Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structure-factor extrapolation using the scalar approximation: theory, applications and limitations

Ulrich K. Genick‡

Brandeis University, Department of Biochemistry, Waltham, MA 02454, USA

‡ Current address: Nestle Research Center, Vers-chez-les-Blancs, Switzerland.

Correspondence e-mail: ulrich.genick@rdls.nestle.com

Received 21 February 2007
Accepted 3 August 2007

For many experiments in macromolecular crystallography, the overall structure of the protein/nucleic acid is already known and the aim of the experiment is to determine the effect a chemical or physical perturbation/activation has on the structure of the molecule. In a typical experiment, an experimenter will collect a data set from a crystal in the unperturbed state, perform the perturbation (*i.e.* soaking a ligand into the crystal or activating the sample with light) and finally collect a data set from the perturbed crystal. In many cases the perturbation fails to activate all molecules, so that the crystal contains a mix of molecules in the activated and native states. In these cases, it has become common practice to calculate a data set corresponding to a hypothetical fully activated crystal by linear extrapolation of structure-factor amplitudes. These extrapolated data sets often aid greatly in the interpretation of electron-density maps. However, the extrapolation of structure-factor amplitudes is based on a mathematical shortcut that treats structure factors as scalars, not vectors. Here, a full derivation is provided of the error introduced by this approximation and it is determined how this error scales with key experimental parameters. The perhaps surprising result of this analysis is that for most structural changes encountered in protein crystals, the error introduced by the scalar approximation is very small. As a result, the extrapolation procedure is largely limited by the propagation of experimental uncertainties of individual structure-factor amplitudes. Ultimately, propagation of these uncertainties leads to a reduction in the effective resolution of the extrapolated data set. The program *XTRA*, which implements SASFE (scalar approximation to structure-factor extrapolation), performs error-propagation calculations and determines the effective resolution of the extrapolated data set, is further introduced.

1. Introduction

For a growing number of crystallographic studies, the goal is not the determination of a new crystal structure. Instead, the goal is to understand how a known structure responds to a chemical or physical stimulus; for example, the binding of a small molecule, exposure to light or changes in pH, redox potential *etc*. A combination of technical and basic physical factors often limits the efficiency of the perturbation and only a fraction of the unit cells in a crystal respond to the stimulus. The experimenter then obtains pairs of data sets where the first data set corresponds to a crystal with 100% of unit cells in their 'native' state *A* and the second data set from a crystal containing a mix of unit cells, some remaining in state *A* and some in the 'perturbed' state *B*. While difference electron-

density maps calculated from such pairs of data sets provide useful information, model building often benefits greatly from a data set corresponding to a fully activated crystal (*i.e.* a crystal in which all molecules are in state *B*).

1.1. Vectorial *versus* scalar extrapolation of structure factors

Given experimental structure-factor amplitudes $|F_A|$ from a 'native' crystal and $|F_{AB}|$ from a crystal containing a mix of unit cells in state *A* and *B*, can one not simply determine the structure-factor amplitudes for a fully activated crystal $|F_B|$ from linear extrapolation of these amplitudes? Strictly speaking, the answer is 'no'. Structure factors are vector quantities. In the polar-coordinate representation, structure factors have two components: the structure-factor amplitude $|F|$ and the phase φ. Unless both amplitudes and phases are known (and the phases for the partially activated state are usually not known), proper vector extrapolation is not possible.

To bypass this problem, Genick *et al.* (1997) proposed $|F_{B\mathrm{ex}}|$ as an approximation of $|F_B|$, where $|F_{B\mathrm{ex}}|$ for each structure factor is obtained by simple linear extrapolation of the structure-factor amplitudes $|F_B|$ and $|F_{AB}|$,

$$|F_{B\mathrm{ex}}| = (|F_{AB}| - |F_A|) \times \frac{1}{f} + |F_A|. \qquad (1)$$

This equation uses only $|F_A|$ and $|F_{AB}|$, which are observed experimentally, and *f*, which is the fraction of unit cells that adopt state *B* in the partially activated crystal (*i.e.* the crystal that gave rise to $|F_{AB}|$). The value of *f* can often be obtained from spectroscopic measurements of the crystal or from the height of electron-density peaks in F_A and $F_{AB} - F_A$ maps.

Despite its apparent heavy-handedness, the approximation indicated in (1), which I refer to as SASFE (scalar approximation to structure-factor extrapolation), has proven to be remarkably successful and popular (for examples, see Genick *et al.*, 1997; Edman *et al.*, 1999, 2002; Cao *et al.*, 1998; Lanyi & Schobert, 2007).

Here, I report a systematic analysis of the error introduced by SASFE and how this error compares with other types of errors commonly encountered (and tolerated) in protein crystallography. I also analyze the effect of experimental errors on the final extrapolated data set and present the program *XTRA* that implements the structure-factor extrapolation, error propagation and statistical analysis of the extrapolation results in a stand-alone program.

1.2. Requirements for the distribution of activated molecules within the crystal

Before launching into an analysis of the extrapolation procedure, it is important to point out that the proposed treatment assumes that molecules in state *A* and *B* are distributed throughout the crystal such that photons scattered by molecules in both states interfere in the formation of the diffraction pattern. If, instead, the crystal were to consist of a small number of large domains, some containing only molecules in form *A* and others containing only molecules in form *B*, such that photons diffracted from the respective domains did not interfere, then the proper extrapolation would be the scalar extrapolation of scattered intensities (*i.e.* $|F|^2$). In practice, in-crystal activation experiments on protein crystals generally appear to generate the former scenario, so that the treatment outlined below will be applicable.

2. Source and magnitude of the error introduced by SASFE

In this section, I discuss the origin of the error introduced by SASFE and how this error depends on the size of the structural change between state *A* and state *B* and on the fraction of molecules converted from state *A* to state *B*.

In order to understand the source of the SASFE error, it is useful to recall how the position of individual atoms affects a given structure factor. The radiation scattered by an individual atom can be described as a wave with a characteristic amplitude and phase. The amplitude depends on the atom's number of electrons and the phase depends on the atom's position relative to all other atoms. The molecular structure factor is simply the sum of those individual waves. This addition of waves can be represented in an Argand diagram, in which each scattered wave is represented as a vector in the complex plane and the wave sum is calculated as the vector sum of the

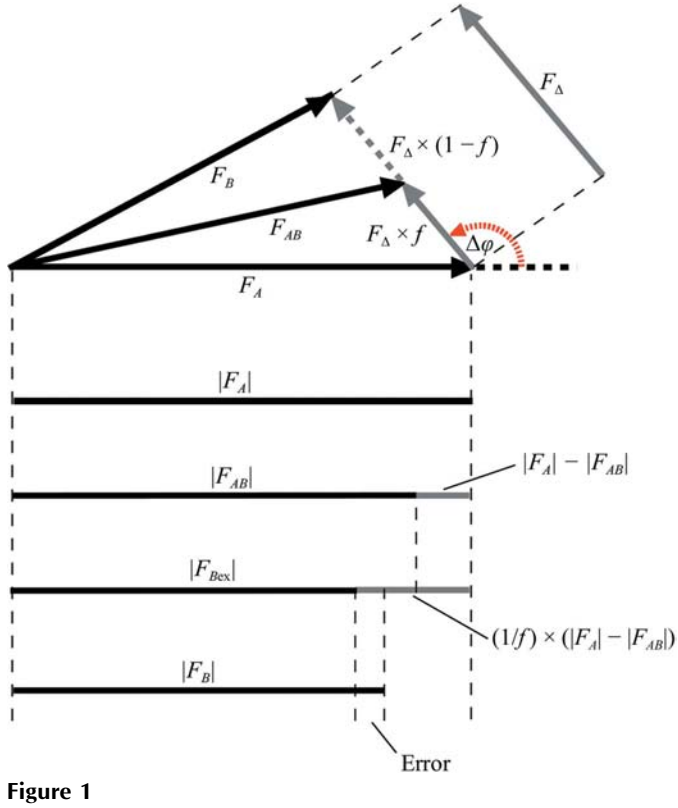

Figure 1
Graphical comparison of SASFE to the the proper vector extrapolation of a structure factor. The example is for the case of a crystal where F_{AB} corresponds to an activation of 50% of unit cells (*i.e.* $f = 0.5$). Therefore, $F_\Delta \times f = F_\Delta \times (1 - f)$ and $|F_A| - |F_{B\mathrm{ex}}| = 2 \times |F_A| - |F_{AB}|$. Vector extrapolation is shown at the top and the scalar approximation at the bottom. The lengths of the bars representing the amplitude of the structure factors are identical in the top and bottom half of the figure.

individual atomic wavevectors. If some of the atoms move, the phases of their atomic scattering vectors change and as a result the molecular structure factor changes. This change can be described as a difference vector. If the difference vector that corresponds to the perturbation of a given fraction f of a crystal's unit cells is known, then the difference vector F_Δ corresponding to the full conversion can be calculated by multiplying this difference vector by $1/f$.

Fig. 1 shows a graphical comparison of a proper vector-based structure-factor extrapolation and SASFE. From the diagram, it is clear that for $\Delta\varphi = 0$ ($\Delta\varphi$ is the difference in phase between F_A and F_Δ) SASFE is identical to the proper vector-based extrapolation and the error ($|F_B| - |F_{B\mathrm{ex}}|$) is therefore zero. When $\Delta\varphi = 180^\circ$ the error is also zero, as long as $F_A > F_\Delta$. However, if $F_{AB} < F_A < F_\Delta$, $\Delta\varphi = 180^\circ$ results in a negative $F_{B\mathrm{ex}}$ and the error is maximal.

