Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Substructure search procedures for macromolecular structures

R. W. Grosse-Kunstleve* and P. D. Adams

Lawrence Berkeley National Laboratory, One Cyclotron Road, BLDG 4R0230, Berkeley, California 94720-8235, USA

Correspondence e-mail: rwgrosse-kunstleve@lbl.gov

Received 28 April 2003
Accepted 12 August 2003

This paper accompanies a lecture given at the 2003 CCP4 Study Weekend on experimental phasing. The first part is an overview of the fundamentals of Patterson methods and direct methods with the audience of the CCP4 Study Weekend in mind. In the second part, a new hybrid substructure search is outlined.

1. Introduction

Traditionally, experimental phasing of macromolecular structures involves heavy-atom soaks and the collection of two or more data sets: the diffraction intensities of the native crystal and those of the derivative(s). This is often referred to as single or multiple isomorphous replacement (SIR, MIR). In recent years, it has become very popular to use crystals containing anomalous scatterers, most notably by selenomethionine substitution. These experiments are known as single or multiple anomalous diffraction experiments (SAD, MAD) or alternatively single anomalous scattering experiments (SAS).

Experimental phasing can be viewed as a divide-and-conquer technique in which the larger problem of determining the complete structure is divided into two steps.

(i) Given the experimental diffraction data, approximate substructure structure factors are computed, *e.g.* difference structure factors. The substructure is solved using methods developed for the solution of small molecules.

(ii) Using the substructure, algebraic or probabilistic methods are used to extrapolate phases for the full structure. The structure-solution process continues with density modification, model building and refinement. In this paper, we focus on the first step above, the determination of the substructure.

2. Estimation of substructure structure factors

2.1. Isomorphous differences

Since the number of atoms in a native macromolecular structure is usually much larger than the number of additional heavy atoms in a derivative, it is a valid approximation to assume $F_H << F_{PH}$, where F_H are the structure-factor amplitudes corresponding to the substructure only and F_{PH} the structure-factor amplitudes of the derivative. This approximation leads to (Blundell & Johnson, 1976*a*)

$$F_H << F_{PH} \Rightarrow F_{PH} - F_P \approx F_H \cos(\varphi_{PH} - \varphi_H). \qquad (1)$$

The cosine term takes on values between −1 and 1. Therefore, the isomorphous differences $F_{PH} - F_P$ are lower estimates of

© 2003 International Union of Crystallography
Printed in Denmark – all rights reserved

the substructure structure factor F_H: the F_H can be larger but they cannot be smaller than the observed isomorphous differences.

2.2. Anomalous differences

Similar considerations lead to the following equation for anomalous differences $F^+_{PH} - F^-_{PH}$ (Blundell & Johnson, 1976*b*),

$$F''_H << F'_{PH'} \Rightarrow F^+_{PH} - F^-_{PH} \approx 2F''_H \sin(\varphi_{PH} - \varphi_H). \quad (2)$$

Here, F''_H are the imaginary contributions to the structure factors of the anomalous scatterers and F'_{PH} is the sum of the structure factors of the macromolecular structure and the real contributions of the anomalous scatterers. The sine term also takes on values between -1 and 1. Therefore, the anomalous differences are lower estimates of the imaginary contributions of the anomalous scatterers.

2.3. F_A structure factors

In the case of multiple anomalous diffraction (MAD) experiments, it is possible to compute better estimates of the substructure factors. These estimates are commonly referred to as F_A structure factors. Various algorithms for the computations of F_A structures are available: *MADSYS* (Hendrickson, 1991), *CCP*4 *REVISE* (Fan *et al.*, 1993), *SOLVE* (Terwilliger, 1994) and *XPREP* (Bruker AXS, Madison, USA). For good MAD data, F_A structure factors usually lead to significantly more efficient determination of the substructure. However, if the MAD data are affected by systematic errors such as intensity changes arising from radiation damage, it is possible that the corresponding F_A structure factors are not suitable for substructure determination. In this case, it is adventageous to attempt substructure determination with the data set collected first (ideally at the peak of the anomalous signal).

