Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Simulation of modulated protein crystal structure and diffraction data in a supercell and in superspace

Jeffrey J. Lovelace,[a] Peter D. Simone,[a] Václav Petříček[b] and Gloria E. O. Borgstahl[a]*

[a]Eppley Institute for Research in Cancer and Allied Diseases, 987696 Nebraska Medical, Omaha, NE 68198-7696, USA, and [b]Institute of Physics, Academy of Sciences of the Czech Republic, Na Slovance 10, 182 21 Praha, Czech Republic

Correspondence e-mail: gborgstahl@unmc.edu

Received 29 November 2012
Accepted 17 February 2013

The toolbox for computational protein crystallography is full of easy-to-use applications for the routine solution and refinement of periodic diffraction data sets and protein structures. There is a gap in the available software when it comes to aperiodic crystallographic data. Current protein crystallography software cannot handle modulated data, and small-molecule software for aperiodic crystallography cannot work with protein structures. To adapt software for modulated protein data requires training data to test and debug the changed software. Thus, a comprehensive training data set consisting of atomic positions with associated modulation functions and the modulated structure factors packaged as both a three-dimensional supercell and as a modulated structure in (3+1)D superspace has been created. The (3+1)D data were imported into *Jana*2006; this is the first time that this has been performed for protein data.

1. Introduction

During the course of data collection, a crystal may produce an unexpected aperiodic diffraction pattern. Currently, when these cases occur the first response is to remove the crystal and to mount another one because there is no software to handle these data. Although the spots may be well defined, they cannot be indexed with standard software, or perhaps only a fraction of them can be indexed. There can be many reasons for this: for example, modulations of position or occupancy, quasicrystals *etc.* Of interest here is the case of a protein crystal with a positional modulation in one direction.

The diffraction pattern of a modulated crystal has a unique signature consisting of main reflections flanked by one or more satellites (Fig. 1). When most protein crystallographers try to process these data, the main reflections will be indexed and predicted but the satellite reflections will be problematic. They may or may not be indexed as a supercell. If the modulation is incommensurate they will not be handled properly at all because they will not have an integral relationship to the main unit cell. Satellite reflections of significant intensity from protein crystals may not be observed if only low-resolution data are collected. If they are not indexed by standard periodic software they might also be ignored/unobserved by the crystallographer, especially if the crystallographer did not watch the images as they were integrated. In this case, only the main reflections are indexed, integrated and used to determine the basic unit cell and three-dimensional space group. Using only the main reflections corresponds to averaging the unit-cell contents over all unit cells; thus, the result is called an ‘average structure’. This

doi:10.1107/S0907444913004630

determined average structure will have good electron density and structure refinement in the regions where the modulation is negligible and poor or absent electron density in the highly modulated regions/domains of the protein structure. The refined B values in these well ordered and modulated areas will be correspondingly low and high, respectively. Since the intensities and spacing of the satellite reflections are not used, structural information about the modulated domains is lost and these components cannot be solved.

Satellites can be successfully indexed using **q** vectors. One software package that can process incommensurate protein diffraction data is *EVAL*15 (Porta *et al.*, 2011; Schreurs *et al.*, 2010). If the **q** vector is a rational number, such as 1/7, the modulation is commensurate with the main structure (Fig. 2). Irrational **q** vectors, such as 1/7.314 . . . , indicate an incommensurate modulation. One or more **q** vectors may be required depending on the number of directions the satellites have relative to the mains. Modulations with one direction require only a single **q** vector. The satellites are indexed along the **q** vector moving out from the main. The **q** vector can lie along a primary axis (*e.g.* along $\mathbf{a}_2^*$ in Fig. 1) or it can be arbitrarily oriented in the crystal. Each **q** vector requires an additional index in the reflection table. For example, in the case of one **q** vector all reflections will have four indexes.

This type of indexing is known as (3+n)D indexing, where n is the number of **q** vectors in the data set [Fig. 1 shows a (3+1)D indexing]. The presence of satellite reflections indicates that there is a disruption in the short-range order of the crystal along the direction(s) of the **q** vector(s) and that this disruption has ordered structure rather than random unrelated positions (Fig. 2). Commensurate structures have a modulation that occurs in an integral number of unit cells (Fig. 2*b*), whereas incommensurate structures do not (Fig. 2*c*). Modulated protein crystals can be grown naturally (*e.g.* fibrous proteins and large protein complexes), can occur owing to the physical stress associated with cryocooling (*e.g.* proteins with large solvent contents) or can be purposefully induced in order to determine time-resolved structures of biological intermediates (the profilin–actin case; see Lovelace *et al.*, 2008).

There are three ways to handle the diffraction pattern of a modulated structure. The first way is to ignore the satellite reflections and use only the main reflections to solve the average structure. The second way is to drop the distinction between main and satellite reflections and to use all reflections equivalently for indexing with a supercell. Modulated structures can be approximated with conversion from **q**-vector indexing to supercell indexing (Wagner & Schönleber, 2009). The third way invokes superspace, uses all reflections and retains the distinction between main and satellite reflections. An excellent review article on the application of superspace to aperiodic crystals has been published (van Smaalen, 2005). This method allows the most accurate refinement of incommensurately modulated structures, but has not been used in protein crystallography. It uses the main reflections to determine the reciprocal unit cell, uses a **q** vector to describe the modulation direction and spacing, and employs superspace groups and modulation functions in the refinement of the crystal structure. Modulated structures can be more accurately solved and refined by making use of superspace. The computer simulation described here employs the supercell method to create the modulated structure and then converts the supercell into superspace. The (3+1)D diffraction and structural data were imported into *Jana*2006. The data generated by this simulation will be used to test future versions of software for the full refinement with stereochemical restraints of incommensurately modulated protein structures in (3+1)D superspace.

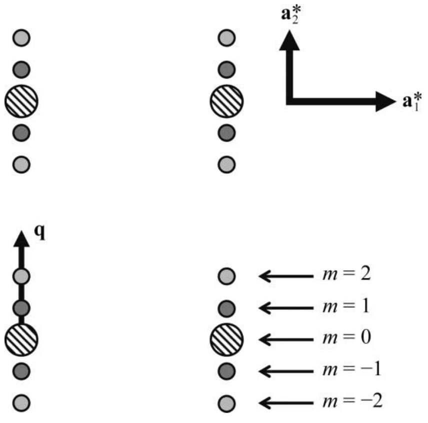

Figure 1
Introduction to (3+1)D modulated diffraction. The main reflections are shown as large hatched circles; the vectors $\mathbf{a}_1^*$ and $\mathbf{a}_2^*$ define the main (or basic) unit cell as derived from the positions of the main reflections. The satellite reflections are shown as gray circles. In this example, the **q** vector is along $\mathbf{a}_2^*$. How the fourth index, m, for the second-order satellites is assigned is shown.

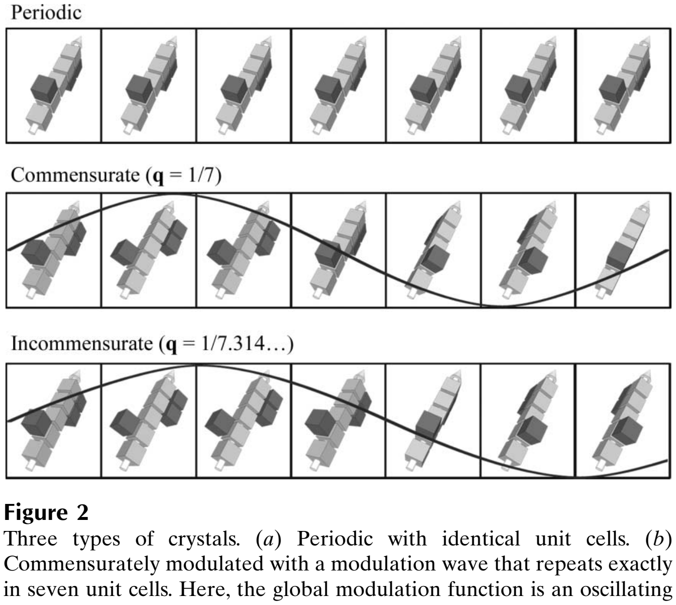

Figure 2
Three types of crystals. (*a*) Periodic with identical unit cells. (*b*) Commensurately modulated with a modulation wave that repeats exactly in seven unit cells. Here, the global modulation function is an oscillating rotation about a vertical axis. In this type of modulation the satellite reflections would have an integral relationship with the mains and could be assigned three-dimensional indices (h, k, l) in register with the main reflections. (*c*) Incommensurately modulated crystal, where the modulation wave does not repeat in phase with the unit cells and (3+1)D indices (h, k, l, m) are required.