To better illustrate the magnitude and behavior of the error, Fig. 2 shows $|F_{AB}|$, $|F_{B\mathrm{ex}}|$, $|F_B|$ and $|F_B| - |F_{B\mathrm{ex}}|$ as a function of $\Delta\varphi$ for two relative magnitudes of F_A and F_Δ. Given the heavy-handedness of the assumption underlying SASFE, the error is surprisingly small. It is also notable that SASFE systematically underestimates structure-factor amplitudes (*i.e.* $|F_B| \geq |F_{B\mathrm{ex}}|$ for all $\Delta\varphi$).

Figs. 3(*a*) and 3(*b*) show how this error (averaged over $\Delta\varphi$) varies as a function of the fraction of activated molecules f and the size of the structural change $|F_\Delta|/|F_A|$.

2.1. Treating $|F_A|$ and $|F_\Delta|$ as distributions amplifies the SASFE error

For any real-world example, the $|F_A|$ and $|F_\Delta|$ values do not adopt fixed values but distributions. In other words, for a given average size of $|F_A|$ and $|F_\Delta|$ we will find combinations of large $|F_A|$ with small $|F_\Delta|$, small $|F_A|$ with large $|F_\Delta|$ *etc.* How does this affect the expected error for given average values of $|F_\Delta|$ and $|F_A|$?

The calculation of $|F_{B\mathrm{ex}}|$ for any reflection with a given Miller index (*hkl*) involves only the $|F_A|$ and $|F_\Delta|$ with the same Miller index, *i.e.* the scattering angle is the same. So any scattering angle-dependent effects can be ignored. This simplifies our considerations and structure factors can be treated in their unitary form. I also focus solely on acentric reflections. The reason for this is simple, with phase angles restricted to 0 and π the SASFE of centric reflections is (at least in the vast majority of cases) identical to the proper vector extrapolation. For acentric reflections, it can be shown (Giacovazzo *et al.*, 2002) that the distribution of unitary structure-factor amplitudes mirrors

$$P(|E|) = 2|E| \exp(-|E|^2). \qquad (2)$$

This function approaches zero for $|E| > 3\langle|E|\rangle$, where it is customarily truncated. By scaling $|F_A|$ and $|F_\Delta|$ by the root-

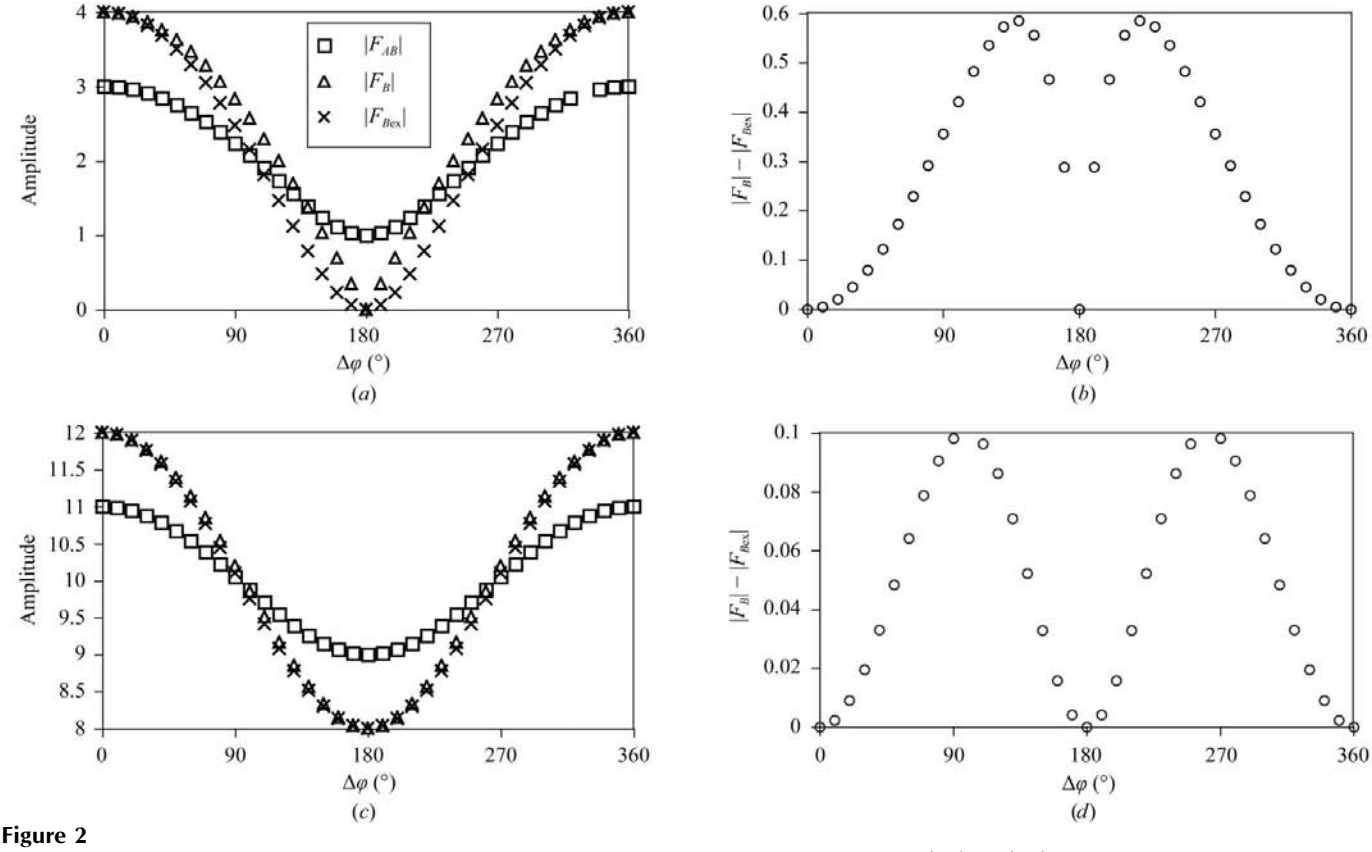

Figure 2
Error introduced by the scalar approximation as a function of the angle $\Delta\varphi$ and the relative sizes of $|F_A|$ and $|F_\Delta|$ ($F_\Delta = F_B - F_A$). The fraction of unit cells in state B is assumed to be 50% (*i.e.* $f = 0.5$). The nomenclature is identical to that used in Fig. 1. (*a*) and (*b*) correspond to $|F_A| = |F_\Delta|$ and (*c*) and (*d*) correspond to $|F_A| = 5 \times |F_\Delta|$, with $|F_\Delta| = 2$ in both cases. (*a*) and (*c*) show $|F_{AB}|$, the true structure-factor amplitude $|F_B|$ and the extrapolated amplitude $|F_{B\mathrm{ex}}|$. (*b*) and (*d*) show the error (*i.e.* $|F_B| - |F_{B\mathrm{ex}}|$) generated by the scalar approximation. Note the differences in vertical scale.

mean-square amplitudes of their distributions, we can then calculate the probability distribution of $|F_A|$ and $|F_\Delta|$ values

$$P(|F_A|) = 2\frac{|F_A|}{\text{r.m.s.}(F_A)}\exp\left\{-\left[\frac{|F_A|}{\text{r.m.s.}(F_A)}\right]^2\right\} \quad (3)$$

and

$$P(|F_\Delta|) = 2\frac{|F_\Delta|}{\text{r.m.s.}(F_\Delta)}\exp\left\{-\left[\frac{|F_\Delta|}{\text{r.m.s.}(F_\Delta)}\right]^2\right\}. \quad (4)$$

To obtain the average error caused by performing SASFE on these distributions of structure factors, one then needs to consider all possible permutations of $|F_A|$ and $|F_\Delta|$ values, calculate the SASFE errors corresponding to each permutation and weight these errors for the relative probability of finding that combination of $|F_A|$ and $|F_\Delta|$. The results of this calculation are shown in Fig. 3(*c*). For $|F_\Delta|/|F_A|$, the treatment of $|F_\Delta|$ and $|F_A|$ as distributions moderately increases the average SASFE error. This effect can be understood by considering the graph in Fig. 3(*b*). This graph indicates that the SASFE error is disproportionately large when $|F_\Delta| \simeq 2|F_A|$. While it is unlikely that the average $|F_\Delta|$ will ever approach twice the size of $|F_A|$, variations in $|F_A|$ and $|F_\Delta|$ within their respective distributions will occasionally result in combinations of $|F_A|$ and $|F_\Delta|$ values such that $|F_\Delta| \simeq 2|F_A|$ and these combinations boost the average error.