3. The phase problem

In the second and third decades of the 20th century, early X-ray crystallographers worked out that the observed diffraction intensities are directly related to the Fourier transformation of the electron density of the crystal structure (not taking Lorentz factors, polarization factors and other experiment-specific corrections into account),

$$I^{1/2} \equiv |F| \propto \mathrm{FT}(\rho). \quad (3)$$

Here, I represents the observed intensities, $|F|$ the structure-factor amplitudes, ρ the electron density and FT a Fourier transformation. The same relation more specifically:

$$F_h = |F_h| \exp(i\varphi_h) = \frac{V}{N} \sum_x \rho(x) \exp(2\pi i h x). \quad (4)$$

Here, h is a Miller index, x the coordinate of a grid point in real space, N the total number of grid points and V the volume of the unit cell. The complex structure factor F is also shown in the alternative representation as a pair of amplitude $|F|$ and phase (φ).

Obviously, it is straightforward to compute the structure-factor amplitudes from the electron density. Given complex structure factors, it is equally straightforward to compute the electron density *via* a Fourier transformation,

$$\rho \propto \mathrm{FT}^{-1}(F). \quad (5)$$

FT^{-1} represents the inverse Fourier transformation. Unfortunately, with current technology it is almost always impractical to directly measure both intensities and phases. Conventional diffraction experiments only produce intensities; the phases are not available. This is colloquially known as the 'phase problem' of crystallography.

4. Techniques for solving the phase problem

4.1. Patterson methods

The Patterson function is defined as the Fourier transformation of the observed intensities,

$$\text{Patterson} \propto \mathrm{FT}^{-1}(I). \quad (6)$$

This is a straightforward calculation requiring only the experimental observations as the input. Patterson (1935) showed that the peaks in this Fourier synthesis correspond to *vectors between atoms* in the crystal structure. Alternatively, the Patterson function can be viewed as a convolution as follows.

(i) Note that the real intensity I is the product of the complex structure factor F and its complex conjugate. Therefore,

$$\text{Patterson} \propto \mathrm{FT}^{-1}(I) = \mathrm{FT}^{-1}(F \cdot F^*). \quad (7)$$

(ii) The next elementary observation is

$$F^* = \mathrm{FT}(\rho_{\text{inverse}}). \quad (8)$$

This follows immediately from the definition of the discrete Fourier transformation (4).

(iii) Now we consider a central theorem of Fourier methods, the convolution theorem (*e.g.* Giacovazzo, 1992),

$$\mathrm{FT}(g) \cdot \mathrm{FT}(h) = \mathrm{FT}[\text{Convolution}(g, h)]. \quad (9)$$

(iv) By substituting ρ and ρ_{inverse}, we arrive at

$$F \cdot F^* = \mathrm{FT}(\rho) \cdot \mathrm{FT}(\rho_{\text{inverse}}) = \mathrm{FT}[\text{Convolution}(\rho, \rho_{\text{inverse}})]. \quad (10)$$

(v) Comparison with (7) leads to the conclusion (Ramachandran & Srinivasan, 1970)

$$\text{Patterson} = \text{Convolution}(\rho, \rho_{\text{inverse}}). \quad (11)$$

4.2. Patterson interpretation in direct space and in reciprocal space

In the classic textbook *Vector Space*, Buerger (1959) demonstrates that under idealized conditions image-seeking procedures are capable of recovering the image of the electron density from the Patterson function. 'Idealized conditions' essentially means fully resolved peaks in the Patterson function. In practice this condition is only fulfilled for very small structures, but it still is possible to extract useful information from real Patterson maps. The basic idea is as follows.

(i) Postulate a hypothesis, for example a putative substructure configuration.

(ii) Test the hypothesis against the Patterson map.

The test involves the computation of vectors between the atoms of the putative substructure and the determination of the values in the Patterson map at the location of these vectors. This involves interpolation between grid points of the map. The interpolated peak heights are usually the input for the computation of a Patterson score. Theoretically, the minimum of all the peak heights found is the most powerful measure, but sum or product functions have also been used (Buerger, 1959). Nordman (1966) suggests using the mean of a certain percentage of the lowest values.

It is also possible to work with the observed intensities in reciprocal space, without transforming them according to (6). Conceptually, the procedure is even simpler.

(i) Postulate a hypothesis, for example a putative substructure configuration.

(ii) Test the hypothesis against the observed intensities.

In this case, the test involves the calculation of intensities for the putative structure and the evaluation of a function comparing these with the observed intensities; for example, the standard linear correlation coefficient (*e.g.* Press *et al.*, 1986). An advantage of this method is that it does not involve interpolations and should therefore be intrinsically more accurate. However, the calculations are much slower than the computation of Patterson scores in direct space if performed in the straightforward fashion suggested here. The key to making the reciprocal-space approach feasible is the fast translation function devised by Navaza & Vernoslova (1995). We were able to show that the fast translation function is typically 200–500 times faster than the conventional translation function. The fast translation function was originally designed for solving molecular-replacement problems, but we have also used it successfully for the determination of substructures (Grosse-Kunstleve & Brunger, 1999).