1.1. Supercell and superstructure

Diffraction from a commensurately modulated crystal can usually be described using a supercell. Here, there is no distinction between main and satellite reflections and all reflections are used equivalently for indexing. There can be close spacing between the reflections and, depending on the choice of supercell, many unobserved reflections. Protein crystallographers have employed this method by solving commensurately modulated structures with very large unit cells and many molecules in the asymmetric unit, and sometimes observe noncrystallographic translational symmetry in such structures. The ratio of observed reflections to refined parameters typically limits the crystallographic refinement for these cases (Zwart *et al.*, 2008; Vila-Sanjurjo *et al.*, 2004).

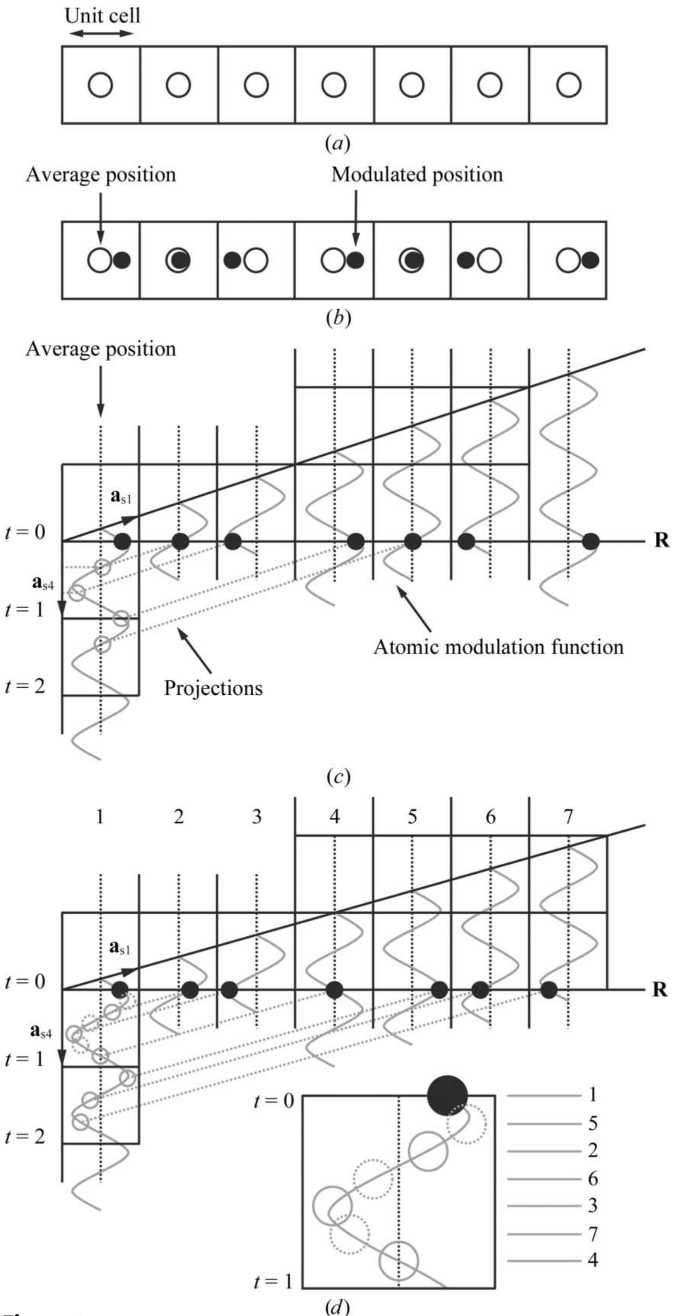

Figure 3
Description of superspace. (*a*) A periodic crystal structure that is composed of one atom (circle) in the asymmetric unit for seven unit cells. (*b*) A modulated system with the actual atomic positions (black dots) displaced from the average positions (circles). A commensurate modulation repeating every three unit cells is drawn. (*c*) Superspace description given as a (1+1)-dimensional drawing of the atomic positions shown in (*b*). A section defined by the superspace vectors $\mathbf{a}_{s1}$ and $\mathbf{a}_{s4}$ is shown, with $\mathbf{a}_{s4}$ perpendicular to $\mathbf{a}_1$ (not shown) and the angle between $\mathbf{a}_1$ and $\mathbf{a}_{s1}$ defined by the α component of $\mathbf{q}$, where $\mathbf{q} = \alpha\mathbf{a}_1^* + \beta\mathbf{a}_2^* + \gamma\mathbf{a}_3^*$ (see p. 253 and Figs. 7*c* and 8 of Wagner & Schönleber, 2009). The average position (dashed vertical lines), actual atomic positions (black circles), atomic modulation functions (gray wavy lines) and projections (gray dashed lines) to *t*-sections are shown. **R** represents physical space and contains all three directions; it is parallel to $\mathbf{a}_1$. The modulation is commensurate with $\mathbf{q} = (1/3)\mathbf{a}_1^* + 0\mathbf{a}_2^* + 0\mathbf{a}_3^*$. (*d*) Superspace diagram showing a 2/7 commensurate modulation with the unit-cell reordering that occurs in the *t*-plot highlighted in the enlarged area from $t = 0$ to $t = 1$. The unit cells 1–7 are encountered in the order 1, 5, 2, 6, 3, 7, 4 as *t* goes from 0 to 1.

1.2. Introduction to superspace

If the crystallographer chooses to retain the distinction between main and satellite reflections, then superspace is used to describe the modulated crystal and the problems associated with average and superstructure approximations are alleviated. Superspace, although conceptually more difficult, provides a much more accurate method for solving modulated structures. The general concepts of superspace are introduced here. However, several conventions first need to be explained.

Classical three-dimensional crystallography represents the unit-cell vectors as **a**, **b** and **c** with their respective coordinates x, y and z. These are now denoted $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$ with coordinates x_1, x_2 and x_3. This notation is used because it is easy to extend it to any number of dimensions. For (3+1)D superspace the axes are $\mathbf{a}_{s1}$, $\mathbf{a}_{s2}$, $\mathbf{a}_{s3}$ and $\mathbf{a}_{s4}$ and the dimensions are x_1, x_2, x_3 and x_4. The fourth dimension is defined to be perpendicular to all three dimensions of physical space. It is convenient to draw two-dimensional (or three-dimensional) sections in superspace (for example $\mathbf{a}_{s1}$ *versus* $\mathbf{a}_{s4}$), as it is impossible to draw all four perpendicular dimensions on paper or on a computer system. Furthermore, it is convenient to define a space line, **R**, which represents all of physical space.

Using these conventions, a (3+1)D modulated structure can be diagrammed. In a periodic crystal the same atom occupies the same location in every unit cell (Fig. 3*a*). An atom that has a modulation of a displacement nature is shown over several neighboring unit cells (black dots in Fig. 3*b*) relative to the average position of the atom (circle). Here, the atomic positions are only shown along a single axis for simplicity. This modulation is commensurately repeating every three unit cells. Also, the displacements relative to the size of the unit cell are exaggerated from what would be observed in order to make the modulation easier to visualize. In each unit cell, the position of the atom is different. A (3+1)D superspace description of the system is shown in Fig. 3(*c*). The superspace description provides a much more complete picture of the displacements than the supercell description (Fig. 3*b* *versus* Fig. 3*c*). The superspace approach is successful for both commensurate and incommensurate modulations. For each **q** vector, an extra dimension ($\mathbf{a}_{s4}$ in Fig. 3*c*) is added to the system that is orthogonal to physical space (**R**). Units along **R**

Table 1
A comparison between a three-dimensional space group $P2_12_12_1$ with (3+1)D superspace groups $P2_12_12_1(00\gamma)$ and $P2_12_12_1(0\beta0)$.

