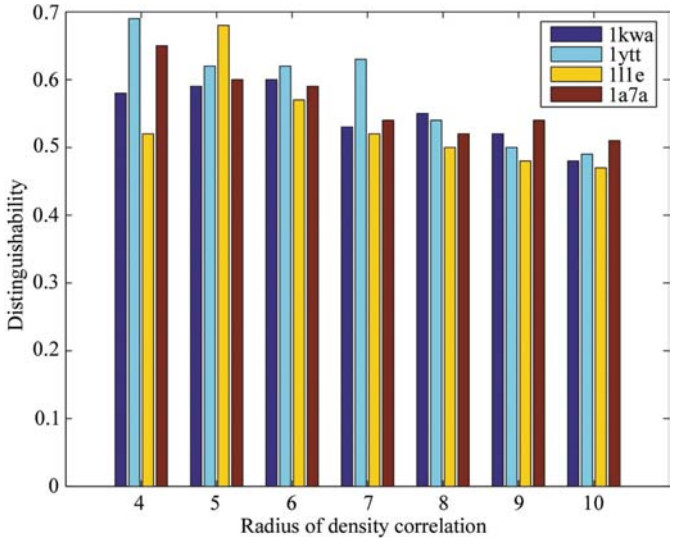

Figure 3
Variation of distinguishability of matching based on local density-correlation radii for four maps.

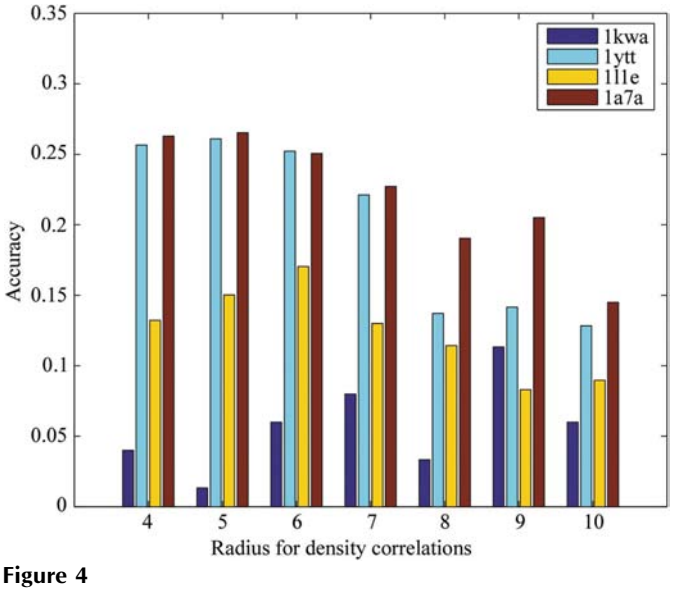

Figure 4
Variation of accuracy over different local density-correlation radii for four maps.

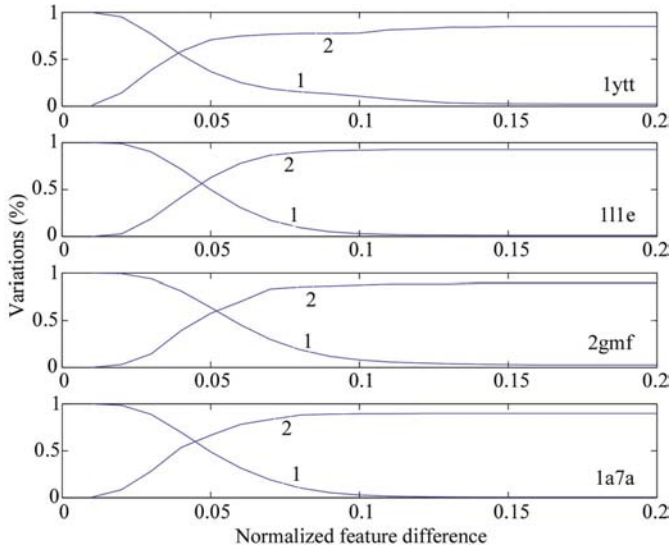

Figure 5
Plot used to determine the feature-difference threshold. Curve 1 for each map shows the percentage of pairs of regions eliminated owing to the feature-difference threshold. Curve 2 for each map shows the percentage of times the true match for a region has a feature difference less than the threshold.

