Not your average density

Gerard J Kleywegt[1] and Randy J Read[2]

Addresses: [1]Department of Molecular Biology, Uppsala University, Biomedical Centre, Box 590, SE-751 24 Uppsala, Sweden and [2]Department of Medical Microbiology and Immunology, University of Alberta, 1–41 Medical Sciences Building, Edmonton, Alberta, Canada T6G 2H7.

E-mail: gerard@xray.bmc.uu.se
E-mail: Randy.Read@ualberta.ca

Structure 15 December 1997, **5**:1557–1569
http://biomednet.com/elecref/0969212600501557

Introduction

It is a curious (and useful) fact that many proteins prefer to form crystals with multiple copies in the asymmetric unit. A recent statistical survey found that this happens in about one-third of protein crystals [1]. For proteins whose crystals do not diffract better than to 2.5 Å resolution, this happens in about half the cases (and 80% of these crystals contain two copies of the molecule per asymmetric unit) [2]. At one time, this phenomenon was generally regarded as a nuisance. The resulting increase in the size of the unit cell meant that there were more closely spaced diffraction spots to collect, and that there were more atoms to refine on computers which were already stretched to their limits. In papers from the 1970s reporting crystal structures, there are numerous examples of researchers throwing away such crystals and going back to the bench! With modern computers and data-collection methods, the practical difficulties have all but disappeared and the advantages of redundant information have come to the forefront.

In this review, we shall focus on the use of non-crystallographic symmetry (or NCS, as this redundancy is known) to obtain better electron-density maps. NCS is also of great importance in improving the parameter to observation ratio in refinement [2,3]. In addition, it provides some information about the effects of crystal packing on protein conformation, and details of inherent flexibility and conformational heterogeneity [4].

The past decade has seen a veritable renaissance of the use of methods that employ geometric redundancies [5] in macromolecular diffraction data. Simultaneously, several new methods have been developed to improve phases. Collectively, these techniques go under the monicker of density-modification or phase-refinement techniques. They all apply heuristics (rules of thumb) as constraints on either the electron-density map (real space) or the complex structure factors (reciprocal space). In addition to improving phases (and thereby the appearance of electron-density maps), these techniques can extend phase information to higher resolution and remove bias from maps calculated using phases derived from an incorrect or incomplete model.

Examples of density-modification methods are listed in Table 1. Density modification is usually carried out as a cyclic process, in which a map is calculated with the initial phase set, this map is modified according to one or more heuristics, and from the modified map, a new and more accurate set of complex structure factors is calculated. The new structure factors can then either be combined with the original ones or used directly to calculate a new map. In practice, a combination of several techniques tends to be more powerful than any individual technique.

Averaging

Molecular averaging is one of the strongest of the density-modification techniques. The underlying theory is well-established, and the utility of the technique has been proven time and again in practice. The foundation for the reciprocal-space formulation of averaging (as well as the general molecular replacement method) was laid in the classic 1962 paper of Rossmann and Blow [6]. The equivalence between real and reciprocal space NCS averaging was suggested by Main in 1967 [7], and in 1974 Peter Colman proved it for the special case of proper symmetry [8]. In the same year, Gérard Bricogne provided a general proof that real and reciprocal space NCS averaging are equivalent in a thorough paper [5], which was followed two years later by a practical implementation of the method [9]. As Colman and Bricogne pointed out, there are definite computational advantages to working in real space (see [10] for a historical discussion).

The underlying assumption of the averaging technique is that chemically identical copies of a molecule in physically distinct (independent) environments will nevertheless have identical conformations. The arrangements of atoms (and their electron density) therefore will be identical. This means that their densities can be averaged to achieve an improvement in the signal-to-noise ratio of the order of $\sqrt{N}$, if N is the number of independent copies. The strategy of improving signal-to-noise through averaging is used in many areas of science, for example to enhance electron microscopic images or (in the old days) continuous-wave NMR spectra of ^{13}C nuclei.

Table 1

Examples of density-modification techniques.

Technique	Heuristic	Remarks	References
Solvent flattening	the solvent in protein crystals is disordered in time and space and therefore has flat density	solvent flipping is even more powerful	[12,19,14]
Molecular averaging	chemically identical molecules in physically distinct environments have identical electron density	–	[5,6]
Iterative skeletonisation	density for a macromolecule is continuous and connected	–	[48]
Histogram matching	distribution of density values for a protein is known and depends only on resolution and a temperature factor	–	[49]
Sayre's equation	atomicity (density equals the squared density multiplied by a shape function)	not usually true for biomacromolecules, but it works in practice	[50]

There are many different circumstances in which a crystallographer may encounter (partially) identical copies of macromolecules:

1. Non-crystallographic symmetry (more than one copy of a molecule or complex in the asymmetric unit); this is a very common phenomenon [1,2], with the number of molecules in the asymmetric unit ranging from 2 up to 120.

2. Multiple crystal forms (each of which may or may not have NCS). This is quite common with current large-scale screening methods for crystallisation conditions, although often only the best diffracting crystal form is used.

3. Multiple-domain NCS, in which there are different spatial relationships involving copies of different domains of a protein (or two proteins in a complex). This occurs, for instance, with crystals of antibody F_{ab} fragments, in which different copies commonly have different hinge angles.

4. Partial NCS, in which only part of the asymmetric unit obeys the NCS relationship.

5. Any combination of the above.

It is important that the different copies of a molecule or domain are independent; technically, this means that the copies of the molecular transforms are sampled in distinct manners. For example, artificially reducing crystallographic symmetry (e.g. by considering a $P2_1$ crystal with one molecule per asymmetric unit as if it were P1 with twofold NCS) does not yield additional information, whereas two non-isomorphous crystal forms of the same molecule do. Less obviously, we shall also demonstrate later that the special case of purely translational NCS does not lead to completely independent copies of a molecule. In general, the averaging method is more effective with increasing number of independent copies (which is why this technique has been so prominent in virus crystallography) and also if the solvent content of the crystal(s) is relatively high.

Theory

The theory underlying molecular averaging is well-established and well-documented. We shall therefore only touch upon it briefly to demonstrate a few issues. The greatest insight is gained by considering what happens in reciprocal space when density is averaged in real space. If the starting map obeys NCS perfectly, the density should not change, so the structure factors should not change. As we shall show, this leads to relationships among the structure factors. In practice, there will be phase errors and the starting map will not obey NCS perfectly, so averaging will change the map and the corresponding structure factors. The equations relating structure factors will then tell us how information spreads in reciprocal space when averaging is carried out in real space.

Solvent flattening is usually carried out at the same time as averaging. In solvent flattening, the density in the solvent region is typically replaced by its average value ρ_s. The equations are derived more easily if we do this by first subtracting ρ_s everywhere in the map, multiplying by zero in the solvent region and by one elsewhere and finally adding ρ_s back at every point in the map at the end of averaging. In fact, in reciprocal space, adding and subtracting a constant density value will only change F(000), so we can effectively ignore ρ_s. The combination of flattening and averaging over N molecules related by NCS can then be expressed in the following equation (illustrated schematically in Figure 1):

$$\rho_{avg}(\mathbf{x}) = \Sigma_{i=1,N}\, M_i(\mathbf{x})\, (1/N)\, \Sigma_{j=1,N}\, \rho(\mathbf{x}_{ij}) \qquad (1)$$

in which $M_i(\mathbf{x})$ is a mask or envelope function that has a value of one inside the volume U_i, enclosing molecule i, and zero elsewhere, $\mathbf{x}$ is a position vector inside the unit

cell, and $\mathbf{x}_{ij}$ is $\mathbf{x}$ transformed by the NCS operator that superimposes molecule i on molecule j. This transformation is expressed by the combination of a rotation matrix C_{ij} and a translation vector $\mathbf{d}_{ij}$:

$$\mathbf{x}_{ij} = C_{ij}\,\mathbf{x} + \mathbf{d}_{ij} \qquad (2)$$

The averaging operation in Equation 1 can be interpreted as taking each point within the mask M_i, fetching densities for each of the N copies of molecule j, averaging and then repeating the process for each of the molecules i.