Space group	$P2_12_12_1$	$P2_12_12_1(00\gamma)$	$P2_12_12_1(0\beta0)$
No.	19	19.1	19.1 reset
Unit-cell constraints	$\alpha = \beta = \gamma = 90^\circ$	$\alpha = \beta = \gamma = 90^\circ$	$\alpha = \beta = \gamma = 90^\circ$
q-vector constraints	Not applicable	Along x_3 axis	Along x_2 axis
Symmetry operators	x_1, x_2, x_3 $\frac{1}{2} - x_1, -x_2, \frac{1}{2} + x_3$ $-x_1, \frac{1}{2} + x_2, \frac{1}{2} - x_3$ $\frac{1}{2} + x_1, \frac{1}{2} - x_2, -x_3$	x_1, x_2, x_3, x_4 $\frac{1}{2} - x_1, -x_2, \frac{1}{2} + x_3, x_4$ $-x_1, \frac{1}{2} + x_2, \frac{1}{2} - x_3, -x_4$ $\frac{1}{2} + x_1, \frac{1}{2} - x_2, -x_3, -x_4$	x_1, x_2, x_3, x_4 $\frac{1}{2} - x_1, -x_2, \frac{1}{2} + x_3, -x_4$ $-x_1, \frac{1}{2} + x_2, \frac{1}{2} - x_3, x_4$ $\frac{1}{2} + x_1, \frac{1}{2} - x_2, -x_3, -x_4$
Special reflection conditions	$h00$: $h = 2n$ $0k0$: $k = 2n$ $00l$: $l = 2n$	$h000$: $h = 2n$ $0k00$: $k = 2n$ $00lm$: $l = 2n$	$h000$: $h = 2n$ $0k0m$: $k = 2n$ $00l0$: $l = 2n$

are in terms of unit cells. In this example, $\mathbf{a}_1$ (not shown in Fig. 3) is along **R**. Units (t) along $\mathbf{a}_{s4}$ are in terms of the unit waves of the modulation. The **q** vector determines the angle that $\mathbf{a}_{s1}$ makes with $\mathbf{a}_1$ [see p. 253 and Figs. 7(c) and 8 of Wagner & Schönleber, 2009]. Fractional coordinates x_4 along $\mathbf{a}_{s4}$ are measured in terms of unit waves and t is the phase of the modulation wave and determines the displacements of atoms along $\mathbf{a}_1$ for the respective unit cells. A periodic atomic modulation function (AMF; gray lines in Fig. 3c) can now be created in superspace that will correctly determine the position of the atom in any unit cell in normal space relative to the average position (black dashed lines). For atoms in superspace that are not modulated the AMFs are straight lines.

In the example shown (Fig. 3c), the modulation waves change through one phase cycle every three unit cells or a **q** vector of $(1/3)\mathbf{a}_1^* + 0\mathbf{a}_2^* + 0\mathbf{a}_3^*$ based on the direction and spacing of the satellites from the mains (Fig. 1). The periodic superspace structure is now defined by two basis vectors: $\mathbf{a}_{s1}$ and $\mathbf{a}_{s4}$. $\mathbf{a}_{s1}$ makes an angle with $\mathbf{a}_1$, the tangent of which is $\alpha/|\mathbf{a}_1|$, where α is the **q**-vector component along $\mathbf{a}_1^*$; in this case, it would be arctangent$[(1/3)/|\mathbf{a}_1|]$. The modulation waves are translated along $\mathbf{a}_{s1}$ (*i.e.* are periodic in $\mathbf{a}_{s1}$) and where they intersect **R** the modulated atomic positions are found. Equivalent superspace positions can be calculated along $\mathbf{a}_{s4}$ for atoms in real space by projecting lines that are parallel to $\mathbf{a}_{s1}$ but pass through the atoms (gray dashed lines in Fig. 3c) to the point where they intersect the modulation function closest to the origin (gray circles). These positions can be further compressed by translating them to their equivalent locations in the modulation wave closest to the origin. Effectively, this means that all atomic displacements for every possible unit cell along the modulation direction can be visualized by looking at one unit wave of the AMF (*i.e.* one t section).

One way to analyse data in superspace is with a t-plot. A t-plot is constructed by looking at the atomic positions as a function of t. A curious feature of superspace is that the order in which states appear in a t-plot is not necessarily the order that they will appear in the crystal. To illustrate this point, Fig. 3(d) shows a diagram of a supercell that has two modulation waves every seven unit cells [$\mathbf{q} = (2/7)\mathbf{a}_1^* + 0\mathbf{a}_2^* + 0\mathbf{a}_3^*$]. In this case the unit cells are labeled 1–7. When their positions are projected into superspace (gray dashed lines to gray circles) they occupy two units in t. The positions in the range $t = 1$ to $t = 2$ are translated to their equivalent locations (dashed gray circles) in the range $t = 0$ to $t = 1$. Here, the unit cells are encountered as a function of increasing t in the order 1, 5, 2, 6, 3, 7, 4 instead of 1, 2, 3, 4, 5, 6, 7 as they would be ordered in the supercell. This order is easier to see in the enlarged $t = 0$ to $t = 1$ region shown in Fig. 3(d).

For modulated systems, superspace provides a more complete description of the atomic displacements than the supercell method. In superspace atoms are described by continuous atomic modulation functions describing a continuum of states instead of just a few discrete points as in the supercell approach.

1.3. Superspace groups and superspace symmetry

Superspace has space groups and symmetry operators that are functionally similar to the normal three-dimensional space groups (van Smaalen, 2007; Dauter & Jaskolski, 2010). Table 1 shows a comparison of three-dimensional space group $P2_12_12_1$ (No. 19) to a related (3+1)D superspace group $P2_12_12_1(00\gamma)$ (No. 19.1) as obtained from Table 9.8.3.5 of Janssen *et al.* (1999). The extra information attached to the superspace group describes the characteristics of the **q** vector. The (00γ) indicates that for this space group **q** vectors can only be along one of the principal axes with $\mathbf{a}_3$ as the standard setting. Additionally, the symmetry operators are listed for the convention in which the modulation is along the $\mathbf{a}_3$ axis; if the modulation is along one of the other axes then the sign would need to be flipped to match that direction. The superspace group $P2_12_12_1(0\beta0)$ is given as an example in Table 1 for a **q** vector along the $\mathbf{a}_2$ axis. One interesting property of this superspace group is the systematic absences. Main reflections are systematically absent as in the three-dimensional space group, but along the modulation direction if the main reflection is systematically absent then the associated satellites are also systematically absent for this specific superspace group.

1.4. Superspace *versus* supercell

For commensurate modulations, both the superspace and supercell approaches provide valid ways to describe how atoms are positioned within the unit cell. Superspace will describe a continuum of structural states along the modulation function in each t section. The states found from the superspace calculation should match the discrete states found using a supercell approximation based on the same data. Thus, unlike incommensurate modulations, a commensurate modulation provides a unique situation in which a structure can be solved using two independent approaches. The simulation calculated here takes advantage of this situation.

For cases with a large supercell, the superspace approach can provide better data-to-parameter ratios. For a case in which the modulation can be described reasonably well using second-order Fourier coefficients, each atom is described by a minimum of 15 parameters in superspace independent of how large the supercell might be. Here, each atom has three coordinates to refine (x_1, x_2 and x_3 for the average position) and four parameters for the AMF for each coordinate (two amplitudes and two phases). In the case of the supercell the numbers of parameters grow as the supercell increases in size. If the required supercell is five times the basic cell or larger, then the superspace approach is better with respect to the data-to-parameter ratio.