4.3. Difference Fourier methods

The popular *SOLVE* program (Terwilliger & Berendzen, 1999) tightly integrates Patterson methods, difference Fourier analysis and phasing. One or two initial substructure sites are determined with Patterson superposition functions. The remaining sites are found by repeated analysis of isomorphous or anomalous difference Fourier maps. These fundamental building blocks are integrated into a high-level procedure that automates decision making using a sophisticated scoring system. *SOLVE* includes all steps including the refinement of experimental phases. A full account of the procedure is beyond the scope of this paper and the reader is referred to the original publication.

4.4. Direct methods

Direct methods were originally developed for the *direct determination of phases* without direct use of stereochemical knowledge. The fundamental approach is to start with a very small set of starting phases and to construct a more complete phase set by applying phase probability relationships. The expanded phase set in combination with the observed structure factors is used to compute an electron-density map that is hopefully interpretable when stereochemical knowledge is taken into account.

The phase probability relations governing the phase-extension procedure are usually based on the well known tangent formula (Karle & Hauptman, 1956). This formula is typically introduced as

$$\tan(\varphi_h) = \frac{\sum_k \left|E_k E_{h-k}\right| \cos(\varphi_k + \varphi_{h-k})}{\sum_k \left|E_k E_{h-k}\right| \sin(\varphi_k + \varphi_{h-k})}. \qquad (12)$$

To avoid distraction, for the moment we will assume that the E values in this formula are analogous the structure factors F introduced above. The derivation of the tangent formula employs the assumptions that the electron density is positive everywhere in the unit cell (positivity) and that all atoms are resolved (atomicity). To understand this, it is useful to rewrite the tangent formula as a simpler but mathematically equivalent expression,

$$E_h \propto \sum_k E_k E_{h-k}. \qquad (13)$$

Comparison with the definition of the convolution (*e.g.* Giacovazzo, 1992) leads us to recognize that

$$\sum_k E_k E_{h-k} \equiv \text{Convolution}(E, E). \qquad (14)$$

Application of the convolution theorem (9) leads to

$$\sum_k E_k E_{h-k} = \text{FT}[\text{FT}^{-1}(E) \cdot \text{FT}^{-1}(E)]. \qquad (15)$$

Application of (5) leads to

$$\text{FT}^{-1}(E) \cdot \text{FT}^{-1}(E) = \rho^2. \qquad (16)$$

Thus, we arrive at

$$E_h \propto \sum_k E_k E_{h-k} = \text{FT}[\text{FT}^{-1}(E)^2] = \text{FT}(\rho^2). \qquad (17)$$

This equation shows that the tangent formula uses positivity and atomicity to introduce a self-consistency argument. Fig. 1 illustrates the essence of direct methods.

(i) Consider a crystal structure of positive point atoms of equal weight (electron density ρ).

(ii) From (4) we know that the Fourier transformation yields complex structure factors E.

(iii) Now consider the square of the crystal structure of point atoms (ρ^2).

(iv) The tangent formula postulates that the Fourier transform of ρ^2 yields structure factors that are directly proportional to the structure factors obtained by transforming ρ. The amplitudes may differ by a constant factor depending on the weight chosen for the point atoms, but the phases are identical.

This argument is essentially the same as that used in the derivation of the Sayre equation (Sayre, 1952; note that the title of this paper begins with 'The Squaring Method'). Sayre's equation is slightly more complex than the tangent formula because it is formulated for atoms with Gaussian shapes rather than point atoms. This describes real crystal structures more closely, but in practice it is often more advantageous to eliminate the shape term and to work with normalized structure factors corresponding to point atoms. Recognizing that the Fourier transformation of an isolated point atom is a constant, it is only a small step to realise that the expected average diffraction intensities of a point-atom structure are independent of the diffraction angle (*cf.* neutron diffraction experiments). Therefore, normalized structure factors can be estimated from observed intensities by enforcing the expected average in resolution shells,

$$E_h^2 = \frac{F_h^2}{\langle F_h^2/\varepsilon_h \rangle}. \qquad (18)$$

To be precise, this equation yields estimates of the quasi-normalized structure factors. The term ε takes the multiplicity of the reflections into account and can be directly computed from the space-group symmetry (simply by counting how often a given Miller index h is mapped onto itself by symmetry; ε must be used instead of the more familiar multiplicity because the latter conventionally takes Friedel symmetry into account).