The equations relating structure factors are derived by taking the Fourier transform of both sides of Equation 1. After some manipulations that will not be reproduced here (but can be found, for instance, in [11]) the following is obtained:

$$\mathbf{F}_{avg}(\mathbf{h}) = (U/NV)\,\Sigma_{\mathbf{k}}\,\mathbf{F}(\mathbf{k})$$
$$\Sigma_{i=1,N}\,\Sigma_{j=1,N}\,\exp(-2\pi i\,\mathbf{k}\cdot\mathbf{d}_{ij})\,\mathbf{G}_i(\mathbf{h}-C_{ij}^{T}\mathbf{k}) \qquad (3)$$

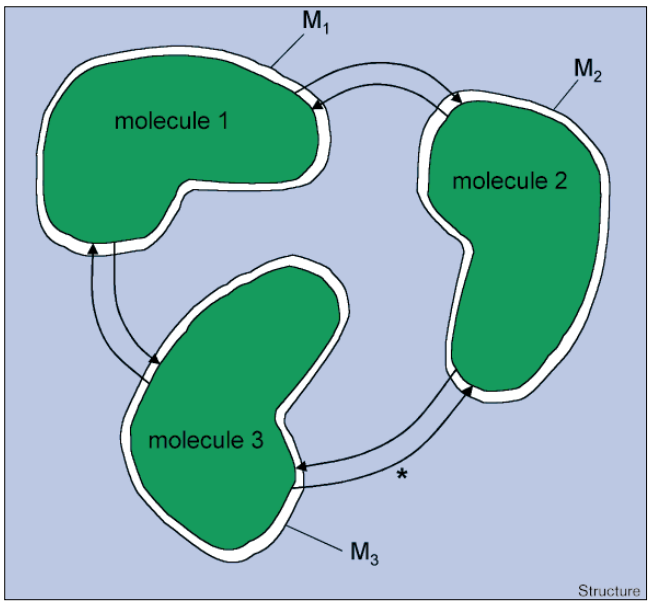
in which $\mathbf{G}$ is the well-known G function or interference function obtained as the Fourier transform of the mask M:

$$\mathbf{G}_i(\mathbf{l}) = (1/U)\int M_i(\mathbf{x})\exp(2\pi i\,\mathbf{l}\cdot\mathbf{x})\,d\mathbf{x} \qquad (4)$$

and C_{ij}^{T} is the transpose of the matrix C_{ij}.

In words, every structure factor resulting from averaging can be expressed as the weighted sum of the structure factors with which it interferes (including itself). (Note that the weighting factors are complex, so they introduce phase shifts.) Rules of thumb about the behaviour of the G function can be obtained by approximating the molecular envelope as a sphere, so that the G function can be expressed analytically. The properties of this function are such that it only assumes appreciable values near the origin (i.e. when $|\mathbf{h}-C_{ij}^{T}\mathbf{k}| \sim 0$ or when $\mathbf{h} \sim C_{ij}^{T}\mathbf{k}$). What does this tell us? Firstly, the obvious, namely that solvent flattening is simply a special case of averaging (only one NCS-related molecule, with C_{11} the identity matrix I and $\mathbf{d}_{11}$ the null translation), and that it is based on the same phenomenon of structure-factor interference. In this case, every reflection $\mathbf{h}$ interferes with other reflections $\mathbf{k}$ that are close to $\mathbf{h}$. In the case of general NCS, every reflection interferes not only with reflections close to it in reciprocal space, but also with other reflections near points related by the inverse NCS rotations. The more NCS-related copies there are of a molecule, the more regions of reciprocal space that are sampled and the more powerful the averaging will be. Secondly, the formula also tells us that averaging becomes more powerful as the G function covers a larger number of neighbouring reflections in each NCS-related volume of reciprocal space. This will happen when the mask can be defined with higher resolution detail. It will also happen when the volume U covering one molecule is small compared to the volume of the asymmetric unit, V. In fact, this explains why solvent flattening (onefold averaging) becomes more powerful as the solvent content increases.

Figure 1



Schematic illustration of NCS averaging, for the general case of improper NCS. Solvent regions (blue) are flattened (i.e. set to the average solvent density before averaging). Within the masks (M_i), electron density for the molecules (green) is replaced by the average of all copies of density related by NCS operations, indicated by the arrows connecting the molecules. An asterisk marks one such arrow, which is the transformation $\mathbf{x}_{32} = C_{32}\mathbf{x} + \mathbf{d}_{32}$ that superimposes molecule 3 on molecule 2.

Gamma correction

In Equation 3, one of the major contributions to a reflection is that of the reflection itself. Although this has been appreciated since the earliest days, it is only recently that this observation has provided a solution to a long-standing problem: how can the phases from averaging or solvent flattening be combined with the original phase information? The modified map is highly correlated to the input map, so it does not provide fully independent phase information, and phase combination is not really justified. On the other hand, Wang argued [12] that much of the power of solvent flattening in single isomorphous replacement comes from the new phase reinforcing one of the two possible phase choices, which requires phase combination. Abrahams [13] realised that, if the contribution of the reflection to itself were subtracted, what remained would be the independent information coming from other reflections, which could legitimately be combined with the starting phase information. He called the contribution of the structure factor to itself γ, hence the name gamma correction. He demonstrated

further that gamma correction could be carried out in real space, by a simple change in the density modification procedure. The factor γ is obtained by separating the $\mathbf{h} = \mathbf{k}$ and i = j terms in Equation 3, to get:

$$\mathbf{F}_{avg}(\mathbf{h}) = (U/V)\ \mathbf{F}(\mathbf{h}) +$$

$$(U/NV)\ \mathbf{F}(\mathbf{h})\ \Sigma_{i=1,N}\ \Sigma_{j\neq i}\ \exp(-2\pi i\ \mathbf{h}\cdot\mathbf{d}_{ij})$$

$$\mathbf{G}_i(\mathbf{h}-C_{ij}^{T}\mathbf{h}) +$$

$$(U/NV)\ \Sigma_{\mathbf{k}\neq\mathbf{h}}\ \mathbf{F}(\mathbf{k})\ \Sigma_{i=1,N}\ \Sigma_{j=1,N}\ \exp(-2\pi\ i\ \mathbf{k}\cdot\mathbf{d}_{ij})$$

$$\mathbf{G}_i(\mathbf{h}-C_{ij}^{T}\mathbf{k}) \qquad (5)$$

so that γ is equal to U/V (i.e. the ratio between the volume of one copy of the molecule and the volume of the asymmetric unit). (If there is a significant rotational component to the NCS operators, $\mathbf{G}_i(\mathbf{h}-C_{ij}^{T}\mathbf{h})$ will be small so the $\mathbf{G}_i$-weighted contributions from $\mathbf{F}(\mathbf{h})$ will be insignificant.) The terms in Equation 5 that contain new information can be collected to define a structure factor, $\mathbf{F}_{new}(\mathbf{h})$, on the same scale as $\mathbf{F}(\mathbf{h})$ and $\mathbf{F}_{avg}(\mathbf{h})$.

$$\mathbf{F}_{avg}(\mathbf{h}) = \gamma\ \mathbf{F}(\mathbf{h}) + (1-\gamma)\ \mathbf{F}_{new}(\mathbf{h}) \qquad (6)$$

$\mathbf{F}_{new}$ is a more appropriate source of phase information for phase combination than $\mathbf{F}_{avg}$. We can solve for $\mathbf{F}_{new}$ to get:

$$\mathbf{F}_{new}(\mathbf{h}) = (1/(1-\gamma))\ \mathbf{F}_{avg}(\mathbf{h}) - (\gamma/(1-\gamma))\ \mathbf{F}(\mathbf{h}) \qquad (7)$$

Equation 7 can be rearranged to show that $\mathbf{F}_{new}$ can be obtained by overshifting in the averaging step.

$$\mathbf{F}_{new}(\mathbf{h}) = \mathbf{F}(\mathbf{h}) + (1/(1-\gamma))\ (\mathbf{F}_{avg}(\mathbf{h}) - \mathbf{F}(\mathbf{h})) \qquad (8)$$

In other words, gamma correction corresponds to multiplying the structure factor change that would be obtained by straight averaging and solvent flattening by a factor of $1/(1-\gamma)$. In real space, Equation 8 becomes:

$$\rho_{new}(\mathbf{x}) = \rho(\mathbf{x}) + (1/(1-\gamma))\ (\rho_{avg}(\mathbf{x}) - \rho(\mathbf{x})) \qquad (9)$$

In the case of solvent flattening (onefold averaging), gamma correction can therefore be seen to correspond to solvent flipping [14] — instead of being flattened, the solvent density is flipped so that points with a higher than average density are assigned lower than average density, and *vice versa*. When the solvent content is 50%, γ is 0.5 and the factor $1/(1-\gamma)$ is 2, which means that the solvent is flipped exactly; deviations from mean solvent density are equal in magnitude but opposite in direction to those in the input map. The theoretical understanding of gamma correction, expressed in Equation 9, allows one to choose an optimal 'flipping factor' on the basis of the solvent content, instead of optimising it by trial and error. In the case of averaging, gamma correction corresponds to using a greater contribution from the other NCS-related densities in computing the average.