1.5. Modulated simulation for methods development

In order to develop methods to solve a modulated protein (Lovelace *et al.*, 2008) training data are needed to test them. The training data can either be a solved structure with associated experimental data or a simulated structure. Since at this time there are no solved modulated protein structures, the best option is a simulated training set. A simulated modulation has previously been reported (Lovelace *et al.*, 2010). This simulation was successful at reproducing a diffraction pattern that was modulated, as well as providing some interesting information about the resolution, relative intensities and diffraction order of satellite reflections *versus* the degree of structural modulation. Unfortunately, it was not a good candidate for use in methods development for several reasons. First, it made use of a global modulation function that limited the application to use in space group $P1$ only. Additionally, the modulation ignored spatial and stereochemical restraints and all of the atoms were converted to nitrogen, resulting in loss of the amino-acid residue and chain information in the model. In reality, a global modulation function will probably not be sufficient to describe a modulated protein crystal structure. It is likely that each atom (or rigid group of atoms) in the crystallographic asymmetric unit will have unique displacements and will require the refinement of its own modulation function.

For methods development, the comprehensive training set needs to consist of a protein model with realistic stereochemical constraints and no steric conflicts. It also needs to have individual atomic modulation functions that can be expressed as Fourier series for the x_1, x_2 and x_3 directions. For the types of modulated protein crystals observed to date, it will require structure factors (amplitude and phase) in (3+1)D superspace. Finally, it is desirable for the simulation to be calculated in a higher symmetry space (and superspace) group, and the simulation will be most useful if the space group is one that has already been observed in a real modulated protein crystal (Porta *et al.*, 2011).

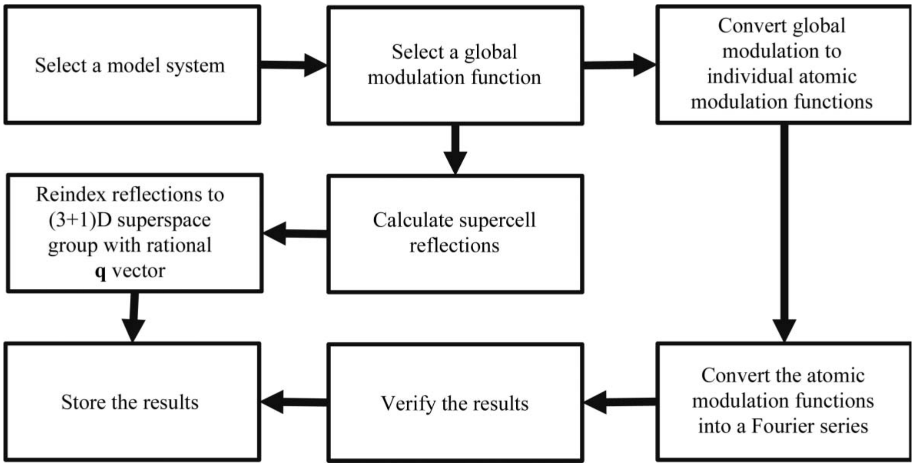
Select a model system
Select a global modulation function
Convert global modulation to individual atomic modulation functions
Reindex reflections to (3+1)D superspace group with rational **q** vector
Calculate supercell reflections
Store the results
Verify the results
Convert the atomic modulation functions into a Fourier series

Figure 4
Data-processing flowchart.

2. Methods

The approach used to create the simulated data set is diagrammed in Fig. 4. Firstly, a model system and a global modulation function were selected. The global modulation function is used to create a commensurate modulation of the protein within a supercell. Structure factors are calculated with the supercell modulated structure. Those structure factors are reindexed in the (3+1)D superspace group and both sets of indexing are saved. The global modulation is converted into individual modulation functions for each atom that is modulated. Through Fourier synthesis, the individual modulation functions are converted to a second-order Fourier series. Atomic positions are calculated for atoms occupying the same volume as the supercell using the superspace approach. To verify that the conversions are correct, the atomic positions from superspace are compared with the supercell positions. If all of the positions line up then the superspace representation is consistent with the supercell representation. Once both of the approaches agree, all of the information is stored in appropriate files for use in methods development.

2.1. Selection of a model system

Orthorhombic profilin–actin crystals can be induced to become incommensurately modulated, with satellite reflections appearing near the periodic main reflections. The modulation vector points in the x_2 direction and the (3+1)D superspace group is 19.1 with setting $P2_12_12_1(0\beta0)$. The modulation is along the $\mathbf{a}_2^*$ direction and has an incommensurate $\mathbf{q}$ vector of 0.2829 . . . (Porta *et al.*, 2011). The closest practical commensurate approximation for this $\mathbf{q}$ vector is 2/7, with about two modulation waves for every seven unit cells. ToxD (PDB entry 1dtx; Skarzyński, 1992) was used because it has the same $P2_12_12_1$ space group but contains fewer atoms than profilin–actin (unit-cell parameters a_1 = 73.52, a_2 = 39.06, a_3 = 23.15 Å). Another advantage is that it has large solvent channels (Fig. 5*b*). To create the modulated superstructure a polypeptide fragment of the ToxD protein was clipped, trimmed and translated to the solvent channel and was then modulated with the global modulation function (Fig. 5). The modulated polypeptide was placed in the solvent channel so that the displacements would not create steric conflicts with any other protein atoms. Several of the amino acids were also mutated to alanine to free up even more space around the modulating atoms.

2.2. Selection of a global modulation function

The next task was to pick a global modulation function. The function took the form of a rotation about an axis. The axis was defined by the C$^{\alpha}$ atoms of the first and last amino acids in the group of modulated atoms. The amount of rotation was determined by the position of the center of mass (COM) of the modulating atoms. The COM was used as it was an easy property to calculate for the atoms independent of orientation. The COM along the x_2 dimension was converted into x_4 using (1). x_4 is converted into an angular amount using (2). $\theta_{\max}$ was defined as 15° because at this amount of rotation the relative intensities between main and satellite reflections visually appeared similar to those observed for profilin–actin.

$$x_4 = \frac{\mathrm{COM}_{x_2}}{a_2} q. \tag{1}$$

$$\theta(x_4) = \theta_{\max} \sin(2\pi x_4). \tag{2}$$

The simplified matrix form of the rotation about an axis equation is fairly straightforward for this type of transformation (3). The rotation is accomplished by translating one of the end points to the origin. The points are then rotated about x_3 until the other end point is in the x_1x_3 plane. The points are rotated about y until the other end point is along the x_3 axis. The points are rotated about x_3 for the desired amount of rotation. Finally, the inverse transforms are applied to return the end points back to their original position. This transform is widely used in three-dimensional graphics. The individual subtransformations can be pre-multiplied, resulting in a simplified matrix form that is shown in (4).

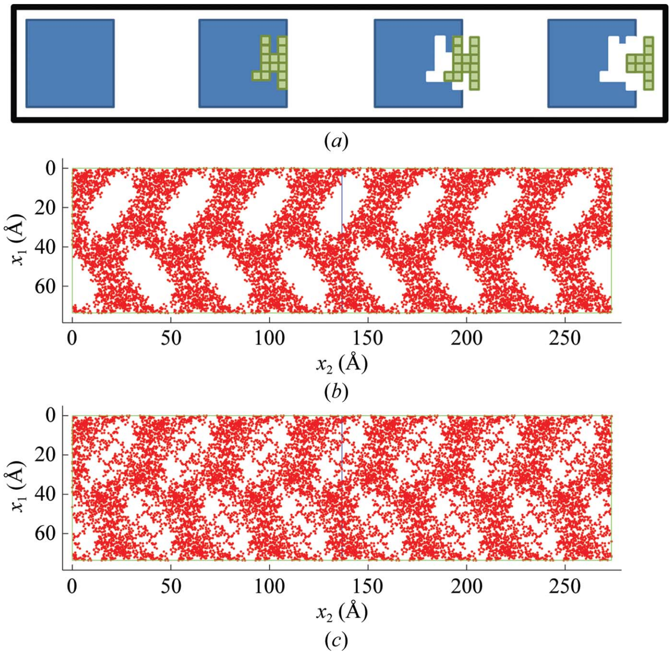

Figure 5
The approach used to create an independent region that can be modulated. (*a*) A schematic diagram showing how a region was selected and displaced out into the solvent area. The boundaries were cleaned up by mutating some of the residues to alanine. (*b*) The initial atomic positions. (*c*) The position after the process is completed.