When plotted as a function of $\langle F_\Delta\rangle$ (see Fig. 3*c*), the distribution-corrected SASFE error adopts a simple monotonic curve. For $\langle F_\Delta\rangle/\langle F_A\rangle \leq 0.4$, calculated curves of the average error approximate a simple quadratic function $\langle|F_B| - |F_{B\text{ex}}|\rangle = k\langle F_\Delta\rangle^2/\langle F_A\rangle$. From Fig. 3(*a*) we know that $\langle|F_B| - |F_{B\text{ex}}|\rangle$ is also linearly proportional to $1 - f$, so we expect $\langle|F_B| - |F_{B\text{ex}}|\rangle = (1 - f)k'\langle F_\Delta\rangle^2/\langle F_A\rangle$. The optimal value of k' was determined by least-squares fits of numerically

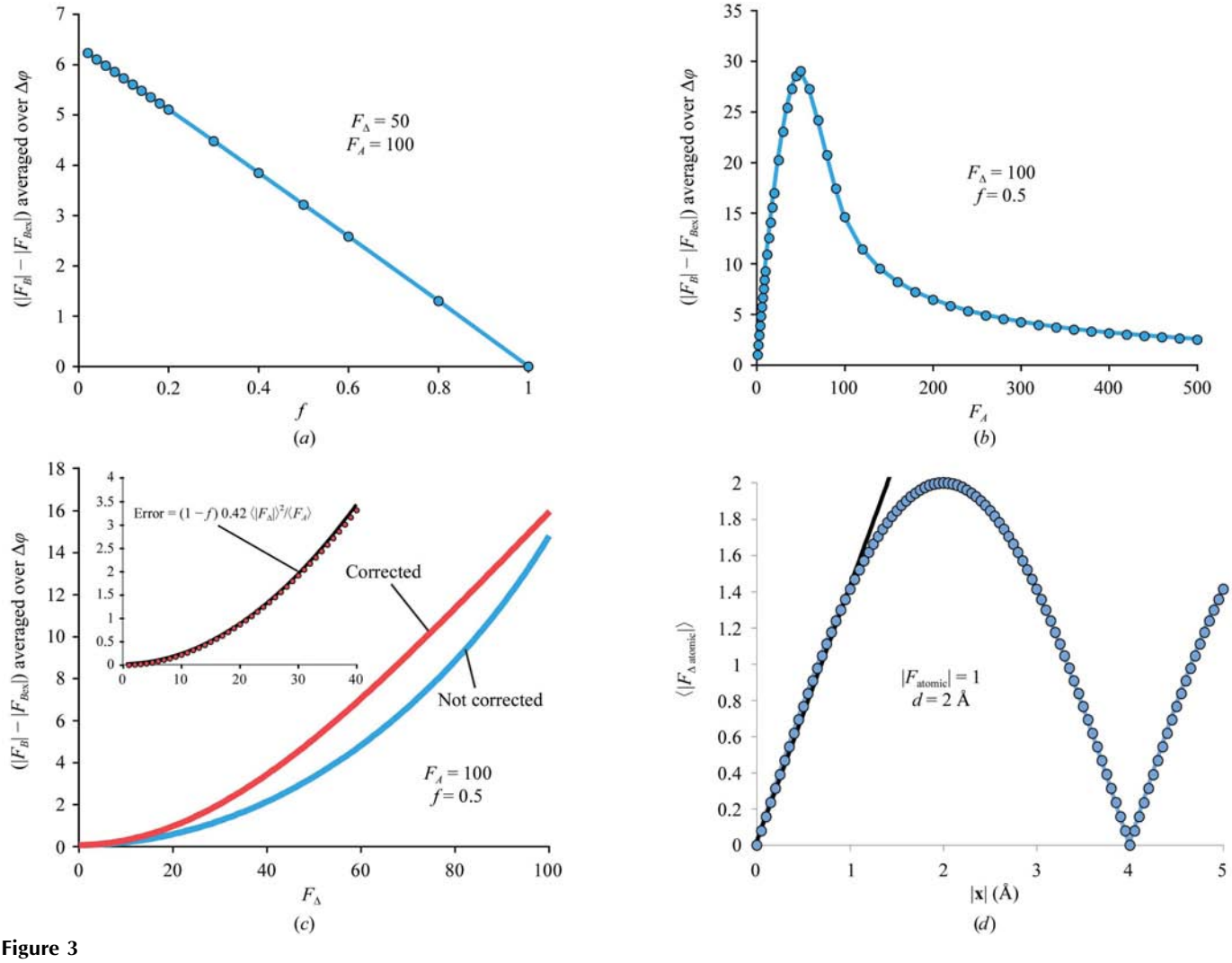

Figure 3
Dependence of the SASFE error (*i.e.* $|F_B - F_{B\text{ex}}|$ averaged over all $\Delta\varphi$) (*a*) as a function of f and (*b*) as a function of the relative size of F_A and F_Δ. (*c*) shows the effect of performing the error calculation not only on the average values of F_A and F_Δ, but also taking into account that both F_A and F_Δ adopt distributions described by (4). The inset in (*c*) shows that for $F_\Delta/F_A < 0.4$ the distribution-corrected curve can be approximated by a quadratic function. In (*d*) the graph represented in blue circles shows how $F_{\Delta\,\text{atomic}}$ varies with the size of random atomic movements of magnitude **x** (see equation 9 and §2.3 for discussion). The plot shows the case of a set of Bragg planes separated by 2 Å and $F_{\text{atomic}} = 1$. The plot shows that for small **x** this function approximates a straight line (shown in solid black) calculated by (10).

determined curves of $\langle|F_B| - |F_{B\mathrm{ex}}|\rangle$ *versus* $\langle F_\Delta\rangle$ for various combinations of $\langle F_\Delta\rangle$ and f. These fits were optimal for $k' = 0.42$ across all values of $\langle F_\Delta\rangle$ and f. The inset in Fig. 3(*c*) shows the quality of the fit for the case of $F_A = 100$ and $f = 0.5$. By dividing both sides of the equation by $\langle F_\Delta\rangle$, we then obtain

$$\frac{\langle F_{B\mathrm{ex}} - F_B\rangle}{\langle F_\Delta\rangle} \simeq 0.42(1-f)\frac{\langle F_\Delta\rangle}{\langle F_A\rangle} \qquad (5)$$

as a simple formula to estimate the SASFE error, *i.e.* the difference between the proper vector extrapolation and the SASFE result $\langle|F_B| - |F_{B\mathrm{ex}}|\rangle$ divided by the size of the structural change $\langle F_\Delta\rangle$.

With this estimator in hand, one can now consider how the ratio $\langle F_\Delta\rangle/\langle F_A\rangle$ varies for different kinds of protein structural changes and then determine how good or bad an approximation SASFE provides in these cases.

2.2. Magnitude of error case 1: binding of a small molecule

Proteins typically contain several thousand non-H atoms. Scattering from each of these atoms contributes to every structure factor in a data set. The amplitude of each atom's contribution to any molecular structure factor is determined by the atom's atomic number, its occupancy and the B factor. The phase of the atomic contribution is determined by the atom's position in the unit cell.

If, for the sake of this analysis, one assumes that for a given scattering angle the amplitudes of each atom's contribution are the same and that the atoms are distributed randomly within a crystal structure, then the Argand diagrams for individual structure factors will take the form of random walks with step size $|F_{\mathrm{atomic}}|$ and a number of steps equal to the number of atoms. Since the average distance between the start and end of a random walk is ~0.8 times the square root of the number of steps times the size of the step, the average structure factor for a protein with n atoms will be $|\overline{F}_{\mathrm{protein}}| \simeq 0.8n^{1/2}_{\mathrm{protein}}|F_{\mathrm{atomic}}|$.

In the case where the transition from state A to state B is simply the binding of a small molecule, $|\overline{F}_A| = |\overline{F}_{\mathrm{protein}}|$ and $|\overline{F}_\Delta| = |\overline{F}_{\mathrm{small\_mol.}}| \simeq 0.8n^{1.2}_{\mathrm{small\_mol.}}|F_{\mathrm{atomic}}|$ and the relative amplitude of the average structure factors is then given by

$$\frac{|\overline{F}_\Delta|}{|\overline{F}_A|} = \frac{|\overline{F}_{\mathrm{small\_mol.}}|}{|\overline{F}_{\mathrm{protein}}|} = \left(\frac{n_{\mathrm{small\_mol.}}}{n_{\mathrm{protein}}}\right)^{1/2} \simeq \left(\frac{\mathrm{MW}_{\mathrm{small\_mol.}}}{\mathrm{MW}_{\mathrm{protein}}}\right)^{1/2}. \qquad (6)$$

For the case of molecular weights of 30 000 Da for the protein and 900 Da for the small molecule and a 25% fraction of activated molecules, one can use (5) to estimate the expected average SASFE error $\langle F_{B\mathrm{ex}} - F_B\rangle/\langle F_\Delta\rangle$ as 5.45%. In other words, the error $\langle F_{B\mathrm{ex}} - F_B\rangle$ introduced by the SASFE approximation is rather small when compared with the magnitude of the structural change $\langle F_\Delta\rangle$.

2.3. Magnitude of error case 2: subtle structural changes in large parts of a molecule

Next, let us consider a scenario where the change from state A to state B involves a movement of existing atoms, *i.e.* a conformational change in the protein. In this case, one can think of a protein as composed of two substructures, one which stays constant and a second containing n_{mobile} atoms that move in the A-to-B transition. Using the same rationale as above, one can then think of the structure-factor contributions of the mobile part to state A and state B as two separate random walks starting from a common origin. Then, $|F_\Delta|$ is the distance between the endpoints of the two random walks and

$$|\overline{F}_\Delta| \simeq |F_{\mathrm{atomic}}| \times 0.8 \times 2^{1/2} n^{1/2}_{\mathrm{mobile}}. \qquad (7)$$

Since protein conformational changes often involve movements, albeit subtle, of a large fraction of the atoms in the unit cell, the estimation of $|\overline{F}_\Delta|$ according to (7) would predict $|F_\Delta|/|F_A|$ ratios and SASFE errors that are quite large. For example, a protein structural change in which 20% of all side chains move would correspond to an $|F_\Delta|/|F_A|$ ratio of

$$\frac{|\overline{F}_\Delta|}{|\overline{F}_A|} = 2^{1/2}\left(\frac{0.2}{1}\right)^{1/2} = 0.63.$$

For this example and the assumption of $f = 0.25$, (5) then predicts a very substantial error of $(|F_B| - |F_{B\mathrm{ex}}|)/|F_\Delta| = 20\%$.