It should be noted that gamma correction is strictly justified only for the first step of averaging, because in subsequent steps the original information from a reflection will come back from its neighbours. Nonetheless, practical experience shows that it overcomes most of the problems associated with phase combination.

Translational NCS

Equation 5 leads us to another important observation, which applies in the case of purely translational NCS (i.e. in which molecules are related by an operator that does not involve any rotation, such that $C_{ij} = I$). As discussed below, translational NCS is quite commonly encountered. The presence of such NCS will not introduce interference with reflections in different parts of reciprocal space, because $\mathbf{k} \sim \mathbf{h}$. But there will be new contributions of the reflection to itself. Mathematically, this can be approached by assuming twofold translational NCS in the gamma correction equation, and carrying out some manipulations to obtain the following:

$$\mathbf{F}_{avg}(\mathbf{h}) = \gamma\ \mathbf{F}(\mathbf{h})$$

$$+ (U/2V)\ \Sigma_{\mathbf{k}\neq\mathbf{h}}\ \mathbf{F}(\mathbf{k})\ \mathbf{G}_1(\mathbf{h}-\mathbf{k})\ \{[1 + \exp(2\pi i\ \mathbf{h}\cdot\mathbf{d}_{12})]$$

$$[1 + \exp(-2\pi i\ \mathbf{k}\cdot\mathbf{d}_{12})]\} \qquad (10)$$

in which

$$\gamma = (U/V)\ [1 + \cos(2\pi\ \mathbf{h}\cdot\mathbf{d}_{12})] \qquad (11)$$

and $\mathbf{d}_{12}$ is the translation vector between the two NCS-related molecules.

The strongest reflections will tend to be those in which $\mathbf{h}\cdot\mathbf{d}_{12}$ is close to an integer, so that the two molecules scatter in phase. In such cases, $\mathbf{F}(\mathbf{h})$ makes a larger contribution to itself in the averaged structure factor, leaving less to come from neighbouring reflections. The most extreme case is when the translation $\mathbf{d}_{12}$ is composed of translations of zero or half of the unit cell edges; then the contributions of the two molecules will cancel for half of the reflections, and they will add up in phase for the other half. Additionally, the contributions will cancel for half of the neighbouring reflections $\mathbf{k}$. In this special case, averaging is equivalent in power to straight solvent flattening. This can be understood in real space because, for reflections that add up in phase, the two copies of the density will be affected in exactly the same way by phase errors. The two copies of density, therefore, do not give truly independent information.

Phase extension

Density modification is often used to propagate phase information from low resolution to a higher resolution that is typically near the diffraction limit. An examination of Equation 3 leads to an important point: one should be cautious about extending phases in large steps, because the G function can only spread phase information over a short distance in reciprocal space [15]. Rules of thumb for conservative phase extension steps can be obtained by considering the analytical G function obtained by approximating the envelope as a sphere.

Practical aspects

In this section, we shall discuss a number of practical aspects related to averaging. In general, the requirements for averaging are:

1. Reasonably complete diffraction data.

2. Non-crystallographic symmetry, or different crystal forms (or both).

3. Initial phases (or a map). The problem of obtaining phases is obviously not specific for averaging and will not be discussed further here.

4. Operators (rotation matrices and translation vectors) relating the various molecules (or domains) inside the asymmetric unit or (in the case of multiple-crystal form averaging) between different crystal forms.

5. A molecular envelope or mask, which is a simple binary 3D function (in real space) indicating the extent of space occupied by one molecule or assembly.

6. Appropriate software. There are many programs available nowadays (Table 2), and, although there are differences in implementation details, they essentially use similar algorithms and all do a good job. The discussion here will largely be independent of the particular choice of program.

Completeness of data

Because the success of averaging depends on information being spread among reflections, the more complete the data the better. Rather than leaving the amplitude of a missing observation as zero, however, it is better to replace it with the structure factor that was extrapolated from other reflections in the previous cycle of averaging. A particularly extreme use was made of such 'amplitude extension' in determining the structure of the trypanosomal glyceraldehyde phosphate dehydrogenase; sixfold NCS was exploited to fill in 63% of the observations, which were missing in a Laue data set [16]. If a molecular replacement model is available, $D\mathbf{F}_c$ (defined in [17]) provides the best guess for the true structure factor in the absence of a measured amplitude.

Table 2

Programs and program suites for molecular averaging.

Program	Author(s)	Reference
AVGSYS	Smith and Hendrickson	[51]
DEMON	Vellieux and collaborators	[27]
DM	Cowtan	[52]
ENVELOPE	Rossmann	[53]
GAP	Stuart and Grimes	[54]
MAGICSQUASH*	Schuller	[55]
MAIN	Turk	[56]
PHASES	Furey	[57]
RAVE†	Kleywegt and Jones	[10, 29]
SKEWPLANES	Bricogne	[9]
SOLOMON	Abrahams and Leslie	[14]
SQUASH	Zhang, Cowtan and Main	[58, 59]

*Derived from SQUASH. †Derived from A [10].

NCS or not?

Whether or not one has NCS (and, if so, how many copies of the molecule there are per asymmetric unit) can often already be deduced once the cell constants and space group of the very first crystal are known, or at least once the first data set has been collected. Typical sources of information include:

1. Values of V_M [18] calculated assuming different numbers of copies of the molecule, possibly combined with knowledge of the oligomer state under physiological conditions or in solution. One should keep in mind, however, that (part of) the symmetry may be crystallographic. For instance, there are many examples in which the monomeric units of a functional dimer are related by a crystallographic twofold rotation axis.

2. The self-rotation function, if it shows one or more clear solutions. One should keep in mind that (part of) the NCS may not involve a rotational component, however.

3. The native Patterson map which, for the above reason, should be calculated as soon as a data set has been collected. If (some) molecules are related by a pure translation, it will give rise to a large number of identical intermolecular vectors, and hence a considerable off-origin peak in the native Patterson. If the translation is pure, this peak will be comparable in height and volume to the origin peak; if a small rotational component is present (e.g. ~3–6°), the peak will be correspondingly weaker, and it may be necessary to restrict the upper resolution limit of the Patterson map to a value between 6 and 10 Å to observe the peak.

If these methods do not lead to a conclusive answer, the analysis has to be postponed until some form of phase information is available. For instance:

1. A model of anomalous or heavy atoms, from which the number of copies of a molecule may be detectable by analysis of the sites (by looking for sets of atoms with similar distances); however, heavy atom binding sites can be perturbed by crystal packing, so they may not reflect the NCS.

2. Manual or automated inspection of the initial map may reveal similar shaped pieces of density or larger volumes of ordered density than can be accounted for by a single molecule. Regions of ordered density can be highlighted most easily by constructing Wang–Leslie envelopes [12,19]. Because such an envelope covers a volume equal to that assumed for the protein part of the unit cell, an envelope should be made corresponding to each plausible number of NCS-related copies of the protein.

3. Molecular replacement, in which the packing of the solutions will often indicate whether or not there is room for another molecule.

Types of NCS

There are three types of NCS, and it can be of practical importance to know which one applies to a particular protein crystal.

1. Proper NCS. This is purely rotational NCS: the NCS operators form a closed group and the N-mer is invariant under all operations of this group. Examples are dimers related by a twofold axis without any screw component and tetramers related by *222* symmetry. The closed group property makes the definition of the molecular envelope simpler — one can simply construct an envelope that covers the entire N-mer, without worrying about the precise delineation of the borders between one molecule and the next. One has to be careful, however, not to assume the NCS to be proper when it is not, because small deviations are not unusual and they can lead to considerable problems in the averaging procedure (see Tête-Favier *et al.* [20], for a case in point).