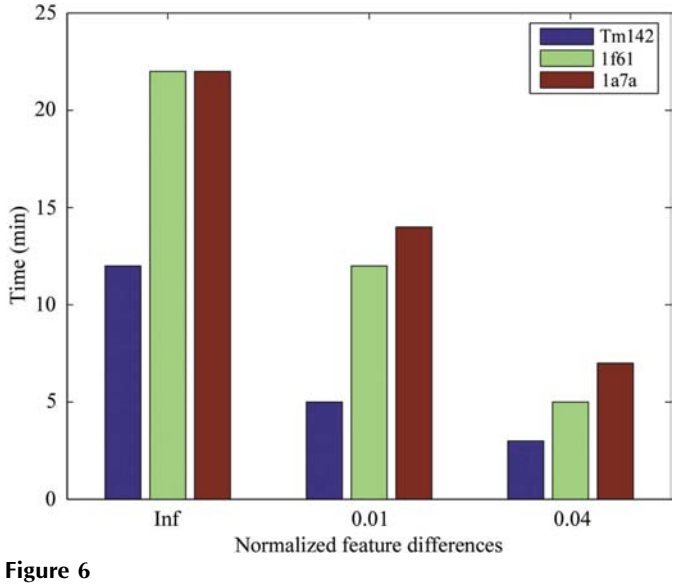

Figure 6
Variation of time for NCS determination by the feature-based method using various feature difference thresholds. 'Inf' means no threshold was used, which simulates the brute-force method.

are obtained much faster (a decrease of almost 60% in computational time for most of the test cases) with fewer density correlation computations (owing to the selection procedure based on feature similarity). The average difference in the accuracies of the brute-force and the feature-based methods is less than 0.02. Once again, the best is around 0.65 Å for 1f61 and the operators obtained for 1kwa (at 6 Å) yield the highest r.m.s.d. value (1.43 Å).

Both the brute-force and the feature-based methods are able to compute the boundaries of each subunit accurately for most of the test cases. All the NCS operators are based on the approximate positions of the Cα atoms and are therefore still unrefined. This results in subunit definitions that are not completely accurate. Additionally in the case of some proteins such as 1f61, 1bkj and 1a7a, all of which are dimers, proper NCS symmetry relates the two subunits. This results in a failure to accurately partition the Cα atoms based on the method presented in this paper, even though the operators were accurately defined. In the case of the trimeric protein 1nf2, there is a mix of proper and improper NCS symmetry. The subunit related by improper NCS is accurately partitioned, whereas in the case of the subunits related by proper NCS, a few Cα atoms (less than 20 atoms) are incorrectly partitioned. Table 3 lists the percentage of Cα atoms of each subunit that were correctly classified using the determined NCS operators.

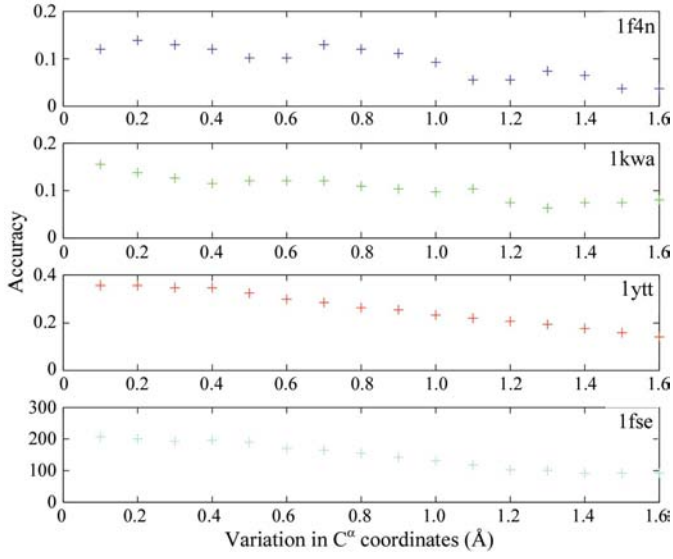

Figure 7
Variation of accuracy with random perturbations in Cα coordinates for the feature-based method, starting from the true Cα coordinates (from refined model).

Table 3
Subunit-boundary definitions using the NCS operators from the two methods.

PDB code	No. of NCS subunits	No. of NCS subunits found	Percentage of Cα atoms accurately assigned using NCS
1l1e	2	2	0.86
2gmf	2	2	1
1kwa	2	2	0.64
1nf2	3	3	1, 0.95
1ytt	2	2	0.94

The rotation-invariant features used in this study attempt to capture the general density patterns in a local neighborhood. Hence, the features are tolerant of minor inaccuracies in Cα positions, as well as insertions and deletions in the backbone. Fig. 7 shows the variation of accuracy of the feature-based method as a function of the variations in Cα coordinates. The position of each Cα atom from the true structure was varied systematically from 0.01 to 1.6 Å (perturbing the coordinates of each atom in random directions). The figure shows that the performance of the algorithm degrades gradually. For example, in 1ytt, the accuracy drops from 0.4 to 0.2 with 1 Å of error. Furthermore, the process used to extend these feature matches is also resistant to these inaccuracies in the backbone, making the feature-based approach a very robust alternative to traditional approaches of NCS-operator determination.