However, (7) assumes a random reshuffling of atomic positions for the mobile portion of the molecule. However, in most protein conformational changes, and certainly in those that are accommodated by a crystal lattice, the majority of atoms will undergo only subtle positional adjustments and such subtle adjustments should lead to smaller F_Δ and consequently smaller errors. The goal of the following section is to derive a rough estimate of the relationship between the amplitude of the atomic displacement and the magnitude of the error introduced by SASFE within the limit of small structural changes. The simplified treatment below treats this effect within the same framework of random walks developed above and while this treatment makes several simplifying assumptions, it should suffice for the specific argument presented here. For a more thorough and extensive treatment of the relationship between atomic motions and resulting structure-factor amplitudes, the reader is pointed to the excellent work on this topic by Read (*e.g.* Read, 1990).

Above, we arrived at (7) by describing $|F_\Delta|$ as a random walk of $2n$ steps with a step size of $|F_{\mathrm{atomic}}|$. This result is mathematically equivalent to a random walk of n steps with step size $2^{1/2}|F_{\mathrm{atomic}}|$, where n is the number of moving atoms and $2^{1/2}|F_{\mathrm{atomic}}|$ is the average size of the difference scattering vector $|F_{\Delta\,\mathrm{atomic}}|$ resulting from the movement of an individual atom. For the case of a structural change that involves only small atomic motions, the number of steps in this random walk that makes up $|F_\Delta|$ will still be the number of moving atoms, but we expect the average step size (*i.e.* the average $|F_{\Delta\,\mathrm{atomic}}|$) to be smaller and this decrease in step size reduces the size of $|F_\Delta|$ compared with the case of atomic movements with random amplitude.

To determine how the size of $|F_{\Delta\,\mathrm{atomic}}|$ depends on the size of an atom's movement, let us consider the case of a crystal in which the A-to-B transition consists of the movement of only a

single atom by vector **x**. Further, let us focus on a single reflection originating from a set of Bragg planes with spacing d. If the movement vector **x** has a component p that is perpendicular to this set of Bragg planes, this movement will result in the rotation of the atomic scattering vector in the plane of the Argand diagram. Let the angle of that rotation be $\Delta\varphi$. A movement that displaces an atom from one Bragg plane to the next (*i.e.* $p = d$) corresponds to a full 2π rotation. Smaller motions result in proportionally smaller changes in $\Delta\varphi$.

Since we stipulate that **x** is randomly oriented relative to the Bragg plane, p (*i.e.* the component of **x** that is perpendicular to the Bragg plane) will be on average $|\mathbf{x}|/2$; therefore,

$$\langle|\Delta\varphi|\rangle = 2\pi\frac{p}{d} = \pi\frac{|\mathbf{x}|}{d}. \tag{8}$$

We can then use the fact that $F_{\Delta\,\mathrm{atomic}}$, $F_{A\,\mathrm{atomic}}$ and $F_{B\,\mathrm{atomic}}$ form a triangle and the law of cosines to calculate $|F_{\Delta\,\mathrm{atomic}}|$ for a set of Bragg planes with spacing d and a random movement of an atom by distance $|\mathbf{x}|$ as

$$\begin{aligned}\langle|F_{\Delta\,\mathrm{atomic}}|\rangle &= (|F_{\mathrm{atomic}}|^2 + |F_{\mathrm{atomic}}|^2 - 2|F_{\mathrm{atomic}}| \\ &\quad \times |F_{\mathrm{atomic}}|\cos\langle\Delta\varphi\rangle)^{1/2} \\ &= |F_{\mathrm{atomic}}|[2(1-\cos\langle\Delta\varphi\rangle)]^{1/2} \\ &= |F_{\mathrm{atomic}}|\left\{2\left[1-\cos\left(\pi\frac{|\mathbf{x}|}{d}\right)\right]\right\}^{1/2}\end{aligned} \tag{9}$$

A plot of (9) reveals (Fig. 3*d*) that for atomic displacements smaller than half of the Bragg spacing d, $\langle|F_{\Delta\,\mathrm{atomic}}|\rangle$ is consistently smaller than the size of $2^{1/2}F_{\mathrm{atomic}}$ resulting from an atomic movement of random amplitude. The plot further shows that in the regime of small displacements where $0 < |\mathbf{x}| < d/2$, $\langle|F_{\Delta\,\mathrm{atomic}}|\rangle$ can be approximated by a simple linear equation

$$\langle|F_{\Delta\,\mathrm{atomic}}\rangle = |F_{\mathrm{atomic}}|\frac{2|\mathbf{x}|}{d}2^{1/2}. \tag{10}$$

For the case of a protein structural change in which the majority of atoms move by a distance of less than half the resolution of the structure, the average length of $F_{\Delta\,\mathrm{atomic}}$ and with it the average length of F_Δ are linearly proportional to the distance of the movement, inversely proportional to the resolution of the reflection and consistently smaller than the F_Δ for a protein structural change involving atomic motions by random distances. A compact mathematical expression for the expected SASFE error that includes the appropriate correction for the case of small atomic movements is given below in (13).

If we apply this correction for small atomic displacements to the case of the protein structural change involving 20% of all side chains and $f = 0.25$, which we discussed above, and assume an average atomic displacement of 0.2 Å, then the SASFE error for reflections with d spacings of 2 Å would be reduced to ∼4%. For reflections with a d spacing of 1.5 Å the predicted error increases slightly to 5.3%, but in either case the error is much smaller than the 20% value we had obtained assuming movements of random amplitude.

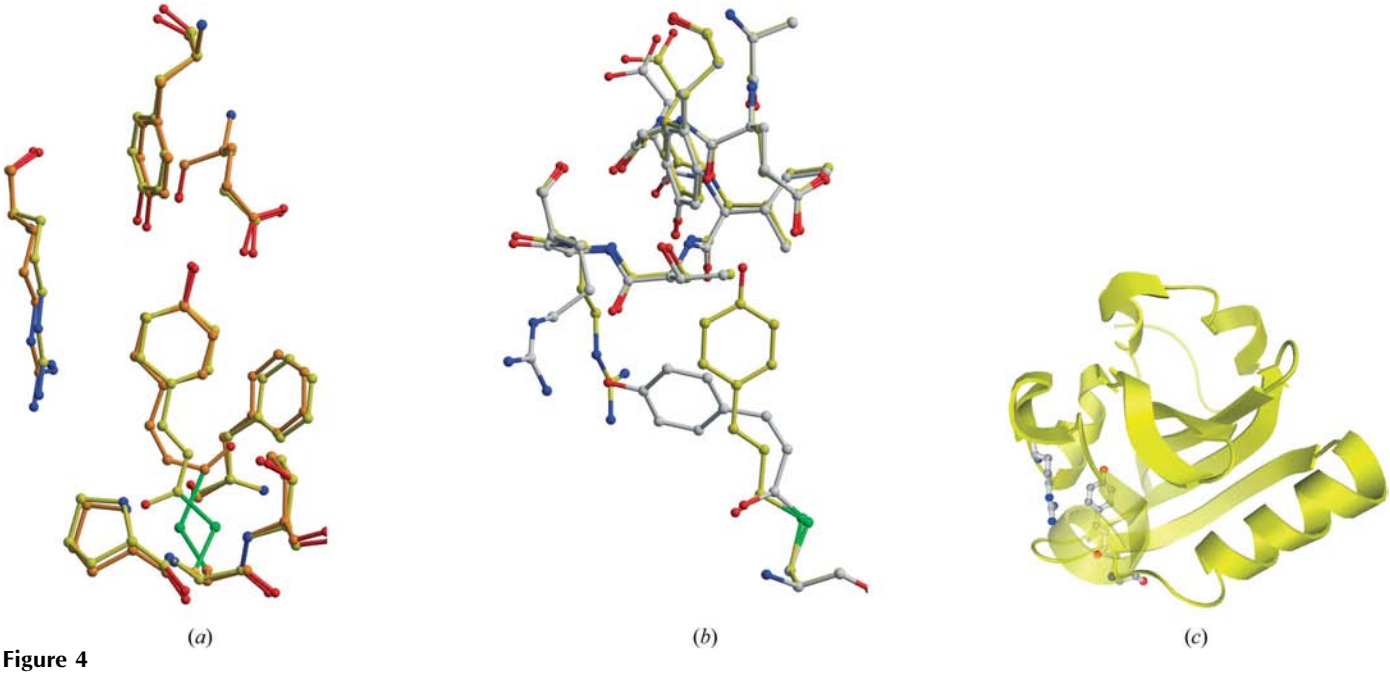

Figure 4
Structural transformations in the bacterial photoreceptor PYP. (*a*) shows the initial chromophore isomerization reaction. This reaction serves as an example of a subtle structural change. PYP's dark-adapted conformation is shown in yellow and the initial photoproduct (I1) is shown in orange. Only atoms within the chromophore undergo substantial structural displacements. (*b*) shows a subsequent structural transformation in which the protein and chromophore rearrange around one another in a more extensive structural change. The dark-adapted conformation (as in *a*) is again shown in yellow, while the light-activated state (I2) is shown in grey. (*c*) shows a ribbon diagram of PYP including a ball-and-stick representation of the chromophore and a neighbouring arginine residue to provide an overall scale for the size of the two structural transitions relative to the size of the protein.

This resolution-dependence of the SASFE error that occurs in the cases of subtle structural changes means that higher resolution reflections are more strongly affected by SASFE error, so that the extrapolated data set will have a lower effective resolution than the data sets from which it has been extrapolated. However, as we will see below, the decrease in effective resolution of extrapolated maps is usually dominated not by this effect, but by the experimental errors of $|F_A|$ and $|F_{AB}|$ and the amplification of these errors by the extrapolation procedure.

To conclude this section, here are the three formulae for the estimation of the SASFE error for the three scenarios discussed above. These equations are an approximation of the error and are most accurate when the square-root term is <0.4.