2. Purely translational NCS. This is not detectable from a self-rotation function, but will easily be detected in a native Patterson map. If two molecules are related by a translation vector (t_x,t_y,t_z), purely translational NCS will give rise to a large off-origin peak at position (t_x,t_y,t_z) in the native Patterson map. This happens, for instance, if a non-crystallographic rotation axis runs parallel to a crystallographic rotation axis with the same rotational component (e.g. twofold axis parallel to a two, four or sixfold axis). As the statistical study by Wang and Janin [1] showed, this is extremely common. In this case, the Patterson peak tells one where the NCS rotation axis is located in the plane perpendicular to the crystallographic symmetry axis. One also has to keep in mind that translational components equal to half of a unit-cell edge will give rise to systematic absences, which will complicate the determination of the space group. In addition, it was shown recently [21] that pairs of enantiomorphic space-groups (e.g. $P6_1$ and $P6_5$) cannot be distinguished if there is a non-crystallographic rotation axis parallel to the crystallographic screw axis.

3. Improper NCS. This is the general case, in which molecules are related by a combination of a rotation and a translation.

Common to all three types of NCS is that they are local (i.e. the symmetry only applies within the asymmetric unit, but does not extend to the rest of the crystal). For this reason, NCS is sometimes referred to as local symmetry, as opposed to crystallographic symmetry (which applies to the entire crystal).

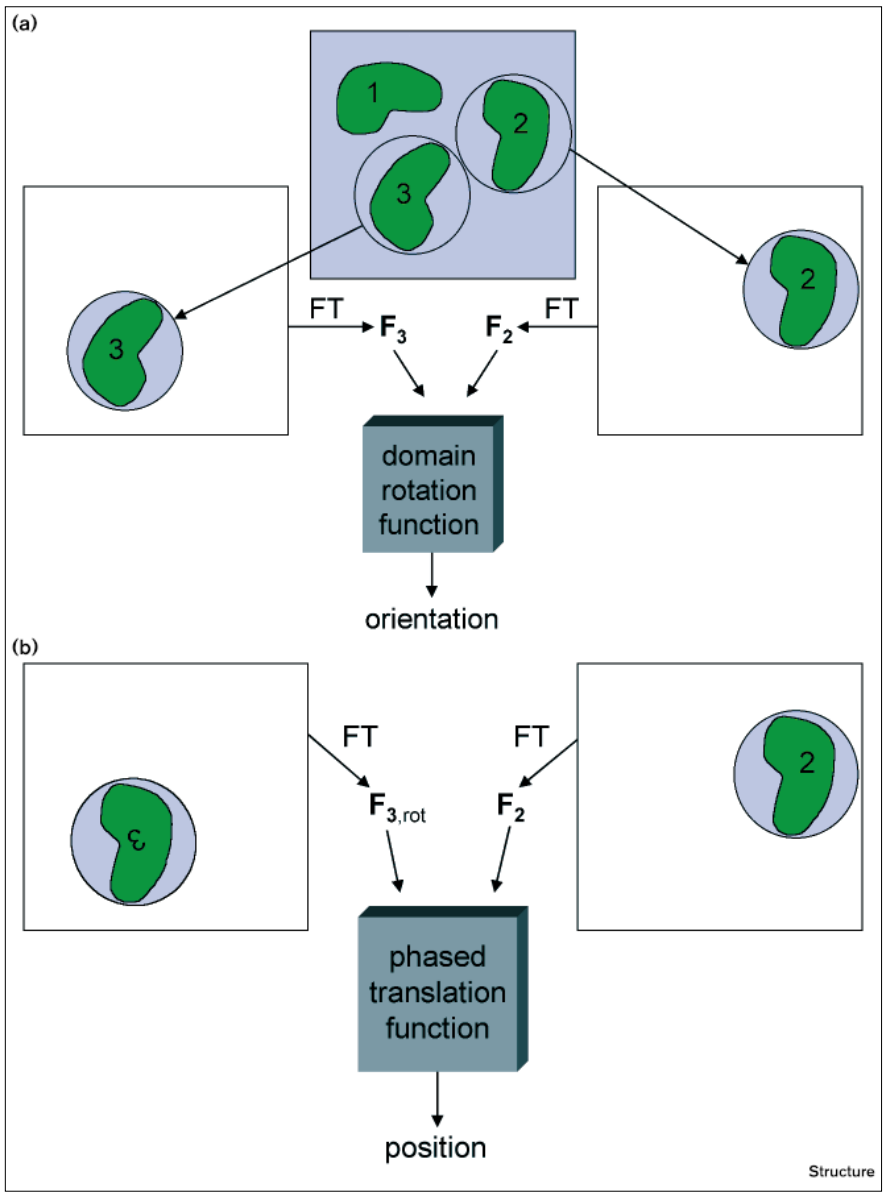
In addition to NCS within a crystal, different crystal forms also supply phase information. Although it is best to have an unrelated crystal form, a non-isomorphous crystal of the same space group also samples the molecular transform differently and the less isomorphous the better. Apart from the difficulty of getting initial phase information, it might even be more valuable to have a non-isomorphous derivative than a perfectly isomorphous one!

NCS operators

The operators that relate copies of a molecule can be derived in various ways: firstly, in molecular replacement cases, the strategy is simple — one just superimposes the various molecules after solving the rotation and translation functions and carrying out rigid-body refinement, which immediately yields the NCS operators. Secondly, information about the rotational part of the NCS operators can generally be obtained from a self-rotation function, which does not require a model or experimental phases. Thirdly, if one has experimental phases, but no model, real-space molecular replacement techniques can be used to find the NCS operations that relate the densities from copies of the molecules. These techniques, the domain rotation function [22] and the phased translation function [22,23] are illustrated schematically in Figure 2. They have been used in solving the structures of pertussis toxin [24] and the Fyn SH3 domain [25]. The same techniques can be used to find operators that relate molecules in different crystal forms (Eleanor J Dodson, personal communication). Finally, sometimes it is possible to derive the operators by comparing the positions of heavy or anomalous atoms.

In most cases, the initial operators are inaccurate, and it is important to optimise them before use. Typically, the

Figure 2



Schematic illustration of real-space molecular replacement techniques. **(a)** The domain rotation function [22] exploits phase information to improve the signal-to-noise ratio in a rotation search. Spheres of density corresponding to two NCS-related molecules (molecules 2 and 3 in this example) are cut out and placed in P1 unit cells, and structure factors are obtained by map inversion. Alternatively, the required structure factors can be obtained in a single step in reciprocal space using the program GHKL (Liang Tong, personal communication). The locations of the molecules can be inferred from a Wang–Leslie envelope [12,19] which highlights regions of ordered density. A cross-rotation function computed with the two sets of structure factors gives the relative orientation, which can be expressed as the matrix C_{32}. **(b)** After the domain rotation function, the sphere of density for molecule 3 is rotated and placed in a P1 unit cell with the same dimensions as the one containing molecule 2. Structure factors derived from this density can then be used in a phased translation function [22,23], which gives the translational part of the NCS operator, $\mathbf{d}_{32}$.

operator is adjusted in small steps to maximise the correlation coefficient between the density inside the mask and the density in the envelope of the NCS-related molecule [10]. This technique was used as early as 1976 to improve the operator between two different crystal forms of hexokinase [26].

Masks

Masks can be obtained in myriad ways. In molecular replacement cases, the task is trivial, once solutions have been obtained. One generates a mask that covers the entire model (e.g. within a radius of 2–3 Å around each atom), and then adjusts it to account for differences between the search model and the actual structure to be solved. When experimental phase information is available, one may skeletonise the initial map and generate a mask around the part of the skeleton that is assumed to constitute one monomer, or one may build a quick-and-dirty initial model and generate a mask around that.

A local correlation map [27] can be constructed to show where a particular set of NCS operators is obeyed, which then provides a mask for the molecule or assembly governed by that set of operators. The map is computed over the entire region in which the NCS operators might apply, which is not generally the same as the asymmetric unit. The value at each point represents the correlation coefficient between spheres of electron density related by the

set of NCS operators. If the set applies to that point, the correlation will be high, otherwise it will be close to zero. The signal-to-noise ratio in the correlation coefficients depends on the size of the spheres; a sphere radius of about 1.5 times the minimum d-spacing to which the phases are trusted works well in practice.