$$\mathbf{A}_n = \mathbf{Trans}^{-1} \cdot \mathbf{Rot}_{x_3}^{-1} \cdot \mathbf{Rot}_{x_2}^{-1} \cdot \mathbf{Rot}[\theta(x_4)] \cdot \mathbf{Rot}_{x_2} \cdot \mathbf{Rot}_{x_3} \cdot \mathbf{Trans} \cdot \mathbf{A}_i. \tag{3}$$

$$\begin{pmatrix} A1n_{x_1} & \cdots & Ain_{x_1} \\ A1n_{x_2} & \cdots & Ain_{x_2} \\ A1n_{x_3} & \cdots & Ain_{x_3} \\ 1 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} \mathrm{TM}_{11} & \cdots & \cdots & \mathrm{TM}_{14} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathrm{TM}_{41} & \cdots & \cdots & \mathrm{TM}_{44} \end{pmatrix} \begin{pmatrix} A1_{x_1} & \cdots & Ai_{x_1} \\ A1_{x_2} & \cdots & Ai_{x_2} \\ A1_{x_3} & \cdots & Ai_{x_3} \\ 1 & \cdots & 1 \end{pmatrix}. \tag{4}$$

Here, the rotation is represented as a (4 × 4) transformation matrix (**TM**). The dimension of the transform is one dimension higher than the space to account for translations. This is a slightly different formulation than that used in crystallography, in which the rotation and translation are stored in a 3 × 4 matrix. The advantage of the 4 × 4 matrix is that all operations are performed using matrix multiplication and reverse operations can be performed by simply inverting the 4 × 4 matrix. In the 3 × 4 matrix formulation translation requires special handling by the program for the extra column to be interpreted as a translation. [As a simple example, (3) is not valid for 3 × 4 matrices.] The atomic coordinates (x_1, x_2, x_3) for each atom are housed columnwise in the **A** matrix, where i is the total number of modulating atoms in the structure. To calculate the new positions (n), the atom matrix (**A**) is multiplied by the transformation matrix (**TM**) to generate the new rotated atomic positions.

2.3. Structure factors and superspace indexing

From the global modulation function it is possible to calculate a supercell that has two modulation waves every seven unit cells. This modulation was selected as a reasonable commensurate approximation to that observed in the profilin–actin crystals. ToxD was extended by sevenfold in the $\mathbf{a}_2$ direction to form a supercell and, using the global modulation function, the displacements were applied to the modulated atoms. *SFALL* was used with the supercell to calculate structure factors (Winn *et al.*, 2011). A comparison between the modulated and periodic ToxD structure factors is shown in Fig. 6. First-order satellites are visible in the modulated data for most of the mains. At higher resolutions second-order satellites are also visible (not shown). The structure factors are converted to (3+1)D indexing, including mains and all first-order and second-order satellites (Fig. 7). In this procedure, the *h* and *l* indices are easily converted since they are the same in both representations. The *k* index is more complicated. All of the *k*s are integers divided by 7, keeping both the integral part of the division and the remainder. The integral part is the value of *k* in (3+1)D space. Remainders are used to determine the *m* values, with a value of 0 for the main reflections ($m = 0$). Remainders of 3 and 5 have *m* values of −1 and −2, respectively, and the *k* value is the integral value incremented by 1. Remainders of 2 and 4 have *m* values of 1 and 2, respectively, and the *k* value is the integral value. Remainders of 1 and 6 are discarded because the third-order satellites in this data set are extremely weak, with I/σ very close to 1. Following this procedure, the three-dimensional structure factors from the modulated supercell can be converted to (3+1)D with the basic unit-cell dimensions. Note that in the three-dimensional indexing (Fig. 7, left) a sevenfold unit cell was used, but in superspace indexing (Fig. 7, right) a **q** vector of $(2/7)\mathbf{a}_2^*$ was used. The *m* indexing of the satellites is sequential with respect to their associated main reflection. Thus, in this case the satellites of the neighboring main reflections interweave between one another (Fig. 7).

2.4. Individual atomic modulation functions

The standard way to describe the modulation for an atom is to break the modulation into three orthogonal Fourier expansions. There is one atomic modulation function (u) for each primary axis (u_{x1}, u_{x2} and u_{x3} corresponding to x_1, x_2 and x_3, respectively) that describes the deviation from the average position as a function of x_4. The global modulation function can be converted to individual modulation functions by rewriting the transformation matrix in (4) as shown in (5). An extra translation is added to make the displacements relative to the average position by translating the atom (A_i) to the origin. This translation individualizes the atomic modulation function for each atom since before this final translation all displacements are still relative to the COM.

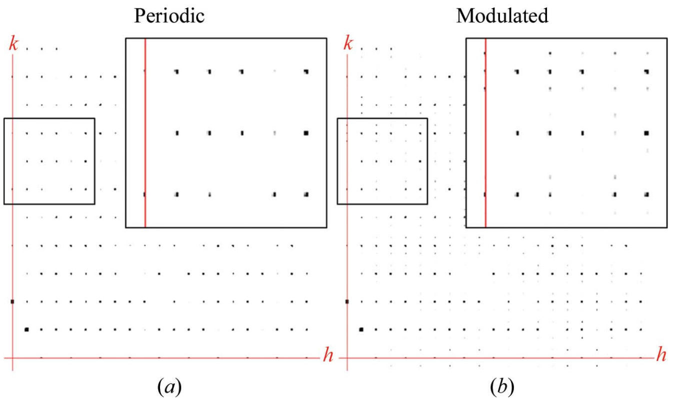

Figure 6
Calculated structure factors in the training set: (*a*) periodic and (*b*) modulated with a **q** vector of $(2/7)\mathbf{a}_2^*$.

$$\begin{pmatrix} u_{x_1} \\ u_{x_2} \\ u_{x_3} \\ 1 \end{pmatrix} = \mathbf{Trans}(-Ai)\cdot\mathbf{PostTM}\cdot\mathbf{Rot}[\theta(x_4)]\cdot\mathbf{PreTM}\begin{pmatrix} Ai_{x_1} \\ Ai_{x_2} \\ Ai_{x_3} \\ 1 \end{pmatrix}. \tag{5}$$

In (5), the **PostTM** matrix is created by multiplying the $\mathbf{Trans}^{-1}$, $\mathbf{Rot}_{x_3}^{-1}$ and $\mathbf{Rot}_{x_2}^{-1}$ matrices because the values in these matrices are dependent only on the location of the rotation axis and therefore are effectively constant for all of the atoms. For the same reason the **PreTM** matrix is created by multiplying the $\mathbf{Rot}_{x_2}$, $\mathbf{Rot}_{x_3}$ and **Trans** matrices. The **Rot** matrix cannot be calculated ahead of time because it is a function of θ, which is a function of x_4. Solving (5) for u_{x_1}, u_{x_2} and u_{x_3} yields

$$\begin{pmatrix} u_{x_1} \\ u_{x_2} \\ u_{x_3} \end{pmatrix} = \begin{Bmatrix} A + B\sin[\sin(Dx_4 + E)] + C\cos[\sin(Dx_4 + E)] \\ F + G\sin[\sin(Dx_4 + E)] + H\cos[\sin(Dx_4 + E)] \\ I + J\sin[\sin(Dx_4 + E)] + K\cos[\sin(Dx_4 + E)] \end{Bmatrix}. \tag{6}$$

For each dimension (x_1, x_2, x_3) five parameters need to be calculated to convert the rotational global modulation function into three orthogonal individual atomic modulation functions for each modulating atom.