Small-molecule binding:

$$\frac{\langle F_{B\mathrm{ex}} - F_B \rangle}{\langle F_\Delta \rangle} \simeq 0.42(1-f)\left(\frac{n_{\mathrm{small\_mol.}}}{n_{\mathrm{protein}}}\right)^{1/2}. \qquad (11)$$

Random conformational change:

$$\frac{\langle F_{B\mathrm{ex}} - F_B \rangle}{\langle F_\Delta \rangle} \simeq 0.42(1-f)2^{1/2}\left(\frac{n_{\mathrm{move}}}{n_{\mathrm{total}}}\right)^{1/2}. \qquad (12)$$

Subtle conformational changes with atomic movements smaller than $d/2$:

$$\frac{\langle F_{B\mathrm{ex}} - F_B \rangle}{\langle F_\Delta \rangle} \simeq 0.42(1-f)\frac{2|\mathbf{x}|2^{1/2}}{d}\left(\frac{n_{\mathrm{move}}}{n_{\mathrm{total}}}\right)^{1/2}. \qquad (13)$$

2.4. Effect of correlation between atomic motions

Throughout this section, the discussion has assumed that atomic motions are random and uncorrelated. Since many of the individual atoms in biological macromolecules are covalently linked to their neighbours and must therefore move together, this assumption of completely random motions seems unrealistic. Even nonlinked atomic neighbors will, owing to their close physical packing, show some degree of correlation in atomic motions. What is the effect of such correlation on the SASFE error? Staying within the framework developed above, correlation of atomic motions will transform the random walk of the atomic difference scattering vectors that make up F_Δ into a biased walk. With an increasing degree of correlation, the length of this walk will increase from scaling with $n^{1/2}$ for the case of uncorrelated motions to scaling with n for a perfectly correlated linear translation of all atoms. For structural changes involving a large number of atoms, this would predict a very substantial increase in F_Δ and potentially a dramatic failure of the SASFE procedure. However, as we will see below, the assumption of random movements delivers adequate estimates of the SASFE error even for examples where the atomic motions are partially correlated.

Why may the effect of correlation between atomic motions be less severe than anticipated by the above picture? While a full treatment of the effect of correlation of atomic motions on the SASFE error would exceed the frame of this work, two factors appear to limit the detrimental effect of correlated atomic motions on the SASFE procedure. Firstly, as mentioned above, the maximal effect of the correlation is only observed if the correlation takes the form of a pure translation of all atoms by an identical displacement vector. From simple considerations of molecular geometry and molecular packing such perfect translations would be expected to be quite rare. For example, the pure translations of a large number of atoms without compensating motions of other atoms would result in the generation of a vacuum. Most structural transitions also involve rotations of molecular groups rather than translations, so that atomic displacements are not strictly correlated and may in fact be anticorrelated.

Secondly, the perfectly correlated translation of a substructure represents a fundamentally different case than the random-walk example discussed above. Such a perfectly correlated translation corresponds to the rotation of the scattering vectors of each of those atoms by exactly the same angle. The effect is then equivalent to the rotation of the scattering vector for the entire substructure by this same angle, but the length of the substructure's scattering vector remains the same. In contrast, in the case of random motions, both the direction and length of the scattering vector for the substructure of moving atoms changes at random and both these changes contribute to the overall size of F_Δ.

3. SASFE applied to two examples of protein conformational change

In the previous section, I examined the source and estimated the magnitude of the SASFE error from first principles. Here, two examples of real protein conformational changes taken from the literature are analyzed. The first example is the initial chromophore double-bond isomerization reaction of the bacterial photoreceptor protein PYP. This reaction involves the movement of just a handful of atoms, most of which move by just a fraction of an angstrom. The second example is a subsequent protein conformational change of PYP in which the chromophore and an arginine side chain move by several angstroms and a handful of additional side chains move by fractions of an angstrom. The goal of this section is to examine whether the error generated by the scalar approximation limits the usefulness of SASFE in these two real-world examples.

3.1. Extrapolation procedure and analysis of errors

The starting points for the calculations are two published structural transformations in the bacterial photoreceptor protein PYP (Fig. 4). To test the error SASFE generates for these real-life examples, I used PDB files containing the coordinates of both the native and activated structural states and set the occupancy of the two states to 0% and 100%, respectively. I then calculated data sets representing the native state $|F_A|$ and fully activated state $|F_B|$, respectively. Finally, I calculated $|F_{AB}|$ corresponding to various f by setting the occupancies of the activated state from 0.01 to 0.9 and

adjusting the occupancy of the native-state atoms correspondingly. Structure-factor amplitudes were calculated from these models using *SHELXL* (Sheldrick & Schneider, 1997).

From each of these sets of $|F_{AB}|$ values, I then extrapolated sets of structure-factor amplitudes ($|F_{B\mathrm{ex}}|$) for a 100% activated crystal according to (1). In addition to calculating the average error $\langle F_B - F_{B\mathrm{ex}}\rangle/\langle F_\Delta\rangle$ for these structural changes, I also calculated R_{ex}. R_{ex} is the equivalent of the standard crystallographic R factor where $F_{B\mathrm{ex}}$ and F_B take the place of F_{obs} and F_{calc}, respectively. R factors are the most common method to assess various errors in crystallography and R_{ex} therefore provides a convenient measure to compare the error introduced by SASFE with other errors encountered in crystallography.

3.1.1. Example 1: double-bond isomerization in the PYP photoreceptor, a very small structural change. The first example is the initial light-driven chromophore isomerization reaction (Genick *et al.*, 1998) of the photoreceptor protein PYP (Kyndt *et al.*, 2004). Only 60 of the protein's 1400 atoms move in this reaction and the average displacement of these 60 atoms is just 0.42 Å. This reaction is therefore a good example of a very subtle protein conformational change. Based on the discussion above, we expect SASFE to perform well for this example. For reflections with a d spacing of 2.0 Å and $f = 0.5$, the average relative SASFE error $\langle F_B - F_{B\mathrm{ex}}\rangle/\langle F_\Delta\rangle$ computed from the actual structure-factor-by-structure-factor extrapolation is 4.4%, which is higher than the estimate (2.6%) obtained according to (13) but substantially lower than the error (6.3%) expected for the random movement of the same number of atoms predicted by (11). I suspect that the slight underestimation of the error by (13) is the result of one heavier than average atom (*i.e.* the S atom linking the chromophore of PYP to the protein) dominating F_Δ and some correlation of atomic motion. Both factors will give the Argand diagram of F_Δ the character of a biased random walk and would increase the average F_Δ and thereby the average SASFE error. While (13) slightly underestimates the average relative SASFE error, it very nicely predicts the resolution-dependence of that error (Fig. 5).

More importantly, the average SASFE error is small, corresponding to less than 5% of the structure-factor amplitude (F_Δ) of the structural change. R_{ex}, shown as a function of f in Fig. 6, confirms this impression. Even with f as low as 0.05 R_{ex} is only 3% (*i.e.* comparable to the R_{merge} of a good macromolecular data set and small relative to a typical R_{cryst}). In other words, even when structure-factor extrapolation is performed from a data set in which only 5% of the molecules were activated, the error introduced by the scalar approximation remains small relative to typical errors in experimental data (*i.e.* R_{merge}).

For this particular example, the plot of R_{ex} *versus* f (Fig. 6*a*) is noticeably biphasic. For $f < 5\%$ R_{ex} increases sharply yet

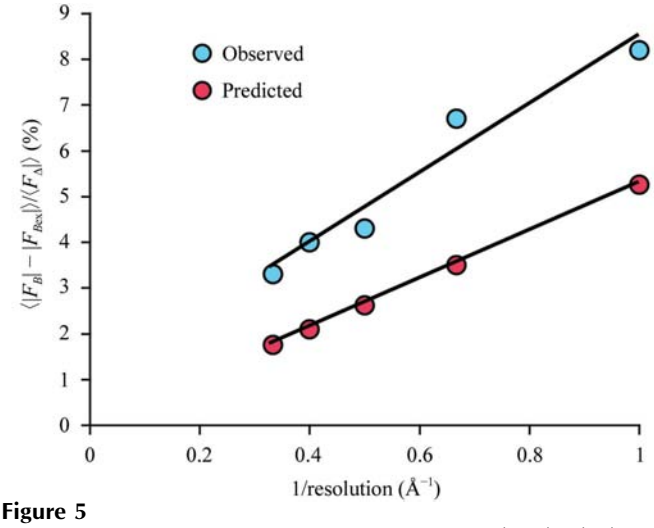

Figure 5
Resolution-dependence of the relative SASFE error $\langle|F_{B\mathrm{ex}}| - |F_B|\rangle/\langle F_\Delta\rangle$ estimated by (13) (red markers) and observed in an explicit 'atom-by-atom' calculation of the same protein conformational change. In this example, the structural change is the formation of PYP's early photocycle intermediate (I1) (see Fig. 4*a*) calculated for $f = 0.5$. The plot shows that (13) provides a reasonable approximation of the relative SASFE error (see text for discussion) and also captures the resolution-dependence of this error.