A related approach is to execute one cycle of mask-less averaging — the NCS operators are applied to all points in the volume in which they might apply. As the operators are only valid for one of the molecules, mask-less averaging is expected to improve the contrast between the volume of that molecule and the rest of the asymmetric unit (where the density will tend to be obliterated). The mask can then be determined with the traditional Wang–Leslie method [12,19].

Onefold masks, which will prevent solvent flattening, should be added to any regions of ordered density that are not covered by an existing mask. This can arise when there are local breakdowns in NCS, and it can be detected by comparison of the combined NCS masks with a Wang mask.

Initial masks, irrespective of the method used to generate them, invariably need to be improved in order to fill internal voids, to remove parts that are isolated from the bulk of the mask, to remove overlap between the mask and the copies generated from it under crystallographic and/or non-crystallographic symmetry and to make sure that all atoms in the molecule are covered by the mask. The last procedure can be done interactively [10], for example using the program O [28], if the model does not yet include all atoms of the molecule. The other tasks can be accomplished with mask editing and improvement programs, such as MAMA [29].

Applications

In this section, we shall highlight some examples of the most significant applications of molecular averaging, used between the late 1960s and the mid 1990s. In the subsequent section, we shall discuss a few cases from our own laboratories in more detail.

The first reported applications of molecular averaging (in real space) date from 1967, chymotrypsin at 2 Å [30] and haemoglobin at 5.5 Å [31]. It was not until 1974, however, that the first case occurred in which use of averaging actually made the difference between solving and not solving the structure. This was for the structure of D-glyceraldehyde 3-phosphate dehydrogenase [32], fittingly solved by Michael Rossmann. The first documented example of multiple crystal form averaging was as early as 1976, when Fletterick and Steitz used two crystal forms to solve the structure of yeast hexokinase [26]. They transformed the density from one crystal form to another, calculated phases from the transformed density and combined them with MIR phases. As part of this work, they also developed software to optimise the operator between the two crystal forms.

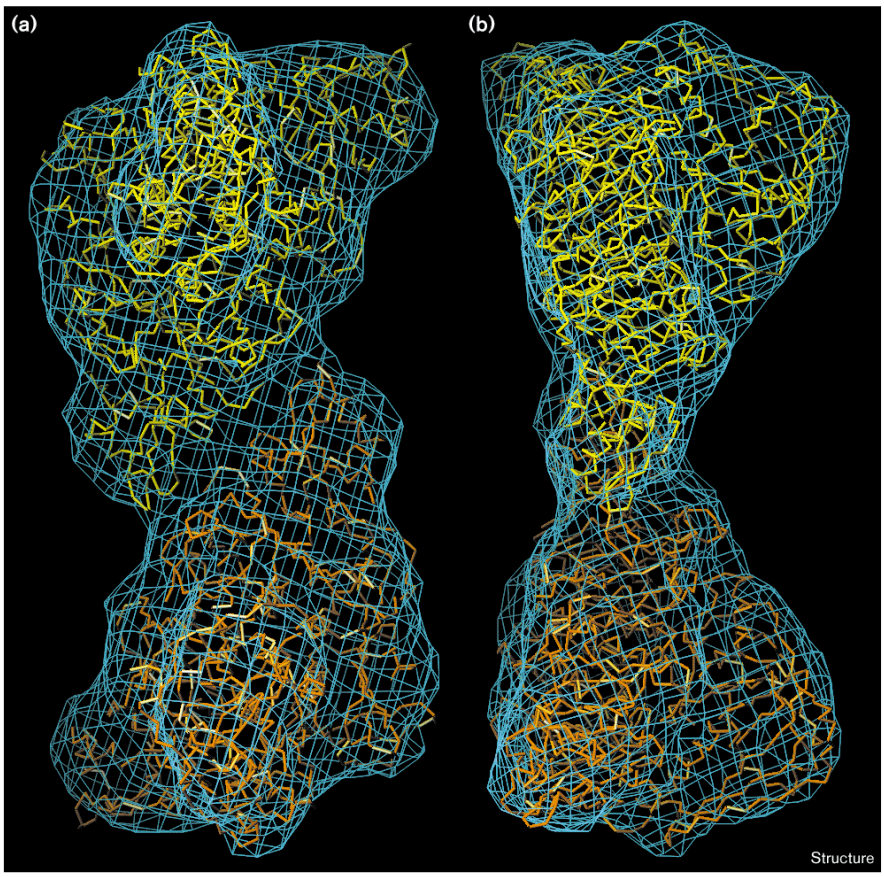
Use of molecular averaging has been a major factor in enabling the structures of viruses to be solved. Indeed, a long-standing but as yet unfulfilled goal is to use the redundancies inherent to virus structures for *ab initio* phasing. In 1978, the first two virus structures, both solved using Bricogne's programs, were reported simultaneously: tobacco mosaic virus (TMV; a disk-shaped virus with 17-fold NCS [33]) and tomato bushy stunt virus (the first icosahedral virus structure [34]). The structure determination of TMV also incorporated the first application of phase extension (from 3.2 to 2.8 Å). An interesting lesson was learned during the structure determination of another virus, MS2 [35]. The phasing started with a model derived from the structure of southern bean mosaic virus (which in fact turned out to be unrelated to the MS2 structure), with phase extension from 13 to 3.4 Å. The resulting map was not interpretable, however. Heavy-atom derivatives therefore were used, the heavy-atom sites located and the phases extended from 8.8 to 3.3 Å. *A posteriori* analysis revealed that the averaging had converged to a set of phases related to the correct phases by a shift of 180° (i.e. the Babinet opposite of the correct phases). At this stage, it was realised that phase extension from low to high resolution can lead to four sets of phases: the correct phases α, their enantiomorphs $-\alpha$, their Babinet opposite $\alpha + 180°$ or the Babinet opposite of their enantiomorphs, $-\alpha + 180°$ [36]. Essentially the same method was used to solve the structures of ϕX174 [37] and cowpea chlorotic mottle virus [38], but in these cases it led to useful maps.

Although phase extension was first used in the structure determination of TMV, a spectacular application of the technique occurred in 1984, in the structure determination of haemocyanin [39]. Exploiting the sixfold NCS, Gaykema *et al.* were able to extend phases from 4.0 to 3.2 Å. At the end of this procedure, roughly half of all reflections to 3.2 Å (31,000 out of 64,000) had no heavy-atom phase information (i.e. their phases had been obtained purely due to the NCS).

The structure determination of GroEL [40] ten years later was interesting for another reason, namely because it started from essentially random phases. A mask covering all seven monomers was used as the starting map, corresponding to a resolution of perhaps 8 Å. The phasing started at 8 Å and phases were slowly extended to 2.7 Å, with periodic improvement of the NCS operators.

An interesting application of multiple crystal averaging led to a much improved structure of HIV-1 reverse transcriptase in complex with a nevirapine analogue [41]. It was found that controlled dehydration of the crystals improved

Figure 3



Application of local correlation to determine a molecular envelope for pertussis toxin. A local correlation map was computed for the non-crystallographic twofold axis in pertussis toxin, using the initial MIR map as input and a sphere radius of 9 Å. A Cα trace from the final refined model is superimposed for comparison. **(a)** A view of the local correlation map down the local twofold axis. **(b)** Same as (a), but looking perpendicular to the local twofold axis.

their diffraction quality. Data sets were collected at different stages of dehydration, and averaging between these data sets (and between domains within the crystals) gave a significant improvement in map quality and a reduction of model bias.