4.5. Convolutions revisited

We have shown that both the Patterson function and the tangent formula underlying direct methods can be interpreted as convolutions. To summarize,

$$\text{Patterson} = \text{Convolution}(\rho, \rho_{\text{inverse}}).$$

The Patterson function is a *convolution in direct space* that leads to *squaring in reciprocal space*: the intensities are proportional to the square of the structure factors (3). Conventionally, the Patterson function is analyzed in direct space using *image-seeking* procedures,

$$E_h \propto \sum_k E_k E_{h-k} = \text{Convolution}(E, E) = \text{FT}(\rho^2).$$

The tangent formula is a *convolution in reciprocal space* that leads to *squaring in direct space*. Employing positivity and atomicity, the tangent formula leads to a *self-consistency argument* and in practice some form of *recycling* (see Fig. 1).

4.6. Dual-space structure-solution methods

The tangent formula alone often does not work efficiently for solving structures with many atoms (see Woolfson, 1961, for some very interesting remarks). The most popular 'direct-methods' programs used in macromolecular crystallography today are the result of an evolution that transformed the pure phase-extension idea into complex multi-trial search procedures. *MULTAN* (Germain *et al.*, 1970) pioneered the multi-trial approach but is still motivated by the phase-extension idea. *RANTAN* (Yao, 1981) and early versions of *SHELX* (Sheldrick, 1985) mark the transition to random-seeded multi-trial approaches that use the tangent formula in a recycling procedure to enforce self-consistency (Fig. 1). *Shake-and-Bake* (Miller *et al.*, 1994) and more recent versions of *SHELX* (Sheldrick & Gould, 1995) introduced the concept of dual-space recycling (Sheldrick *et al.*, 2001). Reciprocal-space phase manipulation based on the tangent formula, or the minimal function in the case of *Shake-and-Bake*, is alternated with direct-space interpretation of Fourier maps. *Shake-and-Bake* picks peaks from the Fourier maps, taking a given minimum distance into account. The peaks are used in a structure-factor calculation to obtain new phases that are entered into the next cycle of phase manipulation. *SHELXD* (Schneider & Sheldrick, 2002) follows a similar approach but typically picks more peaks than are expected (*e.g.* 1.3 times the expected number of sites) and randomly selects the expected number for recycling. This is known as the random omit procedure.

4.7. Direct-methods recycling with Patterson seeding

Conventional direct-methods programs initialize the recycling procedure with random phases or random coordinates. In contrast, *SHELXD* (Schneider & Sheldrick, 2002) uses Patterson seeding to obtain better than random starting phases for the recycling procedure. The fundamental steps in the procedure are the following.

(i) Generation of two-atom fragments. A given number of peaks are picked from a sharpened Patterson map (special positions are omitted). These are considered to be possible vectors between two atoms of the substructure. However, at

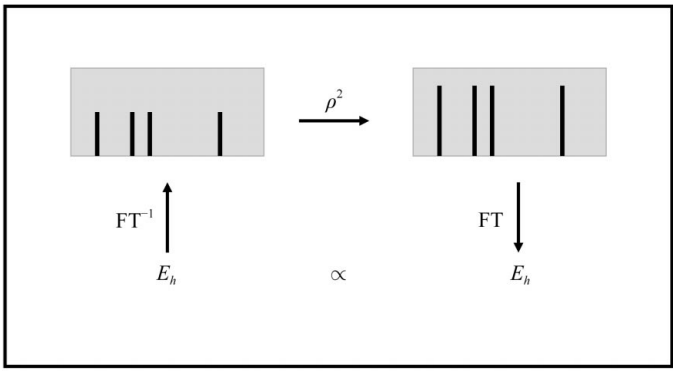

Figure 1
The essence of direct methods. Normalized structure factors correspond to point atoms at rest. Squaring in direct space followed by a Fourier transformation leads to structure factors that are proportional to the original structure factors. The phases are identical.

this stage only the relative orientation of the two atoms is known, not their absolute position in the unit cell. The Nordman (1966) function is used to obtain scores for a number of random translations of the two-atom fragments.