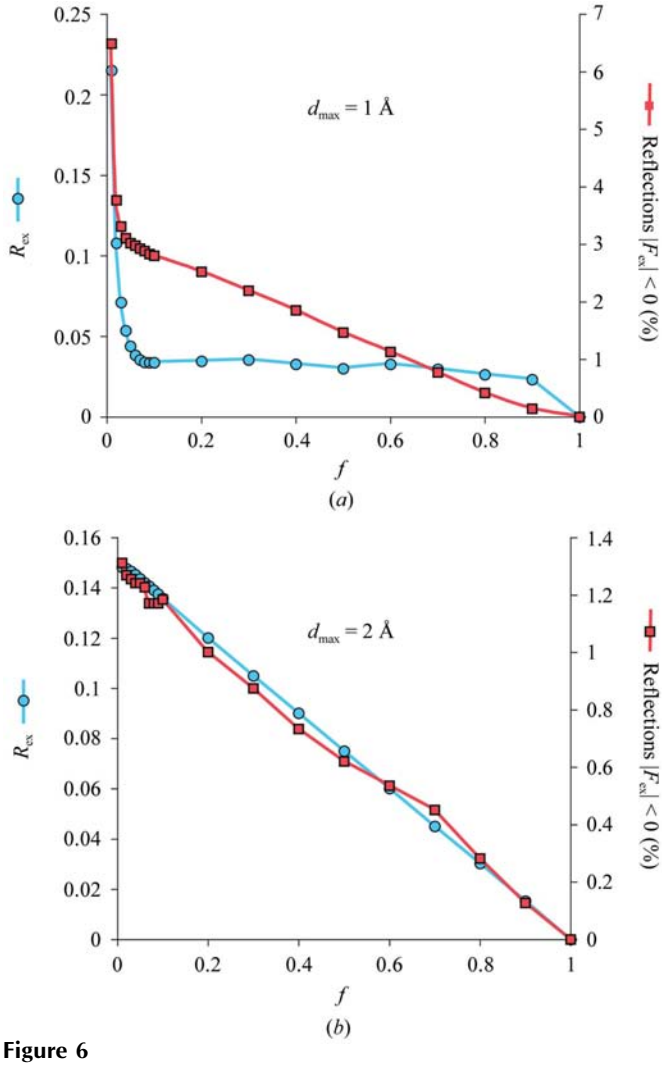

Figure 6
Extrapolation R factor R_{ex} (see text for definition) and fraction of $|F_{B\mathrm{ex}}| < 0$ plotted as a function of f for (*a*) a subtle structural change (dark–I1 conversion in PYP shown in Fig. 4*a*) and (*b*) for a more substantial structural change (dark–I2 conversion shown in Fig. 4*b*).

remains comparable to the $R_{\rm merge}$ of many crystallographic data sets. Surprisingly, for starting occupancies larger than 5% and smaller than 90% $R_{\rm ex}$ is nearly independent of the starting occupancy. Intuitively, one would expect that for increasing starting occupancies $|F_{B\rm ex}|$ should provide an increasingly better approximation of $|F_B|$. $R_{\rm ex}$ should therefore decrease for increasing starting occupancies and not stay constant as is observed in the plot. How can this result be understood?

(1) shows that $|F_{B\rm ex}|$ is not restrained to values larger than zero. For example, for $f = 0.5$ and $|F_A| > 2|F_{AB}|$, $|F_{B\rm ex}| < 0$. Fig. 2 further shows that $|F_{B\rm ex}|$ is always an underestimation of $|F_B|$. So, for cases where structure-factor extrapolation provides a particularly bad estimate of $|F_B|$, $|F_{B\rm ex}|$ is often smaller than zero. Since negative structure-factor amplitudes are physically impossible, I automatically reject $|F_{B\rm ex}|$ values smaller than 0 from the extrapolated data set. Effectively, the rejection of $|F_{B\rm ex}| < 0$ then acts as a filter that rejects $|F_{B\rm ex}|$ values that are particularly bad estimates of $|F_B|$. When plotting the fraction of negative $|F_{B\rm ex}|$ against f (Fig. 6*a*), one sees that for starting occupancies >5% the fraction of $|F_{B\rm ex}| < 0$ falls monotonically from 3% to 0. This in effect means that for smaller f more bad estimates of $|F_{B\rm ex}|$ are rejected, thus raising the average quality of the remaining $|F_{B\rm ex}|$. As f increases, the same $|F_{B\rm ex}|$ values will still provide a bad estimate of their corresponding $|F_B|$, but they may now be just above zero and are therefore retained in the data set and thus degrade the average quality of $|F_{B\rm ex}|$. Apparently, for higher f, the increasing inclusion of bad $|F_{B\rm ex}|$ values cancels the expected overall improvement in $R_{\rm ex}$. Thus, the improvement in the overall quality of the extrapolated data sets intuitively expected for higher values of f manifests itself in the data set's increasing completeness while $R_{\rm ex}$ remains constant.

Fig. 6(*a*) shows the $R_{\rm ex}$ calculated for all data extending to 1 Å resolution. As predicted in §2.2, the error introduced by the extrapolation should depend on the resolution. Specifically, the error introduced by the extrapolation should be greater for higher resolution structure factors. A plot of $R_{\rm ex}$ for the extrapolation with $f = 0.5$ *versus* the resolution (Fig. 7*a*) confirms this expectation. $R_{\rm ex}$ is almost twice as high for 1 Å resolution data than it is for 3 Å data, thus mirroring the results for the resolution-dependence of the average relative SASFE error shown in Fig. 5.

3.1.2. Example 2: chromophore rearrangement in the PYP active site, a modest structural change. After the initial light-activation reaction, PYP undergoes a second structural transformation. In this reaction, the chromophore and the protein active site undergo a more substantial rearrangement involving atomic displacements that exceed 5 Å. A total of 85 atoms move in this structural change and the average distance of their movement is 0.87 Å. The average relative SASFE error $\langle F_B - F_{B\rm ex}\rangle/\langle F_\Delta\rangle$ predicted by (12) for such a case of random atomic movements is 7.3%, which closely matches the observed value of 7.1%. While larger than the error observed in the case of the more modest structural change in the previous example, the overall error is still relatively modest. Fig. 7(*b*) shows that even for very high resolution structure factors and structural changes as substantial as these the SASFE error, as judged by $R_{\rm ex}$, remains small relative to the experimental uncertainties ($R_{\rm merge}$) encountered in all but the highest quality macromolecular data sets. The dependence of $R_{\rm ex}$ on f for this example (Fig. 6*b*) shows an essentially linear decrease in $R_{\rm ex}$ that closely resembles the theoretically predicted dependence of the average SASFE error on f (Fig. 3*a*). The difference in the dependence of $R_{\rm ex}$ on f for the two examples of structural changes shown here is rather striking, where the larger structural change resembles the theoretically expected behaviour much more closely. As discussed above, I suspect that the anomalous behaviour of the first example is the result of the structural change being dominated by the movement of a single S atom. In contrast, the second example involves sizable movements of a substantial number of atoms, so that the resulting changes in structure-factor amplitudes resemble a random structural change and therefore match the assumptions underlying (11)–(13) more closely.

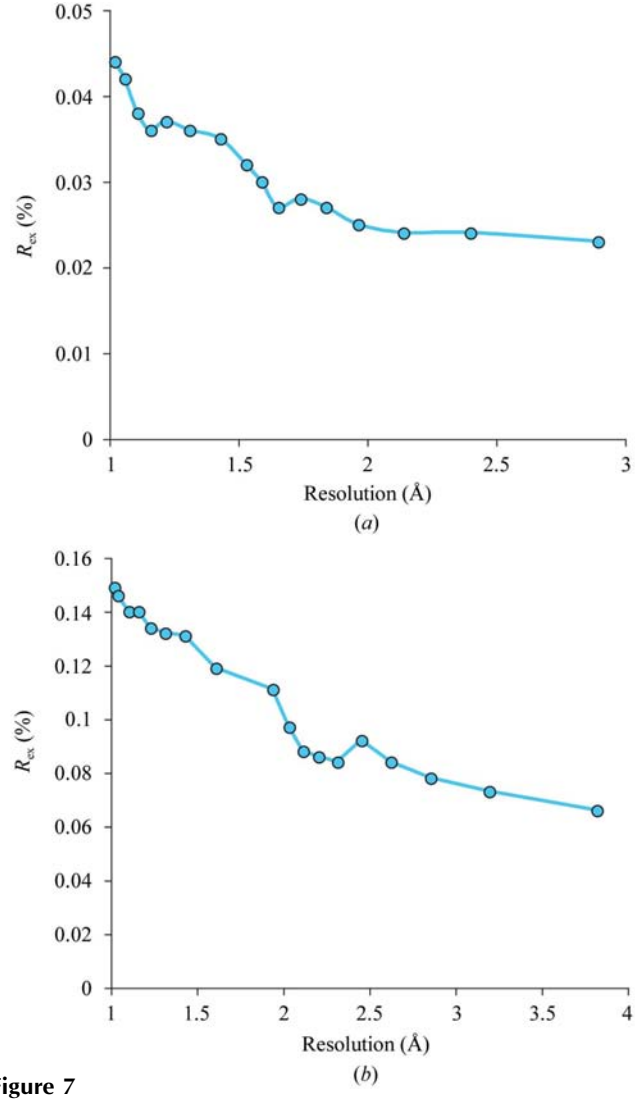

Figure 7
Extrapolation R factor $R_{\rm ex}$ as a function of resolution for (*a*) a subtle structural change (dark–I1 conversion in PYP shown in Fig. 4*a*) and (*b*) for a more substantial structural change (dark–I2 conversion shown in Fig. 4*b*).

3.2. Effect on electron-density maps

One of the principal uses of extrapolated data sets is in the calculation of electron-density maps to aid model building. The top row of electron-density maps in Fig. 8 shows the effect of the extrapolation error on the appearance of electron-density maps. The maps in these figures were calculated using structure-factor amplitudes extrapolated from calculated structure-factor amplitudes in the manner described above (§3.1). Phases were calculated from an atomic model from which the chromophore and the first layer of surrounding residues were removed. The resulting maps mirror the results of the R_{ex} calculation. For extrapolations from data sets in which the occupancy of the activated state is as small as 3% (*i.e.* $f = 0.03$) maps are virtually indistinguishable from the 100% occupancy map ($f = 1$). Only for the extrapolation from an initial occupancy of 1% is there any discernable deterioration in map quality.