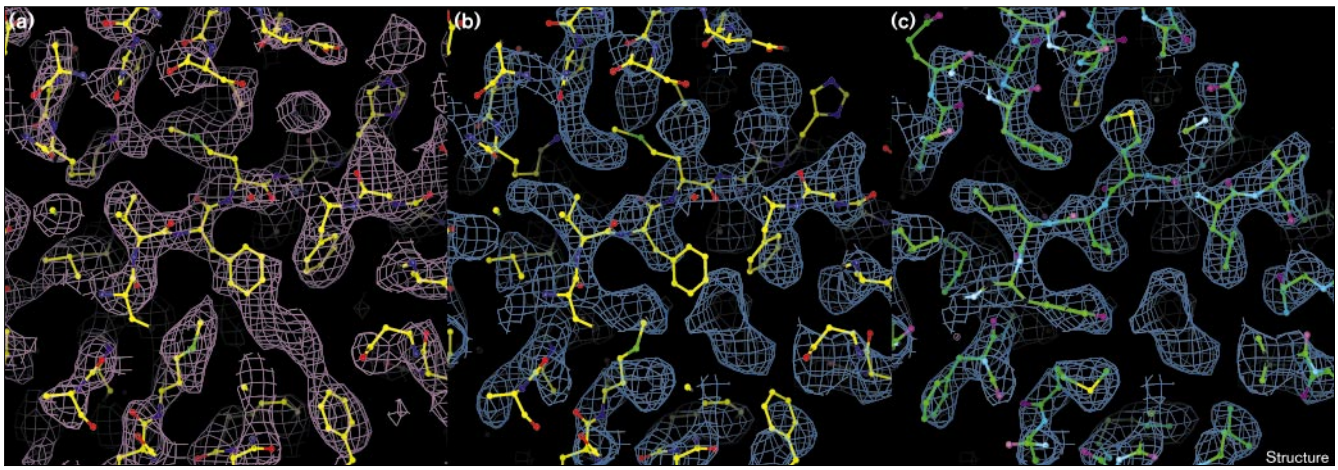
The determination of the structure of human protective protein [42] provides a striking demonstration of how twofold NCS can be exploited in a careful bootstrapping procedure to overcome the deficiencies of a molecular replacement model. The starting model was composed of a core of about two-thirds of the molecule, and sidechains were retained for only the 36% of the residues that are identical in the search model, wheat serinecarboxypeptidase. Bootstrapping started with definition of a generous envelope to cover the missing domain, followed by cycles of averaging, model building and tightening of the envelope in light of new structural detail. Model refinement was delayed until the model was nearly complete, to avoid problems of overfitting. This structure solution probably succeeded because of the high solvent content (63%), the fact that the NCS was rotational, the quality of the data (precise and reasonably complete to 2.2 Å) and, of course, the care taken by the investigators.

Case studies

Taq polymerase

An unsuccessful attempt to determine the structure of an N-terminal deletion mutant of *Taq* polymerase provides a cautionary tale about how and when NCS averaging can fail. Crystals were grown in Edmonton from an expression construct that differed slightly from the one used by Waksman and co-workers to solve the structure [43]. The crystal packing is similar, but in the crystals of the Edmonton construct, a crystallographic twofold axis has shifted slightly to become non-crystallographic. This leads to a doubling of the unit cell and, hence, twofold translational NCS with a translation of about half of a unit cell edge (NEC Duke and RJR, unpublished data). The structure could be solved by molecular replacement, using a model derived from the Klenow fragment of *Escherichia coli* pol I, but the model was poor and incomplete, and maps were too poor to allow productive rebuilding. Twofold averaging failed to yield a significant improvement. Failure can probably be blamed on the nature of the NCS; a translation of half a unit cell edge is the worst case, as discussed above, in which the expected improvement is no better than what would be obtained by solvent flattening alone.

Figure 4



The power of molecular averaging to remove model bias. **(a)** $2F_o–F_c$ map calculated from a backwards traced model of α_{2u}-globulin, which was refined at 3.0 Å resolution without NCS restraints to yield a free R value of 47%. Note that model bias makes the density look almost convincing for several residues, even though the model is completely wrong. **(b)** The fourfold averaged map clearly does not fit the model from which the starting map was calculated. **(c)** Superimposing the correct 2.5 Å model on the averaged map reveals an excellent fit.

Pertussis toxin

The crystal structure of pertussis toxin was solved using MIR phases from poorly isomorphous derivatives, in combination with twofold averaging and solvent flattening [24]. The initial phases had a mean figure of merit of 0.44, and phasing power dropped dramatically beyond 6 Å resolution, so the initial map was essentially uninterpretable. The high solvent content of about two-thirds the unit-cell volume and the fact that the local twofold does not parallel any of the crystallographic symmetry axes were probably essential for success.

The unit-cell volume was sufficient to contain two or three pertussis toxin molecules of 105 kDa in the asymmetric unit. The absence of significant non-origin peaks in a low-resolution native Patterson map ruled out translational NCS. Surprisingly, there are also no significant peaks in self-rotation functions and, even now that we know the correct structure, we cannot find a peak at the correct rotation. (This might be explained by the significant difference in overall B values among the molecules that emerged from the refinement.) The NCS operations could not be determined from heavy-atom positions, because these do not obey the NCS; instead, they were determined using real-space molecular replacement techniques (see Figure 2). The rotational component was determined first with a domain rotation function. The locations of two unique molecules were clear in a Wang–Leslie envelope; therefore, two spheres of density could be isolated and an unambiguous peak of 9.3 times the root mean square (rms) function value was obtained in a cross-rotation function. When one of the spheres was rotated and used in a phased translation function, an even less ambiguous peak was found (70 times the rms deviation). The NCS operation, defined in this way, turned out to be very accurate and was not changed significantly by subsequent refinement.

Because the NCS operator defined a proper twofold rotation, only a single envelope covering the dimer was required. This was obtained by computing a local correlation map with a 9 Å sphere radius, and choosing a contour level that enclosed a slightly generous 44% of the asymmetric unit (Figure 3). Averaging started from 6 Å resolution, with gradual phase extension to 3.5 Å. As the phases improved, the mask was updated with local correlation maps computed with smaller radii (eventually reduced to 7 Å), and contoured to enclose a smaller volume. The final map was sufficient to allow a refinable model to be built, although the availability of structural homologues was a great help.

Shiga-like toxin B subunit–trisaccharide complex

The Shiga-like toxin I B subunit is a symmetrical pentamer that binds to Gb_3 glycolipid molecules on cell surfaces. The structure of the B subunit in complex with a soluble trisaccharide analogue of Gb_3 was solved, using the unliganded B subunit [44] as a molecular replacement model [45]. The crystals diffract weakly to 2.8 Å resolution, but the data beyond 3.5 Å are very poor. A high degree of NCS in this structure was essential for obtaining interpretable results.

There are four pentamers in the asymmetric unit, two of which are related by a translation with a small rotational component of about 6°. The translational NCS was

revealed first by native Patterson maps computed with upper resolution limits of 8–10 Å. Although the pentamers have approximate fivefold symmetry, rigid-body refinement of the 20 monomers revealed that none of the fivefold axes is exact. Twenty separate envelopes therefore had to be defined from the atomic coordinates. NCS is obeyed most poorly in the trisaccharide-binding sites (three per monomer), which are affected by crystal packing, so an additional onefold envelope was defined to surround the 60 carbohydrate binding sites and protect them from solvent flattening. Carbohydrate density was not interpretable in the initial model-phased maps, but it was exceptionally clear in the averaged maps, in spite of the poor data.