2.5. Fourier synthesis of the AMFs

*Jana*2006 uses atomic modulation functions expressed as Fourier series (here denoted by $\mathcal{F}$). Using the known integrals (Gradshteyn & Ryzhik, 2007),

$$\frac{1}{\pi}\int_0^\pi \cos(2n\theta)\cos(z\sin\theta)\,\mathrm{d}\theta = J_{2n}(z), \tag{7}$$

$$\frac{1}{\pi}\int_0^\pi \sin[(2n+1)n\theta]\sin(z\sin\theta)\,\mathrm{d}\theta = J_{2n+1}(z), \tag{8}$$

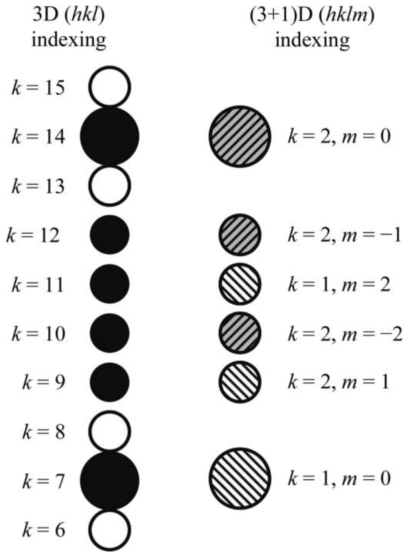

Figure 7
Diagram demonstrating how the reflections from superspace are reindexed into (3+1)D space.

it can be shown that the family of functions in (6) can be converted to Fourier series as follows.

$$u_{x_1} = f_1(x_4) + f_2(x_4) + f_3(x_4), \tag{9}$$

$$\mathcal{F}(u_{x_1}) = \mathcal{F}[f_1(x_4)] + \mathcal{F}[f_2(x_4)] + \mathcal{F}[f_3(x_4)], \tag{10}$$

$$\mathcal{F}[f_1(x_4)] = \mathcal{F}(A) = A, \tag{11}$$

$$\begin{aligned} \mathcal{F}[f_2(x_4)] &= \mathcal{F}\{B\sin[\sin(Dx_4 + E)]\} \\ &= 2B\sum_{n=1}^{\infty} J_{2n-1}(1)\{\sin[(2n-1)E]\cos[(2n-1)Dx_4] \\ &\quad + \cos[(2n-1)E]\sin[(2n-1)Dx_4]\}, \end{aligned} \tag{12}$$

$$\begin{aligned} \mathcal{F}[f_3(x_4)] &= \mathcal{F}\{(C\cos[\sin(Dx_4 + E)]\} \\ &= CJ_0(1) + 2C\sum_{n=1}^{\infty} J_{2n}(1)[\cos(2nE)\cos(2nDx_4) \\ &\quad + \sin(2nE)\sin(2nDx_4)]. \end{aligned} \tag{13}$$

Here, $J_n(x)$ are Bessel functions of the first kind. J is the letter used to denote a Bessel function of the first kind, n is the index and x is the argument, which in this case is 1. In a more generalized case, where the function is not necessarily known, a more straightforward approach is to first fit a curve through the data. The next step is to oversample points on the curve. The last step is to approximate the Fourier coefficients by taking the fast Fourier transform (FFT) of the discrete data and using these terms to approximate the Fourier series. The approximate fits for one of the modulated atoms (Fig. 8) shows the quality of the fit for the first-order (dotted line) Fourier coefficients and the first-order and second-order coefficients (solid line). Here, a coarse subsampling of the real functions is shown as solid circles. Fourier coefficients were calculated by sampling 4096 equally spaced points over a single unit wave of the modulation function. This rate was expected to be many times higher than the Nyquist–Shannon sampling rate (Shannon, 1949) for this modulation. The colors identify the different modulation functions along each dimension, where x_1, x_2 and x_3 are plotted as a function of x_4. The second-order coefficients seem to be a reasonable fit to the data (Fig. 8). Data files for the (3+1)D superspace representation are currently stored as comma-separated values in three different files (reflections, atoms and superspace group/unit-cell information) because there is currently no accepted format for a modulated protein data set. The AMFs in *Jana*2006 do not contain the constant term. This is because the atoms are supposed to be at their average position and would have a constant term of zero. The constant term from the Fourier series was an indication that the initial atomic position was not quite at the average position. The constant terms were added to the atomic positions to move the atoms to the correct average positions. These small shifts in the average structure resulted in small correction to the Fourier phase coefficients, as shown in Fig. 9.

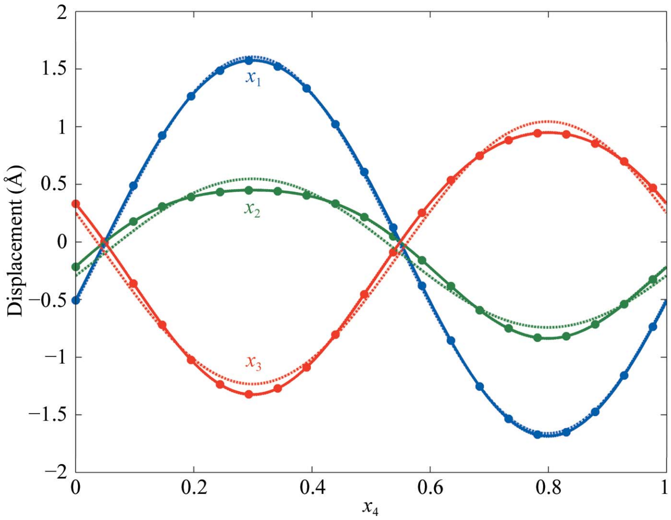

Figure 8
Example Fourier series fit of the three atomic modulation functions with one for each axis (x_1, x_2 and x_3). The actual position is shown subsampled as solid circles. The first-order fit is shown as a dotted line. The second-order fit is shown as a solid line.

3. Results and discussion

In order to verify that the two different representations (superspace and supercell) were correct, atomic positions were calculated for all of the atomic locations inside the supercell including symmetry-related positions. These positions were compared with those calculated using the superspace representation which would fill the same volume. The superspace representation consisted of the average positions of the atoms, their modulation functions and the superspace symmetry operators. If everything has been converted correctly then both representations should predict the same atomic positions.

3.1. Supercell approach

The supercell approach was fairly straightforward. The modified ToxD coordinate file was expanded along the a_2 direction by sevenfold using *PDBSET*. The global modulation function was used to modulate the atoms. The symmetry-related positions were calculated with *PDBSET* and the atoms were translated so that they occupied the sevenfold supercell. Table 2 summarizes a comparison of the structure factors by

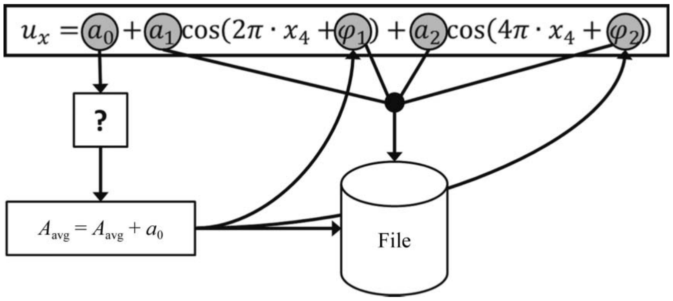
$u_x = a_0 + a_1\cos(2\pi \cdot x_4 + \varphi_1) + a_2\cos(4\pi \cdot x_4 + \varphi_2)$
?
$A_{\rm avg} = A_{\rm avg} + a_0$
File

Figure 9
Application and storage of the Fourier coefficients. The constant term is added to the average structure, resulting in a slight adjustment to the Fourier phase coefficients.

Table 2
Reflection statistics.

Structure factors from the average structure (F_{avg}) and from the modulated structure (F_{mod}). The modulated data were divided into mains, first-order satellites and second-order satellites. The number of reflections for the first and second order is two times larger because there are two of these reflections for each main reflection.