(ii) Extrapolation to the full substructure. Conceptually, a third probe atom is systematically placed on the points of a uniform grid over the asymmetric unit while keeping a trial two-atom fragment fixed at a position that led to a high score. For each grid point, the resulting interatomic vectors are computed, followed by the determination of the corresponding Nordman score. Points with the highest scores are added to the original two-atom fragment to generate the expected number of atoms.

(iii) Correction of defects. Typically, the structures obtained in the previous step contain a considerable number of misplaced atoms. Even the best solutions often have less than half of the atoms correctly placed. These defects are efficiently corrected using dual-space recycling (tangent-formula expansions and random omission of peaks). The standard linear correlation coefficient (*e.g.* Press *et al.*, 1986) between calculated and observed intensities is a very reliable score for ranking the final results of the dual-space recycling procedure.

5. Rapid prototyping of a hybrid substructure search

We have implemented a prototype for a new hybrid substructure search procedure (*HySS*) based on Patterson methods and direct methods as described above, similar to those developed in programs such as *Shake-and-Bake* and *SHELXD*. We used the algorithms already implemented in the Computational Crystallography Toolbox (Grosse-Kunstleve *et al.*, 2002; Grosse-Kunstleve & Adams, 2003*a*) as fundamental building blocks. This included a limited-memory (Langs, 2002) fast translation function (Navaza & Vernoslova, 1995) that we had already implemented for the solution of molecular-replacement problems (Adams *et al.*, 2002). Efficient algorithms for the handling of symmetry, fast Fourier transformations and structure-factor calculations were also readily available. Our main goals were the following.

(i) To test the usefulness of the theoretically more accurate fast translation function for Patterson seeding.

(ii) To replace the random search for two-atom fragment positions with a systematic search.

(iii) To find reliable methods for automatically terminating the search procedure when it is clear that the substructure is solved.

(iv) To minimize the amount of newly written compiled code (C++) to reduce the development time (Grosse-Kunstleve & Adams, 2003*a*).

The following is the core procedure as implemented at the moment, entirely in a high-level interpreted language (Python).

(i) Generation of two-atom fragments. A given number of peaks are picked from the Patterson map computed with quasi-normalized intensities as coefficients (peaks on Harker sections are omitted). For each Patterson vector a two-atom fragment is constructed with the atoms at the endpoints. The limited-memory fast translation function is used to systematically sample the entire asymmetric unit for the best positions.

(ii) Extrapolation to the full substructure. For a given number of peaks in the two-atom translation function, the positioned two-atom fragment is kept fixed in the computation of another fast translation function with a third atom as the probe. The peaks in this function are added to the two-atom fragment to obtain the expected number of substructure sites.

(iii) Correction of defects. Defects in the extrapolated structures are corrected using a direct-space recycling procedure. Because it was faster to implement in our framework, the tangent-formula expansions are performed in direct space simply by squaring, exactly as shown in Fig. 1. This is alternated with application of the random omit procedure. We search for 0.9 times the number of expected sites and randomly select 2/3 for the next recycling step.

HySS introduces the following new experimental features.

(i) Initial recycling in *P*1 symmetry. A somewhat counter-intuitive but consistent observation is that the recycling procedure is often far more efficient if carried out in *P*1 symmetry (Sheldrick & Gould, 1995). Therefore, we expand the extrapolated structures obtained *via* the fast translation function with the third site to *P*1 symmetry. After a given number of recycling cycles (default 10), the fast translation function is used a third time with the entire *P*1 structure as the probe in order to relocate the solution in the original symmetry. At the correct positions with respect to the symmetry elements the *P*1 structure is mapped onto itself, resulting in high correlations. At other positions the atoms are mapped essentially onto random positions for which one can expect low correlations. Therefore, the peaks in the translation function indicate the correct origin of the *P*1 structure with respect to the symmetry elements of the actual space group.

(ii) Recycling only for extrapolated structures with high correlation coefficients. For the most frequently found macromolecular space groups, the recycling procedure for the correction of defects is the most time-consuming step. To save time, we monitor the correlation coefficients of the extrapolated structures and start the recycling procedure only if the correlation coefficient is among the ten highest encountered so far.

(iii) Automatic termination. Conventionally, search procedures are run for a preset number of trials or until they are terminated manually. The correlation coefficients are usually the guide for decision-making. However, using the correlation coefficients alone is sometimes difficult. The absolute values are not necessarily conclusive, especially if the correct solution has a low correlation. Some searches yield tri-modal distributions, so that simply looking for outstanding correlation coefficients can also be misleading. Therefore, the *Shake-and-Bake* suite (Smith, 2002) and recently the *SHELX* suite (Dall'Antonia *et al.*, 2003) include programs for comparing substructures that automatically take allowed origin shifts and change-of-hand operators into account. We have developed a similar procedure as part of the Computational Crystallo-

Table 1
Results obtained with *HySS* and *SHELXD*.