4. Propagation of experimental errors decreases the effective resolution of extrapolated data sets

The first-principles calculations and the two real-world examples above show that SASFE works reliably for a broad range of resolution, occupancy and structural changes and that the error introduced by SASFE is small relative to the experimental uncertainties of structure-factor amplitudes in macromolecular crystallography experiments. However, so far all our calculations have assumed the absence of experimental error in the data. Since all extrapolation procedures have an inherent tendency to amplify experimental errors, it is important to analyze the effect of experimental uncertainties on the quality of the extrapolated structure factors. For the purposes of this analysis, I will treat the experimental error [$\sigma(F_A)$ and $\sigma(F_{AB})$] of the structure factors F_A and F_{AB} as independent of one another so that $\sigma(F_{Bex})$ can be computed using the basic error-propagation rule

$$\sigma_q = \left[\left(\sigma_x \frac{\partial q}{\partial x}\right)^2 + \left(\sigma_y \frac{\partial q}{\partial y}\right)^2\right]^{1/2}. \quad (14)$$

Applied to (1), this computes to

$$\sigma(F_{Bex}) = \left\{\left[\sigma(F_A)\left(1 - \frac{1}{f}\right)\right]^2 + \left[\sigma(F_{AB})\left(\frac{1}{f}\right)\right]^2\right\}^{1/2}. \quad (15)$$

(15) recapitulates the intuitively expected behaviour of $\sigma(F_{Bex})$. For large f, $\sigma(F_{Bex})$ approaches $\sigma(F_{AB})$ [in the extreme case of $f = 1$, $\sigma(F_{Bex}) = \sigma(F_{AB})$], while for small f the error is approximated by $1/f$ times the larger of the two structure-factor errors. Assuming $\sigma(F_A) = \sigma(F_B) = 1$, $\sigma(F_{Bex})$ computes to 2.2 for $f = 0.5$ and ~13.5 for $f = 0.1$. From this, it should be clear that F_{Bex} extrapolated from data sets in which the activated species is present only at low occupancy (*i.e.* small f) will quickly become dominated by noise. Also, $F/\sigma(F) = 2I/\sigma(I)$, so that an n-fold increase in $\sigma(F)$ will lead to an equivalent n-fold decrease in $I/\sigma(I)$. Since the average $I/\sigma(I)$ of crystallographic data sets falls off with resolution, an increase in $\sigma(F)$ and the proportional decrease in $I/\sigma(I)$ will cause the extrapolated data to reach an $I/\sigma(I)$ ratio of 2, at which crystallographic data sets are customarily truncated, at a lower resolution. In other words, both the average quality and the effective resolution of the extrapolated data set will be lower, maybe substantially lower, than those of the two starting data sets.

It turns out that towards the higher end of the resolution range of most macromolecular crystals $\sigma(I)$ is independent of resolution, so that the fall-off of $I/\sigma(I)$ depends largely on the fall-off of I with resolution. For the argument outlined below, one can assume that the shape of a plot of I *versus* resolution for a protein crystal with a given overall B factor will roughly resemble that of a C atom with the same B factor. Therefore, a rough estimate of the effective resolution of an extrapolated data set can be obtained from the value of f, the average B factor

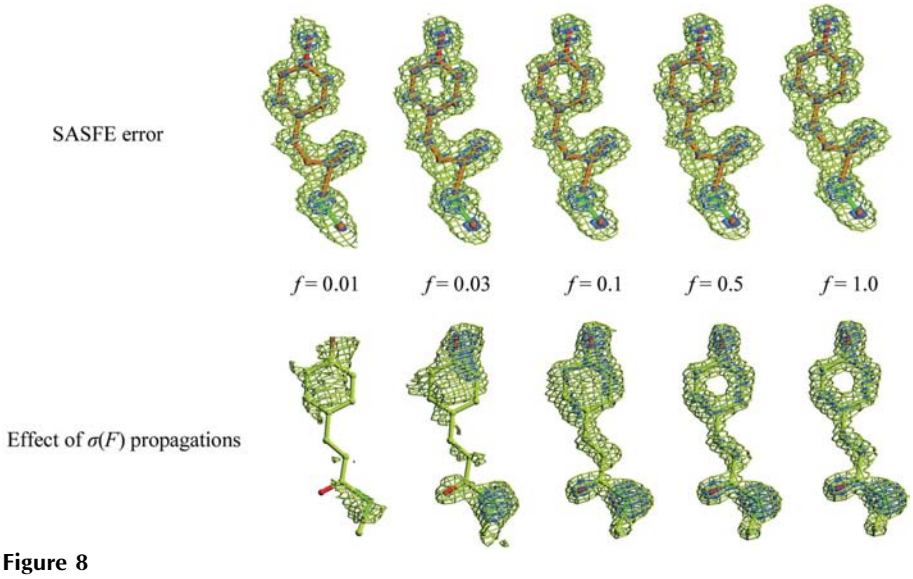

Figure 8
Factors limiting the usefulness of extrapolated maps. OMIT electron-density maps calculated from extrapolated structure factors (F_{Bex}) are shown as a function of f. The top row shows the effect of the error introduced by the SASFE procedure itself. The F_{Bex} values for the maps in this row were obtained to 1 Å resolution by SASFE using calculated structure factors for F_A and F_{AB} based on the atomic models shown in Fig. 4(*a*). The bottom row illustrates the effect of the inevitable amplification of experimental uncertainties during SASFE (see equation 15). The maps are 'mock extrapolations' (*i.e.* $F_A = F_{AB} = F_B = F_{Bex}$ except for experimental error) based on two experimental 1.2 Å resolution data sets from a fully dark-adapted PYP crystal. Therefore, changes in these extrapolated maps arise solely from the amplification of experimental error inherent in any extrapolation procedure. The error introduced by SASFE's scalar approximation causes virtually no deterioration in map quality. In comparison, the propagation of the experimental uncertainties leads to a rapid deterioration of the quality of the extrapolated maps and the apparent resolution.

and the effective resolution of the experimental data sets. For example, according to (15) for an extrapolation from $f = 0.3$, $\sigma(F_{Bex}) \simeq 4\sigma(F_A)$ [assuming $\sigma(F_A) \simeq \sigma(F_{AB})$]. If both the F_A and F_{AB} data sets extend to 2 Å [*i.e.* $I/\sigma(I) = 2$ at 2 Å] and the overall B factor is 20 Å^2, we can use the atomic scattering curve of carbon to predict that the fourfold increase in $\sigma(F)$ caused by the extrapolation procedure decreases the effective resolution of the extrapolated data set [*i.e.* the resolution where $I/\sigma(I) > 2$] to ∼2.6 Å.

The above discussion is presented to provide an order-of-magnitude estimate of the extrapolation's effect on resolution for the purpose of experimental planning or the evaluation of extrapolations where the experimental data are not accessible.

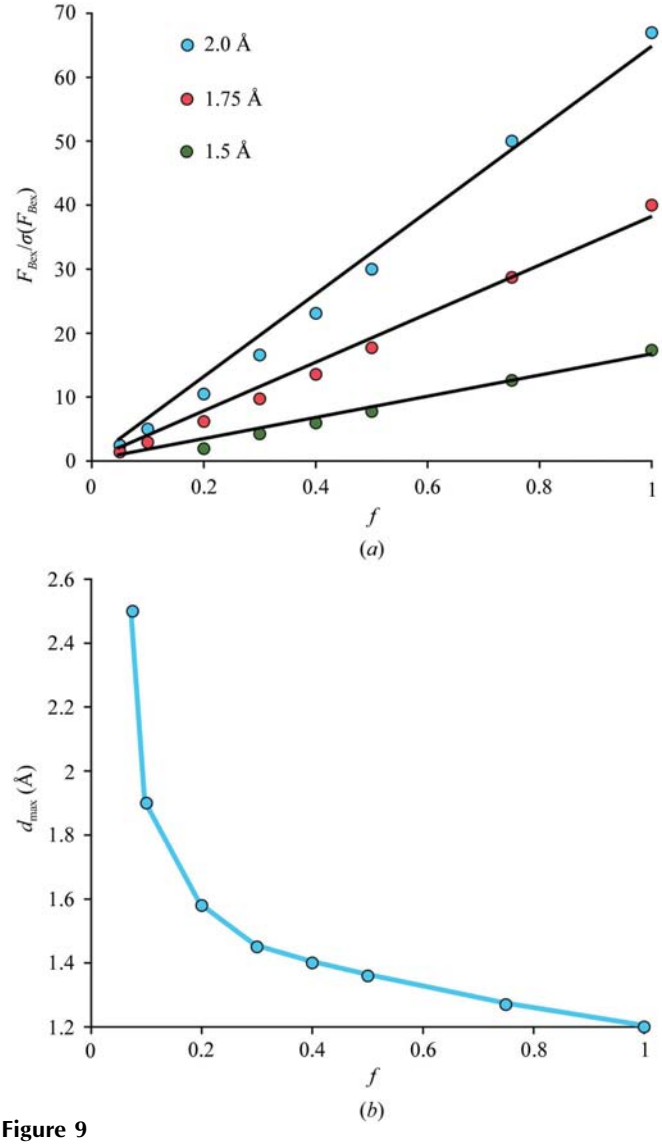

Figure 9
Effect of the propagation of experimental uncertainties on the quality of extrapolated data sets. (*a*) shows the $F/\sigma(F)$ ratio for reflections at three resolutions as a function of f. (*b*) shows the effect of f on the effective resolution of the extrapolated data sets [*i.e.* the resolution below which $F/\sigma(F) > 4$]. The graphs are based on a 'mock extrapolation' of a high-quality 1.2 Å resolution data set [see Fig. 8 for an explanation of the 'mock-extrapolation' procedures and an OMIT map calculated from the original data ($f = 1$)].