Endoglucanase I

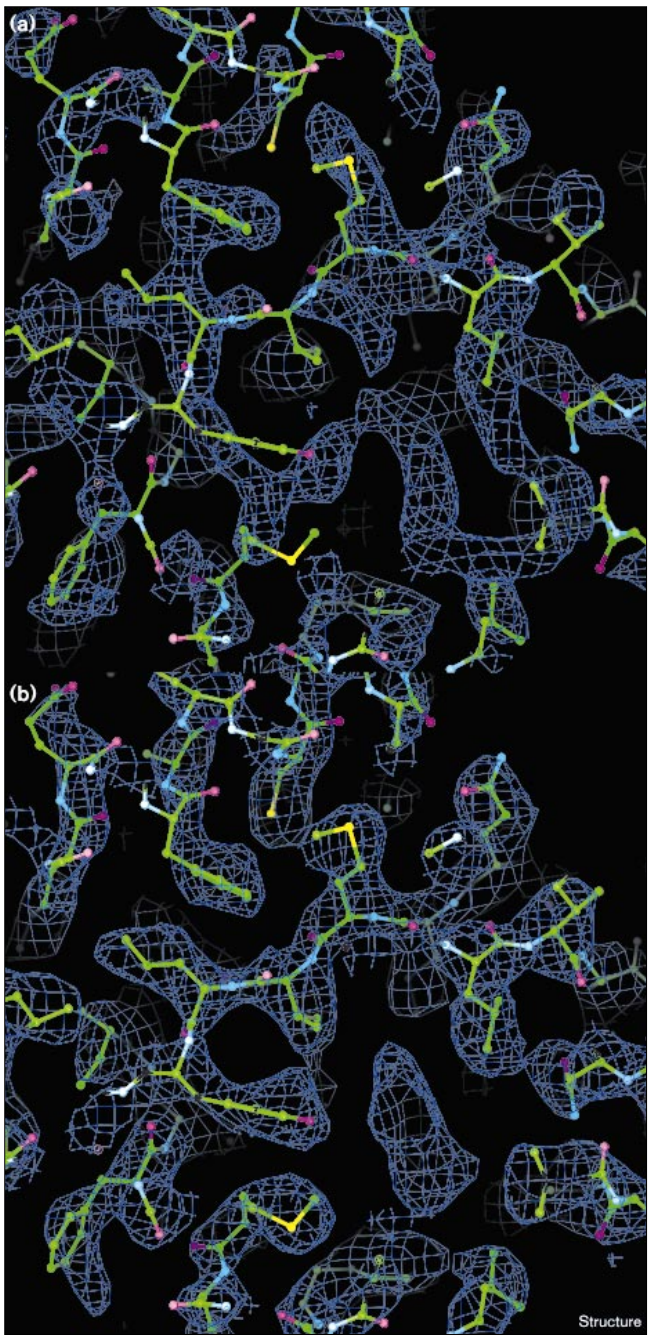
The structure determination of endoglucanase I [46] was complicated by the fact that only low-resolution data were available (initially to 4.0 Å, later to 3.6 Å). A tentative molecular replacement solution was obtained, the correctness of which could only be verified by means of molecular averaging. A polyalanine version of the search model was used to calculate a map that, not surprisingly, was poor and had little sidechain detail. Twofold averaging at 4.0 Å resolution, however, resulted in a much improved map that showed density for many bulky sidechains, as well as for some parts of the structure that differed from the search model. The twofold NCS was also essential for the refinement to succeed. (The final model and its averaged density are available as a collection of VRML 'worlds', courtesy of Tom Taylor, on the World Wide Web at URL http://alpha2.bmc.uu.se/vrml/.)

α_{2u}-Globulin

With the structure of α_{2u}-globulin (GJK *et al.*, unpublished data), which has fourfold NCS and was refined to 2.5 Å resolution, we have carried out some experiments in order to demonstrate the power of molecular averaging to remove model bias, and to investigate the power of averaging starting from near-random phases.

To demonstrate the ability of molecular averaging to overcome model bias (which arises if phases derived from an atomic model are used to calculate maps), we used a model of this protein that was intentionally traced backwards [47]. This model was refined to 3.0 Å resolution, without any NCS constraints or restraints (leading to a free R value of 47%). This refined model, which has very few scatterers in the right place, was used to generate a mask, to calculate NCS operators and to calculate a 3.0 Å map (Figure 4a). Subsequently, fourfold averaging was applied, and the resulting map obviously does not agree with the backwards traced model (Figure 4b). When the correct 2.5 Å model is superimposed, however, it clearly fits the averaged density very well, despite the low resolution and the poor starting phases (Figure 4c).

Figure 5



The power of molecular averaging when starting from near-random phases. **(a)** Map obtained from fourfold molecular averaging at 3.0 Å, starting from a map that consisted only of the joint envelopes of the four α_{2u}-globulin monomers. The correct 2.5 Å model is shown superimposed and clearly does not fit this map. **(b)** Map obtained from the same starting point as in (a), but here the phases were gradually extended from 8.0 to 3.0 Å. The correct 2.5 Å model is superimposed and obviously fits the averaged density very well.

As discussed above, the structure of GroEL was solved using sevenfold averaging procedures starting from near-random phases. To investigate if even the fourfold NCS of

α_{2u}-globulin might be sufficient to allow this, we generated a mask around the correct 2.5 Å model, and used a starting map that consisted only of the combined masks of the four monomers. Averaging starting at 3.0 Å does not lead to a correct set of phases (Figure 5a). If the averaging is started at 8.0 Å and the phasing is gradually extended to 3.0 Å, however, we again obtain a map that fits the 2.5 Å refined model extremely well (Figure 5b). Hence, in this case even mere fourfold NCS suffices to proceed from near-random 8 Å phases to correct 3 Å phases. Of course, this will only constitute a method to solve structures, when an algorithm is devised to construct a precise and accurate envelope with the same near-random phases!

Acknowledgements

GJK would like to thank T Alwyn Jones and Lars Liljas (Uppsala) for illuminating discussions and many pointers to relevant literature. RJR thanks Bart Hazes and Navraj S Pannu for stimulating discussions. We both thank our colleagues for allowing us to share unpublished results in the case histories. GJK is supported by the Swedish Foundation for Strategic Research (SSF), and its Structural Biology Network (SBNet). RJR is a Scholar of the Alberta Heritage Foundation for Medical Research. The work discussed in the case histories from RJR was supported by the Medical Research Council of Canada and an International Research Scholar award from the Howard Hughes Medical Institute.