Times are in CPU s (Intel Xeon, 2.7 GHz).

	HYSS				*SHELXD*							
Structure	Asked	Got	Correct	Time	Asked	Got	Correct	Trials	Time	Δ correct	Δ time	Speed-up
1167B	8	7	7	14	8	8	8	100	89	1	75	6.5
AA041	3	3	2	15	3	3	2	100	107	0	93	7.4
AEP-transaminase	66	66	64	1541	66	66	66	100	1037	2	−504	0.7
CP-synthase	16	15	14	52	16	16	16	100	332	2	281	6.4
Gpatase	22	19	17	218	22	22	20	100	881	3	663	4.0
KPHMT	160	138	133	8496	160	160	152	500	13007	19	4511	1.5
MEV-kinase	6	6	4	15	6	6	6	100	83	2	68	5.5
MP217	16	14	11	68	16	16	15	100	208	4	140	3.0
MP883	50	42	21	1494	50	50	37	200	1128	16	−366	0.8
NSF-D2	9	8	7	44	9	8	8	100	632	1	588	14.3
NSF-N	6	4	1	76	6	3	3	100	183	2	108	2.4
P32	9	9	9	29	9	9	9	100	118	0	90	4.1
P54	6	5	5	3005	6	6	6	100	1002	1	−2003	0.3
SEC17	3	1	0	219	3	3	2	100	204	2	−15	0.9
TM142	18	16	14	409	18	18	17	100	390	3	−19	1.0
TM384	4	3	2	76	4	4	2	100	327	0	251	4.3
UT-synthase	24	24	21	397	24	24	22	100	689	1	291	1.7

graphy Toolbox (for some comments regarding this procedure, see Grosse-Kunstleve & Adams, 2003*a*; details will be published elsewhere; the source code has been fully available for some time, including a web interface at http://cci.lbl.gov/cctbx/). We have embedded this procedure into the substructure search and use it in combination with the correlation coefficients to automatically terminate a search under certain conditions. Our current rule set is as follows.

1. All search results with correlations below 0.1 are discarded.

2. The difference between the top two correlations must be less than 0.05.

3. The lesser of the top two correlations must be at least 2.0 times the smallest correlation encountered or greater than a sliding threshold starting with 0.2.

4. If all the previous conditions are fulfilled, the substructures with the top two correlation coefficients are compared. The search is terminated if more than 2/3 of the number of expected sites match. Otherwise, the sliding threshold is increased by 0.05 up to a limit of 0.3.

(iv) Minimalistic command-line interface. The current implementation of our search procedure works directly with some common reflection file formats [*e.g.* merged *SCALEPACK* (Otwinowski & Minor, 1997) files]. The procedure is started with one command, with the file name for the reflection file, the expected number of sites and a label for the scattering type (*e.g.* `phenix.hyss w3.sca 8 Se`) as arguments. Apart from the reflection file, no other input is required.

6. Results

Table 1 shows our results with the *HySS* prototype and a comparison with results obtained with *SHELXD* on exactly the same difference data (computed as part of our procedure). Since *SHELXD* does not have a mechanism for automatic termination, we set the number of trials to 100 in most cases, a value that is probably a typical choice in practice. However, with the MP883 data *SHELXD* did not find a solution in the first 100 trials. A solution appeared only when we restarted the search for another 100 trials. The KPHMT entry is a special case that is discussed below.

As expected, the raw runtime per trial of our implementation is longer than that of *SHELXD*. This is mainly because of two reasons. Firstly, to minimize development time our prototype is written entirely in an interpreted language (Python). The worst-case expected performance penalty is about a factor of 100. However, the core components (such as the fast Fourier transformations) of the Computational Crystallography Toolbox are written in a compiled language. Therefore, the actual overall performance penalty is usually at most a factor of 2 or 3. Secondly, the two-atom fast translation searches are exhaustive searches compared with the random sampling performed by *SHELXD*.