Fig. 8 shows graphically the accuracy of the superpositon obtained by the transformations output from the feature-based method for three example proteins: 2gmf (two subunits), 1a7a (two subunits) and 1p32 (three subunits).

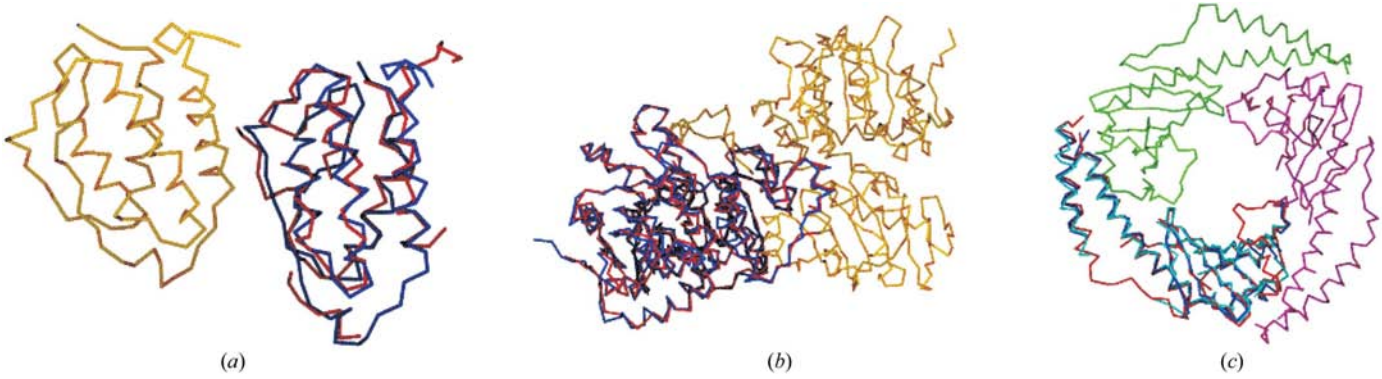

Figure 8
Superposition of NCS-related subunits obtained using the transformation matrices for the proteins (*a*) 2gmf (two identical subunits), (*b*) 1a7a (two identical subunits) and (*c*) 1p32 (three identical subunits). The transformed subunits are superposed onto the original subunits. Backbone models (Cα traces only) are shown, with different subunits shown in different colors, with one subunit superposed onto its symmetry copy using the operators found. In the case of (*c*), both the green and purple subunits are shown superposed onto the red subunit.

Table 4
Comparison of map correlation before and after NCS averaging by *DM* using operators found by the feature-based algorithm.

	Map correlation	
PDB code	Before NCS averaging	After NCS averaging
1a7a	0.845	0.859
1bkj	0.443	0.600
1kwa	0.475	0.531
1ytt	0.667	0.692

The NCS operators obtained by the feature-based algorithm were input to *DM* to perform NCS averaging (using masks generated automatically from the operators in *CCP*4). This averaged density was then compared with the $2F_o - F_c$ map. Table 4 shows the improvement in the map correlation after NCS averaging for a subset of maps. For most of the maps, the map correlation increases by around 10%. In the case of 1bkj this increase is almost 25%. This shows that the NCS operators identified by the two methods are able to make an improvement to the map quality and thereby increase the potential for solving the structure.

3.1. Case study 1: PGDH

The structure of *Mycobacterium tuberculosis* phosphoglycerate dehydrogenase (PGDH) was solved at 3.1 Å (Dey *et al.*, 2005; PDB code 1ygy) using SeMet MAD phasing. There are two subunits in the ASU and the NCS operators were unknown when the structure was initially solved; hence, no NCS averaging had been performed. Each of the subunits has four well defined domains: a nucleotide-binding domain, a substrate-binding domain, a regulatory domain and an intervening domain. The nucleotide-binding, substrate-binding and regulatory domains are homologous to those found in *E. coli*. The intervening domain is not found in the *E. coli* PGDH. Additionally, the structure indicates the presence of two different conformations among the subunits in the ASU. Superposition of the various domains between the two subunits shows that if the nucleotide-binding and the substrate-binding domains are used as reference, the other two domains are rotated by approximately 180°.