Given actual F_A and F_{AB} data sets, the effective resolution of the extrapolated data set would be determined by performing the error-propagation calculation on a structure-factor-by-structure-factor basis as implemented in the program *XTRA* (see §6) while monitoring the $I/\sigma(I)$ ratio of reflections by resolution shell.

In general, it would be useful to truncate extrapolated maps to the resolution at which the $I/\sigma(I)$ ratio of the extrapolated data set falls below 2. Truncation of the extrapolated data set at this new effective resolution will both reduce noise from high-resolution structure factors dominated by error and provide a realistic judgment of whether features observed in a map are significant or dominated by noise.

Comparison of electron-density maps calculated with real-life extrapolated data (not shown) with those extrapolated from calculated F values (Fig. 8, top row) strongly support the notion of the above analysis and indicate that the practical limitation for the usefulness of SASFE is not the limitation of SASFE itself, but the amplification of experimental errors.

To test the effect of experimental uncertainties of structure-factor amplitudes on electron-density maps independent of the error introduced by the SASFE procedure, I performed a series of 'mock extrapolations'. These extrapolations were performed for various values of f using a high-quality 1.2 Å resolution experimental data set for the dark-adapted state of PYP as the starting point. To generate the two starting data sets (*i.e.* F_A and F_{AB}), I took the experimentally observed data set and calculated a 'sigma-shuffled' second data set. Briefly, for each reflection in this sigma-shuffled data set $F_{\rm shuffled} = F_{\rm obs} + [\sigma(F) \times r]$, where r is a Gaussian-deviate random number. Assuming that the $\sigma(F)$ values represent reasonable estimates of the true experimental uncertainties of their corresponding $F_{\rm obs}$, the shuffled data set represents a statistically probable data set one might have observed had one measured a second data set on the same crystal. In the mock extrapolation I then used $F_{\rm obs}$ and $F_{\rm shuffled}$ as F_A and F_{AB} so that (within experimental error) $F_A = F_{AB} = F_{Bex}$. Therefore, maps resulting from this mock-extrapolated data (Fig. 8, bottom row) reflect the effect of the amplification of experimental uncertainties independent of the SASFE error.

Comparison of the electron-density maps in Fig. 8 indicates that the deleterious effect of the propagation of experimental uncertainties dominates the error introduced by the SASFE procedure itself. The dependence of the overall quality of these mock-extrapolated maps closely resembles the trend my laboratory has observed for many extrapolated maps. Fig. 9 shows how the $F/\sigma(F)$ ratio in these mock-extrapolated data sets varies with f. Fig. 9(*a*) indicates that the average $F/\sigma(F)$ ratio across a range of resolutions is roughly linearly proportional to f. Fig. 9(*b*) shows how this decrease in $F/\sigma(F)$ translates into the effective resolution of the data set, *i.e.* the resolution, below which $F/\sigma(F) > 4$ [*i.e.* $I/\sigma(I) > 2$]. The results of this analysis indicate that maps extrapolated from data sets in which the activated molecular species is present at less then 15–20% are likely to be significantly compromised and will only be useful in the rare case of truly exceptional starting data.

5. Other sources of error in extrapolated electron-density maps

In addition to the errors specific to the scalar approximation of a vector extrapolation, which are discussed above, extrapolated electron-density maps are also subject to general sources of error and bias. One such source of error is phase bias towards the structure of the unperturbed state of the molecule. While this error is very small for the two examples treated here and is likely to remain small for most applications of SASFE, cases may exist where this error may become significant. Such cases may occur when protein structural changes are very large and involve a very large fraction of a structure's atoms or when the structural model for the unperturbed structure is very incomplete or of low quality. For such cases, it may be important to choose appropriate procedures to reduce model bias. The article by Read (1997) provides an excellent and authoritative overview of the problem of phase bias and techniques for mitigating this bias.

Throughout this work, I have assumed that f is known exactly, but obviously knowledge of f is also subject to error. Basic error-propagation rules applied to (1) show that $\sigma(F_{B\mathrm{ex}})$ depends on $\sigma(f)$ according to

$$\sigma(F_{B\mathrm{ex}}) = \sigma(f)\big||F_{AB}| - |F_A|\big|/f^2$$

and with $|F_\Delta| = \big||F_{AB}| - |F_A|\big|/f$ we then obtain

$$\frac{\sigma(F_{B\mathrm{ex}})}{|F_\Delta|} = \frac{\sigma(f)}{f}. \qquad (16)$$

So, for example, in the case of $f = 0.5 \pm 0.05$, the resulting relative uncertainty in $F_{B\mathrm{ex}}$ [*i.e.* $\sigma(F_{B\mathrm{ex}})/F_\Delta$] is a tolerable 10%. Electron-density maps extrapolated with purposefully incorrect values of f support this analysis. Maps (not shown) were extrapolated from simulated data sets representing 40/60, 50/50 or 60/40 ratios of native and activated states of PYP using f values ranging from 0.4 to 0.6. These maps look very similar and support identical structural interpretations. Also, even excessive mis-estimates of f do not simply lead to noisy or difficult to interpret maps. Instead, an overestimation of f will simply lead to a map in which state B appears to have elevated B factors. In the case of a drastic overestimation, state A remains noticeably in the extrapolated map. Correspondingly, an underestimation of f will lead to an extrapolated map in which the activated state is overemphasized, giving the appearance of very low B factors for the activated species. By inspecting relative peak heights for individual atoms or side chains and iteratively adjusting f until the heights of these electron-density features are the same in the unperturbed and extrapolated maps, it is actually possible to obtain rather precise estimates of the occupancy of state B in the partially activated crystal.

6. *XTRA*: a software implementation of SASFE

Structure-factor extrapolation according to (1) has been implemented in the program *XTRA*. This program takes two structure-factor files in `SHELX.hkl` format (*i.e.* h k l F sigF R-free_flag) and generates an extrapolated data set containing the extrapolated structure-factor amplitudes. The $\sigma(F_{B\mathrm{ex}})$ are calculated according to (15). The program rejects structure factors with amplitudes smaller than zero and displays the number of reflections rejected by this criterion at the end of the output. *XTRA* also generates a table of $F/\sigma(F)$ by resolution shell for the two input data sets and the extrapolated data set.

When running *XTRA*, it is essential that the two input data sets (F_A and F_{AB}) have previously been scaled to one another and that the scaling routine has corrected differences in the overall B factor between the two data sets. Failure to do so will make the extrapolation results meaningless. Further, estimation of the extrapolated data set relies on the correct estimation of $\sigma(F)$ during data merging. Finally, since the extrapolation procedure leads to a consistent underestimation of $|F_{B\mathrm{ex}}|$ (see §2), the SASFE procedure slightly alters the overall scale of the data set. Therefore, if subsequent computations assume the extrapolated data to be on an absolute scale, it will be necessary to rescale the extrapolated data set, even when both input data sets were initially on the absolute scale.

The program is available from the author as a precompiled standalone command-line application for Linux on X86 architecture or as source code in C.

7. Conclusions

SASFE, the scalar approximation to structure-factor extrapolation, is commonly used in crystallographic experiments aimed at determining the response of a protein structure to a physical or chemical stimulus. In these experiments, SASFE allows the estimation of the structure-factor amplitudes for a crystal in the fully activated state from experimentally observed structure-factor amplitudes of a non-activated and a partially activated crystal. Both first-principles-based analysis of the error introduced by SASFE and the simulation of SASFE methods for two examples of actual in-crystal activation experiments show that SASFE is surprisingly reliable and robust. As a result, the usefulness of SASFE is not limited by SASFE-specific errors, but by the amplification of experimental errors inherent to all extrapolation procedures. This amplification of experimental errors effectively reduces the resolution of the extrapolated data set relative to that of the two experimentally observed data sets.

The author thanks Pierre Damien Coureux for collecting experimental data and Bruce Foxman, Nikolaus Grigorieff and Jane Kondev for helpful discussions.

References

Cao, Y., Musah, R. A., Wilcox, S. K., Goodin, D. B. & McRee, D. E. (1998). *Protein Sci.* **7**, 72–78.

Edman, K., Nollert, P., Royant, A., Belrhali, H., Pebay-Peyroula, E., Hajdu, J., Neutze, R. & Landau, E. M. (1999). *Nature (London)*, **401**, 822–826.

Edman, K., Royant, A., Nollert, P., Maxwell, C. A., Pebay-Peyroula, E., Navarro, J., Neutze, R. & Landau, E. M. (2002). *Structure*, **10**, 473–482.
Genick, U. K., Borgstahl, G. E., Ng, K., Ren, Z., Pradervand, C., Burke, P. M., Srajer, V., Teng, T.-Y., Schildkamp, W., McRee, D. E., Moffat, K. & Getzoff, E. D. (1997). *Science*, **275**, 1471–1475.
Genick, U. K., Soltis, S. M., Kuhn, P., Canestrelli, I. L. & Getzoff, E. D. (1998). *Nature (London)*, **392**, 206–209.
Giacovazzo, C., Monaco, H. L., Artioli, G., Viterbo, D., Ferraris, G., Gilli, G., Zanotti, G. & Catti, M. (2002). *Fundamentals of Crystallography*. Oxford University Press.
Kyndt, J. A., Meyer, T. E. & Cusanovich, M. A. (2004). *Photochem. Photobiol. Sci.* **3**, 519–530.
Lanyi, J. K. & Schobert, B. (2007). *J. Mol. Biol.* **365**, 1379–1392.
Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.
Read, R. J. (1997). *Methods Enzymol.* **277**, 110–128.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.