Table 1 shows that the complete absence of application-specific low-level optimizations is in many cases offset by the high-level decision to automatically terminate the search. This is not always the case, but even in the worst case the performance penalty is less than a factor of 2, except for P54 (factor of 3) which is a special case and is discussed below. In general, searches terminate quickly; in one case (NSF-D2) the search is completed more than 14 times faster than *SHELXD*, although at the expense of one incorrect site.

We observe that for three structures (MP883, NSF-N, SEC17) the solutions produced by our procedure may not contain enough correct sites for successful phasing. We have correlated these failures with the presence of high thermal displacement factors. *SHELXD* accounts for this condition through variable occupancy factors. Currently, we are only

using fixed occupancies in the recycling procedure, but we are planning to implement optimization of occupancy factors or thermal displacement factors to model the substructure more accurately.

For KPHMT (160 expected sites) we used data that were merged with the *CCP*4 programs *SCALA* and *TRUNCATE* (von Delft, private communication). With these data *SHELXD* did not produce a correct solution after 100 trials. Repeating the search with 500 trials one (and only one) correct solution appeared after 343 trials (9317 s). However, *HySS* reproducibly found two solutions with 138 matching sites (133 correct) in 8496 s. We attribute this to the systematic sampling made possible by the fast translation function. It should be noted, however, that *SHELXD* produces one correct solution every ∼20 min on average with a differently merged data set (using *XPREP*) after careful manual selection of the resolution limit (von Delft, private communication). We have not yet analyzed these data.

We deliberately include a structure with a rare cubic symmetry in Table 1 (P54) to highlight the least desirable property of the fast translation function: the runtime scales with the fourth power of the number of symmetry operations (Navaza & Vernoslova, 1995). P54 crystallizes in a space group with 24 symmetry operations ($P4_132$; No. 213). In this case, 87% of the total runtime is consumed by the three applications of the fast translation function embedded in the search procedure. This is the worst case possible for macromolecular structures, since non-primitive settings (*e.g.* *F*432) can easily be transformed to non-standard primitive settings for the purpose of the search procedure.

7. Conclusions

The most important result of our experiments is that reliable automatic termination of the substructure search is possible. We followed this route because it is a very easy one to take in our flexible object-oriented framework. Fully embedding the independently developed substructure comparison into the search procedure only required the addition and modification of a few lines in the scripted source code. Of course, it is also possible to include automatic termination in other programs such as *Shake-and-Bake* or *SHELXD* and this would result in procedures that are in general faster than ours. However, because of the very limited abstraction facilities of the implementation language (Fortran) used by the other programs this would probably require significantly more development time.

In order to minimize development time, we have implemented our phase-manipulation procedure as a combination of Fourier transformations and squaring in real space (Fig. 1). This approach is relatively slow compared with the alternative reciprocal-space implementation (using the well known triplets usually associated with direct methods) because only the largest normalized structure factors are actually used; typically, most of the structure factors in reciprocal space are assumed to be zero. In addition, we are restricted in our choices of phase-manipulation protocols. Sophisticated protocols, such as those used in *SHELXD*, are prohibitively slow if implemented in real space. However, to address this issue we have already implemented a fast triplet generator (Grosse-Kunstleve & Adams, 2003*b*) suitable for integration into *HySS*.

We will continue the development of our procedure by refining the recycling algorithm to take variable occupancies or thermal displacement factors into account. Another focus will be parallelization of the systematic search afforded by the fast translation function. This is mostly trivial, but some minimal inter-process communication (*e.g.* through a shared file) is required for the automatic termination. We also plan to address the disproportionate share of the total runtime consumed by the fast translation functions in very high symmetries (P54 in Table 1). A relatively simple but very efficient measure would be to follow the example of *SHELXD* and to implement the extrapolation scan (currently accounting for 43% of the total runtime) by an application of the Nordman function. We believe that this would not significantly affect the behavior of the search procedure because the fast translation function would still be used to exhaustively sample the two-atom translations (currently accounting for 13% of the runtime). In such high-symmetry cases, initial recycling in *P*1 may not be warranted given the runtime penalty for recovering the origin with respect to the original symmetry (31% of the runtime for P54). We have not yet investigated under which specific conditions it is more beneficial to use *P*1 recycling. We will also continue to enhance our minimalistic command-line user interface to work directly with all common reflection file formats. This should make it possible to solve most substructures without the need to prepare any other input files or the need to run external programs. Within the Phenix package (Adams *et al.*, 2002) the graphical interface provides a convenient mechanism to adjust parameters for difficult searches.