The 3.1 Å experimental data set was input to the feature-based algorithm to determine the NCS operators between the two molecules (represented by chains *A* and *B*). As described earlier, there are two NCS transformations that exist between the two molecules. The approach developed in this paper requires the specification of the number of transformations expected. For this particular test case, the number of transformations was required to be set to three[1]. The first transformation finds the NCS relationship between the nucleotide and substrate-binding domains of chain *A* with those of chain *B*, the second finds the NCS relationship between the nucleotide and substrate-binding domains of chain *B* with those of chain *A* and the third transformation finds the NCS relationship between the regulatory and intervening domains of chain *A* with those of chain *B*. The first and the third transformations are then selected to perform NCS averaging using *DM*.

Using the feature-based algorithm described in this paper, the substrate- and nucleotide-binding domains between the two subunits are superposed with an r.m.s.d. of 1.03 Å and the regulatory and intervening domains are superposed with an r.m.s.d. of 0.89 Å.

3.2. Case study 2: SecE2

This is an unpublished structure (A. Arockiasamy & J. Sacchettini) from *M. tuberculosis* annotated as SecE2 (Rv0397). It was solved at 1.8 Å using platinum MAD phasing. There are 12 subunits in the ASU and the NCS operators were not used to solve the structure. The structure was solved by using its homology with a dodecine from *M. tuberculosis*. This spherical dodecamer has a total of 852 residues and is a conserved protein whose function is unknown. *CAPRA* built 95% of these Cα atoms.

Eight unique NCS operators were determined using the pattern-recognition algorithm described here. These operators were successfully extended using the methology described earlier to determine the relationships among all the 12 subunits with the following r.m.s.d. values (in Å): 0.89, 0.9, 0.91, 1.03, 1.04, 1.05, 1.06, 1.06, 1.08, 1.3 and 1.46.

4. Conclusion

This paper describes two related approaches to determining NCS relationships in a map with multiple NCS-related subunits. Both approaches use local density correlation to detect symmetry between regions of the map and this metric has proved to be a robust and accurate indicator of NCS.

The brute-force algorithm determines symmetry-related regions based on all-against-all comparison using local density correlations and was able to accurately determine the NCS operators in all but one of the test cases. It has advantages over traditional methods that rely on heavy-atom sites. However, it is a computationally intensive algorithm. The feature-based algorithm was designed to address this concern. It uses rotation-invariant features to characterize regions in the map and reduces the correlation computations required by filtering out pairs of regions whose feature vectors have high difference. The feature-based filtering makes the approach more efficient than the brute-force algorithm described here.

The two approaches can be applied to unrefined maps (although solvent flattening is usually still necessary for automated backbone tracing), are tolerant of noise and require neither the location of heavy atoms nor any other information such as the amino-acid sequence. The use of local density correlation over regions of 5 Å makes these algorithms robust. Despite the fact that these approaches use only a rough approximation of the Cα chains, the results obtained from both the algorithms are nearly as good as the results

[1] In cases when the number of NCS operators (N) is unclear, it is possible to search for M ($\geq N$) operators. The additional operators can be filtered out based on the number of regions that can be superposed using the operators.

obtained by previous methods that determine NCS operators from previously built models.

Given the accuracy of the algorithms even for medium-quality maps, they hold great promise for determining NCS operators and performing symmetry averaging to automatically improve the quality of electron-density maps. The feature-based algorithm is available online at http://textal.tamu.edu/NCS/index.html.

This work was supported in part by NIH grant GM-63210 and the Robert A. Welch Foundation (to JCS). The authors wish to thank Dr Paul Adams (Lawrence Berkeley National Laboratory) for supplying the data sets used in this study.

References

Bailey, S., Dodson, E. & Phillips, S. (1988). Editors. *Proceedings of the CCP4 Study Weekend. Improving Protein Phases.* Warrington: Daresbury Laboratory.

Bricogne, G. (1974). *Acta Cryst.* A**30**, 395–405.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, N., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Burling, F. T., Weis, W. I., Flaherty, K. M. & Brünger, A. T. (1996). *Science*, **271**, 72–77.

Coutsias, E. A., Seok, C. & Dill, K. A. (2004). *J. Comput. Chem.* **25**, 1849–1857.

Cowan, S. W., Newcomer, M. E. & Jones, T. A. (1993). *J. Mol. Biol.* **230**, 1225–1246.