				F_{mod}†		
Resolution (Å)	n_{ref}	R_{merge}† (%)	F_{avg}‡	Mains	First order	Second order
36.7–2.20	3695	8.69	123.12 (120.94)	120.77 (119.78)	15.01 (10.94)	6.50 (3.96)
2.20–1.73	3694	16.94	46.83 (27.92)	45.20 (27.08)	6.63 (3.97)	4.40 (2.49)
1.73–1.51	3695	19.14	26.00 (15.18)	25.18 (14.69)	3.57 (2.03)	2.72 (1.50)
1.51–1.37	3694	20.68	17.25 (9.55)	16.43 (9.18)	2.42 (1.37)	1.89 (1.06)
1.37–1.27	3695	18.94	13.15 (7.28)	12.70 (6.99)	1.77 (0.97)	1.35 (0.75)
1.27–1.19	3695	19.61	10.11 (5.64)	9.74 (5.47)	1.33 (0.75)	1.03 (0.57)
1.19–1.13	3694	18.27	8.15 (4.45)	7.80 (4.33)	1.04 (0.57)	0.81 (0.44)
1.13–1.08	3695	18.17	6.46 (3.62)	6.26 (3.51)	0.80 (0.44)	0.64 (0.35)
1.08–1.04	3694	18.29	5.10 (2.74)	4.92 (2.66)	0.63 (0.35)	0.50 (0.27)
1.04–1.00	3695	18.42	4.01 (2.18)	3.88 (2.10)	0.49 (0.28)	0.40 (0.22)

† R_{merge} calculated using *SFTOOLS* as 200 × $\sum|F_{\mathrm{avg}} - F_{\mathrm{mod}}(\mathrm{mains})|/\sum F_{\mathrm{avg}} - F_{\mathrm{mod}}(\mathrm{mains})$. ‡ Average structure-factor intensities for the resolution shell with standard deviations in parentheses.

comparing the average factors with the mains and the first-order and second-order satellites of the modulated structure. A similar analysis was performed on modulated profilin–actin crystals (see Table 2 of Porta *et al.*, 2011). *SFTOOLS* from the *CCP*4 software suite was used to calculate an R_{merge} of 13.5% between F_{avg} and F_{mod}(mains), which was overall better than that observed for profilin–actin crystals (~30%). The merging *R* value between F_{avg} and F_{mod}(mains) increases with resolution as expected. This is because modulations affect the high-resolution data more than the low-resolution data (Lovelace *et al.*, 2010). The F_{avg} are more intense than F_{mod}(mains) at all resolution levels. The ratio of first-order satellites over mains has an average ratio of 0.12, which is lower but comparable to the ratio observed for the profilin–actin system (0.20). Overall, the simulated diffraction data recapitulate what was observed with real modulated protein diffraction data.

3.2. Superspace approach

The superspace approach started with the modified ToxD coordinates. The ToxD coordinates were converted to the average structure by adding the constant portion of the FFT-derived Fourier fits to the atoms. Symmetry-related positions were determined by the superspace symmetry-related operators. The average structure was translated to the next unit-cell position along *b* and the symmetry-related position procedure was repeated. After six translations the resulting atomic positions should be equivalent to those from the supercell approach. Fig. 10 shows the results using these two approaches. In this case the supercell positions are shown as blue circles and the superspace positions are shown as red dots. The top view is looking down the modulation axis and the bottom view is looking along the modulation axis. Fig. 10 only shows the atoms that are modulating in the structure. Both approaches predict the same locations. This indicates that the global modulation function has been correctly converted into individual atomic modulation functions and that the superspace symmetry operators have been applied correctly.

3.3. Superspace symmetry operators

Table 1 shows a comparison of the superspace (3+1)D symmetry operators with the standard symmetry operators. From the table it may not be clear how to use the superspace symmetry operator with the modulation function. In order to demonstrate how to apply the superspace symmetry operators, an operator from Table 1 will be described. Firstly, an atom R_i from the atoms in the average structure R_{avg} is selected. The notation $R_{i(xn)}$ is used to denote the x_n coordinate of atom *i*. We selected the second symmetry operator ($1/2 - x_1$, $-x_2$, $1/2 + x_3$, x_4). This operator needs to be corrected for the direction of the **q** vector. In *International Tables for Crystallography* this superspace group is written for the case in which the **q** vector is along *c* (x_3). In this case the modulation is along *b* (x_2), resulting in the symmetry operator becoming $1/2 - x_1$, $-x_2$, $1/2 + x_3$, $-x_4$ (De Wolff *et al.*, 1981). The sign of x_4 needs to match the sign along the direction of the **q** vector (x_2). A new position for the atom (R_n) is calculated from the symmetry operator, where $R_n = 1/2 - R_{i(x_1)}$, $-R_{i(x_2)}$, $1/2 + R_{i(x_3)}$. x_4 is calculated by taking the dot product of **q** and R_n. In the example, the value of x_4 is $(2/7){\cdot}R_{n(x_2)}$. The final position (R_{f}) is determined by adding the atomic modulations as determined by the atomic modulation functions. The sign of the modulation function (*u*) is equal to the sign of the operator along the dimension, resulting in a final position $R_{\mathrm{f}} = R_{n(x_1)} - u_{x_1}(-x_4)$, $R_{n(x_2)} - u_{x_2}(-x_4)$, $R_{n(x_3)} + u_{x_3}(-x_4)$.

3.4. Representation of the modulation

Superspace data are often represented as a *t*-plot movie. Two *t*-plot movies (provided as Supplementary Material[1]) show the atomic positions for different *t*-sections from zero to one. One *t*-plot shows the AMFs using only first-order FFT coefficients and the second shows AMFs using first-order and second-order FFT coefficients. Effectively, the *t*-plot shows one period of the modulation function and therefore all possible atomic positions. It is important to note that although the *t*-plot is representative of all possible positions, consecutive unit cells along the modulation will have a different displacements than those shown in the *t*-plot. The reason for this reordering was discussed in §1.2.

3.5. Import into *Jana*2006

To refine and solve a modulated protein structure will require adding capabilities to an existing package. There are two possibilities. The first is to modify a protein package to

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: RR5034). Services for accessing this material are described at the back of the journal.

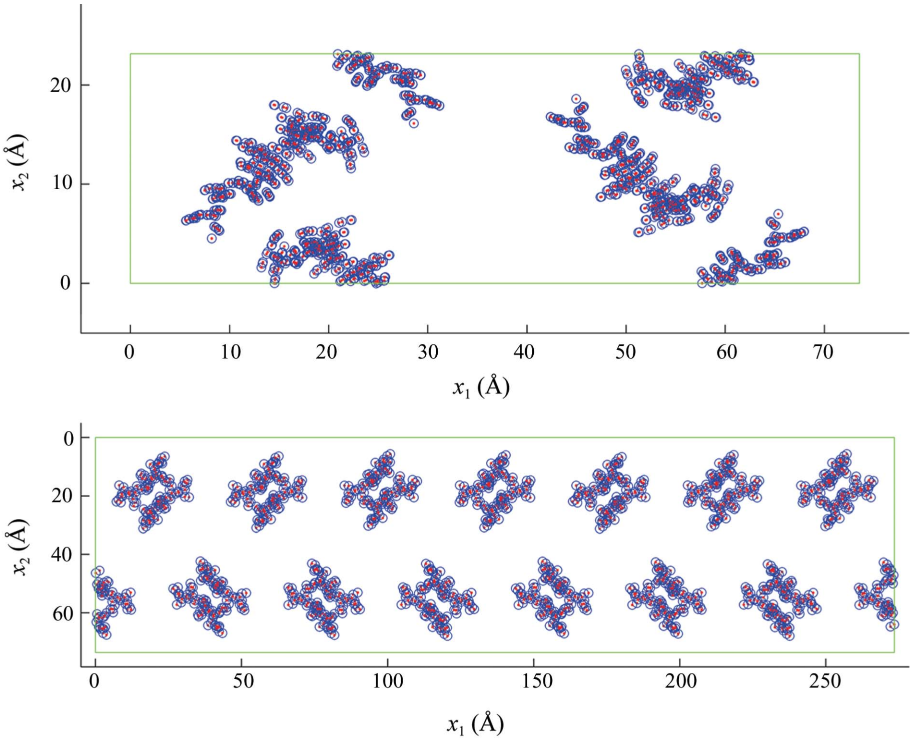

Figure 10
Comparison of atomic positions generated using the supercell approach (circles) and the superspace approach (dots). The top view is looking down the modulation axis. The bottom view is looking along the modulation axis. Only the modulating atoms are displayed.

understand superspace groups. The second is to modify a small-molecule package that can process modulated data. *Jana*2006 is the premier software package for the refinement and solution of modulated small-molecule structures. As a first step toward this integration we decided to see how difficult it would be to have *Jana*2006 input and display the modulated protein data.