References

1. Wang, X. & Janin, J. (1993). Orientation of non-crystallographic symmetry axes in protein crystals. *Acta Cryst. D* **49**, 505–512.
2. Kleywegt, G.J. (1996). Use of non-crystallographic symmetry in protein structure refinement. *Acta Cryst. D* **52**, 842–857.
3. Kleywegt, G.J. & Jones, T.A. (1995). Where freedom is given, liberties are taken. *Structure* **3**, 535–540.
4. Brünger, A.T. (1997). X-ray crystallography and NMR: complementary views of structure and dynamics. *Nat. Struct. Biol.* **4**, 862–865.
5. Bricogne, G. (1974). Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Cryst. A* **30**, 395–405.
6. Rossmann, M.G. & Blow, D.M. (1962). The detection of sub-units within the crystallographic asymmetric unit. *Acta Cryst.* **15**, 24–31.
7. Main, P. (1967). Phase determination using non-crystallographic symmetry. *Acta Cryst.* **23**, 50–54.
8. Colman, P.M. (1974). Noncrystallographic symmetry and the sampling theorem. *Z. Krist.* **140**, 344–349.
9. Bricogne, G. (1976). Methods and programs for direct-space exploitation of geometric redundancies. *Acta Cryst. A* **32**, 832–847.
10. Jones, T.A. (1992). A, yaap, asap, @#*? A set of averaging programs. In *Molecular Replacement.* (Dodson, E.J., Glover, S. & Wolf, W., eds.), pp 91–105, SERC Daresbury Laboratory, Daresbury, U.K.
11. Vellieux, F.M.D. & Read, R.J. (1997). Non-crystallographic symmetry averaging in phase refinement and extension. *Methods Enzymol.* **277**, 18–53.
12. Wang, B.C. (1985). Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol.* **115**, 90–112.
13. Abrahams, J.P. (1997). Bias reduction in phase refinement by modified interference functions: introducing the gamma correction. *Acta Cryst. D* **53**, 371–376.
14. Abrahams, J.P. & Leslie, A.G.W. (1996). Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Cryst. D* **52**, 30–42.
15. Arnold, E., *et al.*, & Rossmann, M.G. (1987). The structure determination of a common cold virus, human rhinovirus 14. *Acta Cryst. A* **43**, 346–361.
16. Vellieux, F.M.D., *et al.*, & Hol, W.G.J. (1993). Structure of glycosomal glyceraldehyde-3-phosphate dehydrogenase from *Trypanosoma brucei* determined from Laue data. *Proc. Natl. Acad. Sci. USA* **90**, 2355–2359.
17. Read, R.J. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Cryst. A* **42**, 140–149.
18. Matthews, B.W. (1968). Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497.
19. Leslie, A.G.W. (1987). A reciprocal-space method for calculating a molecular envelope using the algorithm of B.C. Wang. *Acta Cryst. A* **43**, 134–136.
20. Tête-Favier, F., Rondeau, J.M., Podjarny, A. & Moras, D. (1993). Structure determination of aldose reductase: joys and traps of local symmetry averaging. *Acta Cryst. D* **49**, 246–256.
21. Ha, Y. & Allewell, N.M. (1997). Equivalence of pairs of enantiomorphic space groups in the presence of non-crystallographic symmetry. *Acta Cryst. A* **53**, 400–401.
22. Colman, P.M., Fehlhammer, H. & Bartels, K. (1976). Patterson search methods in protein structure determination: β-trypsin and immunoglobulin fragments. In *Crystallographic Computing Techniques.* (Ahmed, F.R., Huml, K & Sedlacek, B., eds.), pp 248–258, Munksgaard, Copenhagen.
23. Read, R.J. & Schierbeek, A.J. (1988). A phased translation function. *J. Appl. Cryst.* **21**, 490–495.
24. Stein, P.E., Boodhoo, A., Armstrong, G.D., Cockle, S.A., Klein, M.H. & Read, R.J. (1994). The crystal structure of pertussis toxin. *Structure* **2**, 45–57.
25. Noble, M.E.M., Musacchio, A., Saraste, M., Courtneidge, S.A. & Wierenga, R.K. (1993). Crystal structure of the SH3 domain in human Fyn; comparison of the three-dimensional structures of SH3 domains in tyrosine kinases and spectrin. *EMBO J.* **12**, 2617–2624.
26. Fletterick, R.J. & Steitz, T.A. (1976). The combination of independent phase information obtained from separate protein structure determinations of yeast hexokinase. *Acta Cryst. A* **32**, 125–132.
27. Vellieux, F.M.D.A.P., Hunt, J.F., Roy, S. & Read, R.J. (1995). DEMON/ANGEL: a suite of programs to carry out density modification. *J. Appl. Cryst.* **28**, 347–351.
28. Jones, T.A., Zou, J.Y., Cowan, S.W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Cryst. A* **47**, 110–119.
29. Kleywegt, G.J. & Jones, T.A. (1994). Halloween ... masks and bones. In *From First Map to Final Model.* (Bailey, S., Hubbard, R. & Waller, D.A., Eds.), pp 59–66, SERC Daresbury Laboratory, Daresbury, U.K.
30. Matthews, B.W., Sigler, P.B., Henderson, R. & Blow, D.M. (1967). Three-dimensional structure of tosyl-chymotrypsin. *Nature* **214**, 652–656.
31. Muirhead, H., Cox, J.M., Mazzarella, L. & Perutz, M.F. (1967). Structure and function of haemoglobin. 3. A three-dimensional fourier synthesis of human deoxyhaemoglobin at 5.5 Ångstrom resolution. *J. Mol. Biol.* **28**, 117–156.
32. Buehner, M., Ford, G.C., Moras, D., Olsen, K.W. & Rossmann, M.G. (1974). Structure determination of crystalline lobster D-glyceraldehyde-3-phosphate dehydrogenase. *J. Mol. Biol.* **82**, 563–585.
33. Bloomer, A.C., Champness, J.N., Bricogne, G., Staden, R. & Klug, A. (1978). Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between subunits. *Nature* **276**, 362–368.
34. Harrison, S.C., Olson, A.J., Schutt, C.E., Winkler, F.K. & Bricogne, G. (1978). Tomato bushy stunt virus at 2.9 Å resolution. *Nature* **276**, 368–373.
35. Valegård, K., Liljas, L., Fridborg, K. & Unge, T. (1990). The three-dimensional structure of the bacterial virus MS2. *Nature* **345**, 36–41.
36. Valegård, K., Liljas, L., Fridborg, K. & Unge, T. (1991). Structure determination of the bacteriophage MS2. *Acta Cryst. B* **47**, 949–960.
37. McKenna, R., Xia, D., Willingmann, P., Ilag, L.L. & Rossmann, M.G. (1992). Structure determination of the bacteriophage ϕX174. *Acta Cryst. B* **48**, 499–511.
38. Speir, J.A., Munshi, S., Wang, G., Baker, T.S. & Johnson, J.E. (1995). Structures of the native and swollen forms of cowpea chlorotic mottle virus determined by X-ray crystallography and cryo-electron microscopy. *Structure* **3**, 63–78.
39. Gaykema, W.P.J., Hol, W.G.J., Vereijken, J.M., Soeter, N.M., Bak, H.J. & Beintema, J.J. (1984). 3.2 Å structure of the copper-containing, oxygen-carrying protein *Panulirus interruptus* haemocyanin. *Nature* **309**, 23–29.
40. Braig, K., *et al.*, & Sigler, P.B. (1994). The crystal structure of the bacterial chaperonin GroEL at 2.8 Å. *Nature* **371**, 578–586.
41. Esnouf, R., Ren, J., Jones, Y., Stammers, D. & Stuart, D. (1996). Removing bias from a model for HIV-1 reverse transcriptase by real-space averaging between different crystal forms. In *Macromolecular Refinement.* (Dodson, E., Moore, M., Ralph, A. & Bailey, S., Eds.), pp 153–161, CCLRC Daresbury Laboratory, Daresbury, U.K.
42. Rudenko, G., Bonten, E., d'Azzo, A. & Hol, W.G.J. (1996). Structure determination of the human protective protein: twofold averaging reveals the three-dimensional structure of a domain which was entirely absent in the initial model. *Acta Cryst. D* **52**, 923–936.

43. Korolev, S., Nayal, M., Barnes, W.M., Di Cera, E. & Waksman, G. (1995). Crystal structure of the large fragment of *Thermus aquaticus* DNA polymerase I at 2.5 Å resolution: structural basis for thermostability. *Proc. Natl. Acad. Sci. USA* **92**, 9264–9268.
44. Stein, P.E., Boodhoo, A., Tyrrell, G.J., Brunton, J.L. & Read, R.J. (1992). Crystal structure of the cell-binding B oligomer of verotoxin-1 from *E. coli*. *Nature* **355**, 748–750.
45. Ling, H., *et al*., & Read, R.J. (1998). The structure of the Shiga-like toxin I B-pentamer complexed with its receptor, Gb_3. *Biochemistry,* in press.
46. Kleywegt, G.J., *et al*., & Jones, T.A. (1997). The crystal structure of the catalytic core domain of endoglucanase I from *Trichoderma reesei* at 3.6 Å resolution, and a comparison with related enzymes. *J. Mol. Biol*., **272**, 383–397.
47. Kleywegt, G.J. & Brünger, A.T. (1996). Checking your imagination: applications of the free R value. *Structure* **4**, 897–904.
48. Wilson, C. & Agard, D.A. (1993). PRISM: automated crystallographic phase refinement by iterative skeletonization. *Acta Cryst. A* **49**, 97–104.
49. Zhang, K.Y.J. & Main, P. (1990). Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. *Acta Cryst. A* **46**, 41–46.
50. Zhang, K.Y.J. & Main, P. (1990). The use of Sayres equation with solvent flattening and histogram matching for phase extension and refinement of protein structures. *Acta Cryst. A* **46**, 377–381.
51. Bolin, J.T., Smith, J.L. & Muchmore, S.W. (1993). Considerations in phase refinement and extension: experiments with a rapid and automatic procedure. *Abstracts Am. Cryst. Assoc. Series 2* **21**, 51.
52. Cowtan, K. (1994). dm: an automated procedure for phase improvement by density modification. *CCP4/ESF-EACBM Newsletter on Protein Crystallography* 34–38.
53. Rossmann, M.G., *et al*., & Lynch, R.E. (1992). Molecular replacement real-space averaging. *J. Appl. Cryst.* **25**, 166–180.
54. Grimes, J.M. (1995). *Structural Studies of Bluetongue Virus*. D. Phil. Thesis, University of Oxford.
55. Schuller, D.J. (1996). MAGICSQUASH: more versatile non-crystallographic symmetry averaging with multiple constraints. *Acta Cryst. D* **52**, 425–434.
56. Turk, D. (1992). *Weiterentwicklung eines Programms für Molekülgraphik und Elektronendichte-Manipulation und seine Anwendung auf verschiedene Protein-Strukturaufklärungen*. Ph.D. Thesis, Technische Universität, München.
57. Furey, W. & Swaminathan, S. (1997). PHASES-95: a program package for processing and analyzing diffraction data from macromolecules. *Methods Enzymol.* **277**, 590–620.
58. Cowtan, K.D. & Main, P. (1993). Improvement of macromolecular electron-density maps by the simultaneous application of real and reciprocal space constraints. *Acta Cryst. D* **49**, 148–157.
59. Zhang, K.Y.J. (1993). SQUASH: combining constraints for macromolecular phase refinement and extension. *Acta Cryst. D* **49**, 213–222.