Cowtan, K. D. & Main, P. (1993). *Acta Cryst.* D**49**, 148–157.

Crowther, R. A. (1967). *Acta Cryst.* A**22**, 758–764.

Dey, S., Grant, G. A. & Sacchettini, J. C. (2005). *J. Biol. Chem.* **280**, 14892–14899.

Doyle, D. A., Lee, A., Lewis, J., Kim, E., Sheng, M. & MacKinnon, R. (1996). *Cell*, **85**, 1067–1076.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* New York: Wiley.

Eicken, C., Sharma, V., Klabunde, T., Lawrenz, M. B., Hardham, J. M., Norris, S. J. & Sacchettini, J. C. (2002). *J. Biol. Chem.* **277**, 21691–21696.

Gopal, K., Pai, R, Ioerger, T. R., Romo, T. & Sacchettini, J. C. (2003). *Proceedings of the Fifteenth Conference on Innovative Applications of Artificial Intelligence*, pp. 93–100. Menlo Park, CA, USA: AAAI.

Gopal, K., Romo, T. D., Sacchettini, J. C. & Ioerger, T. R. (2004). *Computational Systems Bioinformatics Conference*, pp. 255–265. Los Alamitos, CA, USA: IEEE.

Huang, C., Smith, C. V., Glickman, M. S., Jacobs, W. R. & Sacchettini, J. C. (2002). *J. Biol. Chem.* **277**, 11559–11569.

Ioerger, T. R. & Sacchettini, J. C. (2002). *Acta Cryst.* D**58**, 2043–2054.

Ioerger, T. R. & Sacchettini, J. C. (2003). *Methods Enzymol.* **374**, 244–270.

Jiang, J., Zhang, Y., Kranier, A. R. & Xu, R. M. (1999). *Proc. Natl Acad. Sci. USA*, **96**, 3572–3577.

Kabsch, W. (1978). *Acta Cryst.* A**34**, 827–828.

Kleywegt, G. J. & Jones, T. A. (1994). *Proceedings of the CCP4 Study Weekend. From First Map to Final Model*, edited by S. Bailey, R. Hubbard & D. Waller, pp. 59–66. Warrington: Daresbury Laboratory.

Lu, G. (1999). *J. Appl. Cryst.* **32**, 365–368.

Lunin, V. Yu. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.

McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Cryst.* D**61**, 458–464.

Mackay, A. L. (1984). *Acta Cryst.* A**40**, 165–166.

McKee, E., Kanbi, L. D., Childs, K. C., Grosse-Kunstleve, R. W., Adams, P. A., Sacchettini, J. C. & Ioerger, T. R. (2005). *Acta Cryst.* D**61**, 1514–1520.

Main, P. & Rossmann, M. G. (1966). *Acta Cryst.* **21**, 67–72.

Muirhead, H., Cox, J. M., Mazzarella, L. & Perutz, M. F. (1967). *J. Mol. Biol.* **28**, 117–156.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Rossmann, M. G. (1972). *The Molecular Replacement Method.* New York: Gordon & Breach.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.

Rozwarski, D., Diederichs, K., Hecht, R., Boone, T. & Karplus, P. (1996). *Proteins*, **26**, 304–313.

Sharma, V., Sharma, S., Bentrup, K. H., McKinney, J. D., Russell, D. G., Jacobs, W. J. & Sacchettini, J. C. (2000). *Nature Struct. Biol.* **7**, 663–668.

Shin, D. H., Choi, I.-G., Busso, D., Jancarik, J., Yokota, H., Kim, R. & Kim, S.-H. (2004). *Acta Cryst.* D**60**, 903–911.

Shin, D. H., Roberts, A., Jancarik, J., Yokota, H., Kim, R., Wemmer, D. E. & Kim, S.-H. (2003). *Protein Sci.* **12**, 1464–1472.

Tanner, J. J., Lei, B., Tu, S. C. & Krause, K. L. (1996). *Biochemistry*, **35**, 13531.

Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.

Terwilliger, T. C. (2002*a*). *Acta Cryst.* D**58**, 2213–2215.

Terwilliger, T. C. (2002*b*). *Acta Cryst.* D**58**, 2082–2086.

Turner, M. A., Yuan, C. S., Borchardt, R. T., Hershfield, M. S., Smith, G. D. & Howell, P. L. (1998). *Nature Struct. Biol.* **5**, 369.

Vellieux, F. M. D. & Read, R. J. (1997). *Methods Enzymol.* **277**, 18–53.