Three files from the simulation were provided to *Jana*2006. The first was a text file containing the superspace group parameters and unit cell. The second contained atoms as comma-separated values. The values included atom type, $x_{1\rm avg}$, $x_{2\rm avg}$, $x_{3\rm avg}$, modulation flag, first-order and second-order Fourier coefficients and phases in x_1, x_2 and x_3. The third file contained the 3+1 indexing, intensities and phases.

After some trial and error, the information was successfully imported into *Jana*2006. Fig. 11 shows a two-dimensional slice of the four-dimensional electron density from superspace. The slice shows the atomic displacement along x_2 as a function of x_4 modulation (two modulation periods are displayed; x_4 goes from 0 to 2). The contours for density are from high (yellow) to low (blue). The atom which is represented by a function in superspace is shown as a red line.

In superspace functions are fitted to the density as opposed to individual atoms. The function follows the peak of the density. For proteins this change will most likely require some rethinking as to how stereochemical constraints are implemented.

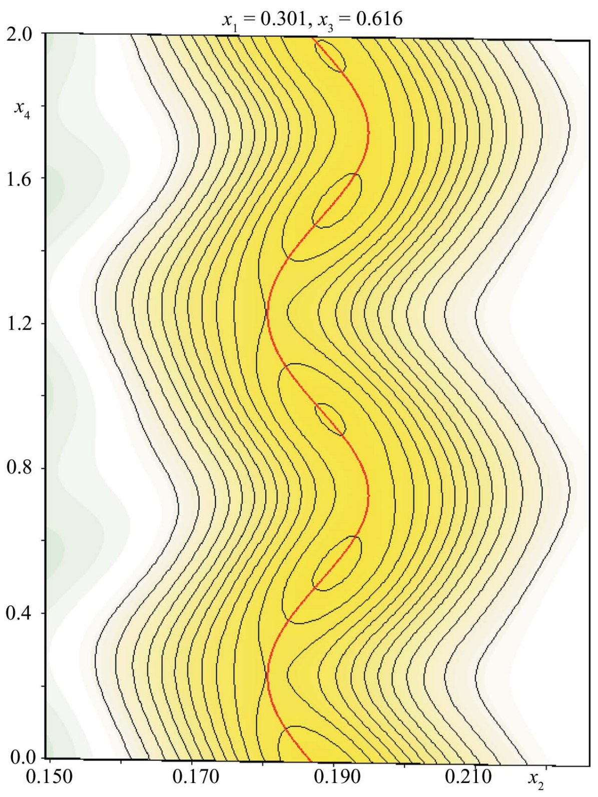

Figure 11
A section of superspace electron density for the N atom of residue 31 in chain *A* of ToxD, showing density along x_2 as a function of the modulation x_4. The atomic modulation function representing the position of the atom within the superspace density is shown by the red line.

4. Conclusions

A more robust simulated modulated protein data set has been created beyond what has been performed previously. This data set uses the same superspace group as the incommensurately modulated profilin–actin crystals. The modulation was performed in such a way as to preserve stereochemical constraints. The data set consists of a true three-dimensional modulation along one of the crystal axes. The modulation was performed in a commensurate fashion so that it can be used to evaluate the software modifications that will be required in order to allow the refinement of modulated protein structures. The data are available as a collection of text files that can be imported into other processing packages. The simulation was successfully imported into *Jana*2006, as shown by the inclusion of the two-dimensional slice from the four-dimensional superspace electron density (Fig. 11). We are now faced with the challenge of modifying *Jana*2006 to refine a protein structure using stereochemical restraints and the other algorithms that protein crystallographers use to deal with the relatively low-resolution diffraction data collected from protein crystals (noncrystallographic symmetry restraints, use of rigid groups *etc.*).

In the future, this data set could be used to work through the problems that will be encountered in the development of software to refine and solve modulated structures. With the simulation correctly imported into *Jana*2006, it should also be possible to simulate and study incommensurate modulations. Using incommensurately simulated data would allow us to better understand how well a supercell approximation to an incommensurately modulated data set will refine by varying the modulation closer or further from the closest commensurate case.

As a final thought, at present there is no accepted format to store modulated protein data. There are two options available. The first would be to extend the PDB format (Bernstein *et al.*, 1977) to store the atoms, modulation functions, unit cell and superspace group information. Proposing an extension to the PDB format is nontrivial. Additionally, this approach would still require a separate file to store the structure factors. The second approach would be to use CIF (Hall *et al.*, 1991). CIF has the advantage that in principle all of the information can be stored in one file. In practice, the mmCIF and msCIF dictionaries need to be blended into the same file or a version of the msCIF needs to be adapted to DDL2 for proteins.

This work was funded by the National Science Foundation (grant MCB-0718661), the Nebraska Research Initiative and the Eppley Cancer Center (support grant P30CA036727). We would like to personally thank Professor Sander van Smaalen (Bayreuth University) and to thank the aperiodic crystallographic community as a whole for useful discussions and encouragement.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Dauter, Z. & Jaskolski, M. (2010). *J. Appl. Cryst.* **43**, 1150–1171.

Gradshteyn, I. S. & Ryzhik, I. M. (2007). *Table of Integrals, Series, and Products.* New York: Academic Press.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.

Janssen, T., Janner, A., Looijenga-Vos, A. & Wolff, P. M. D. (1999). *International Tables for Crystallography*, Vol. *C*, edited by A. J. C. Wilson & E. Prince, pp. 899–947. Dordrecht: Kluwer Academic Publishers.

Lovelace, J. J., Murphy, C. R., Daniels, L., Narayan, K., Schutt, C. E., Lindberg, U., Svensson, C. & Borgstahl, G. E. O. (2008). *J. Appl. Cryst.* **41**, 600–605.

Lovelace, J. J., Winn, M. D. & Borgstahl, G. E. O. (2010). *J. Appl. Cryst.* **43**, 285–292.

Porta, J., Lovelace, J. J., Schreurs, A. M. M., Kroon-Batenburg, L. M. J. & Borgstahl, G. E. O. (2011). *Acta Cryst.* D**67**, 628–638.

Schreurs, A. M. M., Xian, X. & Kroon-Batenburg, L. M. J. (2010). *J. Appl. Cryst.* **43**, 70–82.

Shannon, C. E. (1949). *Proc. Inst. Radio Eng.* **37**, 10–21.

Skarzyński, T. (1992). *J. Mol. Biol.* **224**, 671–683.

Smaalen, S. van (2007). *Incommensurate Crystallography*. Oxford University Press.

Smaalen, S. van (2005). *Z. Kristallogr.* **219**, 681–691.

Vila-Sanjurjo, A., Schuwirth, B.-S., Hau, C. W. & Cate, J. H. D. (2004). *Nature Struct. Mol. Biol.* **11**, 1054–1059.

Wagner, T. & Schönleber, A. (2009). *Acta Cryst.* B**65**, 249–268.

Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.

Wolff, P. M. de, Janssen, T. & Janner, A. (1981). *Acta Cryst.* A**37**, 625–636.

Zwart, P. H., Grosse-Kunstleve, R. W., Lebedev, A. A., Murshudov, G. N. & Adams, P. D. (2008). *Acta Cryst.* D**64**, 99–107.