So far we have not paid any attention to low-level optimization of the *HySS*-specific algorithms. Our prototype implementation relies on high-level code reuse in an object-oriented framework. It is unclear how much development time should be devoted to low-level optimizations. Even for the largest substructure in Table 1 the total runtime is measured in hours using a single CPU (approximately 2 h 20 min). Many synchrotron beamlines are equipped with multi-CPU clusters. Automatic searches run in parallel will often finish without human intervention after only a few minutes on such clusters. Therefore, it is unlikely that *HySS* will be a rate-limiting step in the overall procedure leading from the diffraction data to the refined structure, even without low-level optimizations.

8. Source code availability

HySS is implemented as part of the *PHENIX* package and will be made available for download at http://www.phenix-online.org/. All source code will be available free of charge for non-profit use. The core components (forming the bulk of the source code) are available as unrestricted open source at http://cctbx.sourceforge.net/.

We would like to thank G. Sheldrick for answering our questions and for challenging us to study the *SHELXD* source code. Suggestions by T. Terwilliger and R. Read resulted in improvements to our procedure. The KPHMT data were generously provided by F. von Delft. We thank M. Adams for permission to use unpublished data (P54) for testing. We are grateful to the following authors for making data available: Berkeley Structure Genomics Center, A. Brunger, G. Gilliland, E. Gordon, O. Herzberg, J. Sacchettini and J. Smith. Our work was funded in part by the US Department of Energy under contract No. DE-AC03-76SF00098 and by NIH/NIGMS under grant number 1P01GM063210.

References

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1948–1954.

Blundell, T. L. & Johnson, L. N. (1976*a*). *Protein Crystallography*, ch. 6.2. London: Academic Press.

Blundell, T. L. & Johnson, L. N. (1976*b*). *Protein Crystallography*, ch. 7.6. London: Academic Press.

Buerger, M. J. (1959). *Vector Space.* New York: Wiley.

Dall'Antonia, F., Baker, P. J. & Schneider, T. R. (2003). *Acta Cryst.* D**59**, 1987–1994.

Fan, H.-F., Woolfson, M. M. & Yao, J.-X. (1993). *Proc. R. Soc. London Ser. A*, **442**, 13–32.

Germain, G., Main, P. & Woolfson, M. M. (1970). *Acta Cryst.* B**26**, 274–285.

Giacovazzo, C. (1992). *Fundamentals of Crystallography*. IUCr/ Oxford University Press.

Grosse-Kunstleve, R. W. & Adams, P. D. (2003*a*). *IUCr Computing Commission Newsletter 1.* http://www.iucr.org/iucr-top/commun/ccom/newsletters/2003jan/.

Grosse-Kunstleve, R. W. & Adams, P. D. (2003*b*). *IUCr Computing Commission Newsletter 2.* http://www.iucr.org/iucr-top/commun/ccom/newsletters/2003jul/.

Grosse-Kunstleve, R. W. & Brunger, A. T. (1999). *Acta Cryst.* D**55**, 1568–1577.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.

Hendrickson, W. A. (1991). *Science*, **254**, 51–58.

Karle, J. & Hauptman, H. A. (1956). *Acta Cryst.* **9**, 635–651..

Langs, D. A. (2002). *J. Appl. Cryst.* **35**, 505.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *J. Appl. Cryst.* **27**, 613–621.

Navaza, J. & Vernoslova, E. (1995). *Acta Cryst.* A**51**, 445–449.

Nordman, C. E. (1966). *Trans. Am. Crystallogr. Assoc.* **2**, 29–38.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Patterson, A. L. (1935). *Z. Kristallogr. A*, **90**, 517–542.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1986). *Numerical Recipes. The Art of Scientific Computing.* Cambridge University Press.

Ramachandran, G. N. & Srinivasan, R. (1970). *Fourier Methods in Crystallography*, p. 36. New York: Wiley.

Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.

Sheldrick, G. M. (1985). *SHELXS86. Program for the Solution of Crystal Structures.* University of Göttingen, Germany.

Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423–431.

Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R. & Usón, I. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 233–245. Dordrecht: Kluwer Academic Publishers.

Smith, G. D. (2002). *J. Appl. Cryst.* **35**, 368–370.

Terwilliger, T. C. (1994). *Acta Cryst.* D**50**, 11–16.

Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.

Woolfson, M. M. (1961). *Direct Methods In Crystallography*, ch. 7.3. Oxford University Press.

Yao, J.-X. (1981). *Acta Cryst.* A**37**, 642–644.