Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 25 October 2006
Accepted 6 January 2007

An integer minimal principle and triplet sieve method for phasing centrosymmetric structures

Alexander B. Smith,[a] Hongliang Xu[b] and Nikolaos V. Sahinidis[a]*

[a]Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, 600 South Mathews Avenue, Urbana, IL 61801, USA, and [b]Hauptman-Woodward Medical Research Institute and Department of Structural Biology, State University of New York at Buffalo, 700 Ellicott Street, Buffalo, NY 14203, USA. Correspondence e-mail: nikos@uiuc.edu

In this paper, a new integer minimal principle model for centrosymmetric structures is presented; one which fully accounts for reciprocal-space phase shifts present in non-symmorphic space groups. Additionally, characterization of false minima of the model is done in terms of even and odd triplets. Based on this characterization, a triplet sieve method is proposed. First, Gaussian elimination using only a subset of reliable triplets is employed for phasing. Triplet subsets are generated using a progressively smaller set of the strongest reflections. Several phase solution sets are generated by enumerating the degrees of freedom present. To facilitate computational evaluation of the quality of these phase solutions, these phase sets are passed into the crystallographic software *SnB*, which expands the reflection set in two cycles. The final solution is identified *via* statistics of two crystallographic figures of merit. Computational results are presented for a variety of structures.

1. Background and theory

Full characterization of a crystal structure requires both amplitude and phase information for a sufficient number of structure factors. In a traditional X-ray diffraction experiment, the diffraction intensity is related to the amplitude of the structure factor. Unfortunately, the phase of a given structure factor cannot be determined from measurements of the diffraction intensity alone. The lack of phase information in this context has been coined 'the phase problem of X-ray crystallography'. While structure determination from X-ray diffraction data is employed almost on a routine basis worldwide, it is often a major challenge to solve the phase problem. The minimal principle, a method for phasing, involves the minimization of a cosine figure of merit, originally proposed by Debaerdemaeker & Woolfson (1983).

Model M1

$$\min R_{\min} = \sum_t A_t[\cos(\Phi_t) - \omega_t]^2 \Big/ \Big(\sum_t A_t\Big) \qquad (1)$$

$$\text{s.t.}\ \phi_{\mathbf{H}_t} + \phi_{\mathbf{K}_t} + \phi_{-\mathbf{H}_t-\mathbf{K}_t} = \Phi_t, \quad \forall t \in T, \qquad (2)$$

$$\phi_{\mathbf{H}_m} \in [0, 2\pi], \quad \forall m \in M, \qquad (3)$$

where M denotes the total number of reflections from an X-ray diffraction experiment after all the symmetry-equivalent reflections have been removed, T denotes the set of triplet phase invariants, $A_t = 2N^{-1/2}|E_{\mathbf{H}}||E_{\mathbf{K}}||E_{-\mathbf{H}-\mathbf{K}}|$, $\omega_t = I_1(A_t)/I_0(A_t)$, $|E|$ is a normalized structure-factor amplitude, N is the number of atoms in the unit cell, I_n is a modified Bessel function of order n, and $\mathbf{H}$ and $\mathbf{K}$ denote Miller indices. The unknowns in this model are the phases $\phi_{\mathbf{H}_m}$ for all $m \in M$ and the triplet invariants Φ_t for all $t \in T$. It has been demonstrated (Miller *et al.*, 1993) that a set of phases that minimizes $R_{\min}$ and satisfies proper atomicity constraints on electron density corresponds to the correct set of phases for the true crystal structure in all practical cases.

Solution of the minimal principle formulation is non-trivial, since the objective function is non-linear and non-convex. Hence, the solution space contains many local optima. Most contemporary algorithms for solving the minimal principle involve some combination of local search and stochastic optimization techniques, neither of which guarantee a globally optimal solution. Towards resolving the multiple local minima difficulty in this context, the minimal principle was reformulated for centrosymmetric structures into a mixed-integer linear program by Vaia & Sahinidis (2003). It is well known that the phases of a centrosymmetric structure are restricted to either 0 or π, provided the center of inversion is at the origin. This allows for modeling of the invariants modulo π with binary variables α and β:

$$\Phi_t = 2\alpha_t + \beta_t. \qquad (4)$$

The value of the cosine of the invariant can then be written in terms of β, since this variable determines whether or not the argument is odd or even:

$$\cos(\Phi_t) = 1 - 2\beta_t. \qquad (5)$$

This can then be substituted into the original minimal principle to derive the integer minimal principle of Vaia & Sahinidis (2003).

Model M2

$$\min R_{\min} = \sum_t A_t[4\beta_t + 1 - 2\omega_t + \omega_t^2]/\big(\sum_t A_t\big) \tag{6}$$

$$\text{s.t.}\, \phi_{\mathbf{H}_t} + \phi_{\mathbf{K}_t} + \phi_{-\mathbf{H}_t-\mathbf{K}_t} = 2\alpha_t + \beta_t, \quad \forall t \in T \tag{7}$$

$$\alpha_t, \beta_t \in \{0, 1\}, \quad \forall t \in T \tag{8}$$

$$\phi_{\mathbf{H}_m} \in [0, 2\pi], \quad \forall m \in M. \tag{9}$$

A non-zero global optimum of this integer minimal principle can be found in polynomial time using methods developed by Vaia & Sahinidis (2005).

The computational results in Vaia & Sahinidis (2003) and Vaia & Sahinidis (2005) demonstrated that using the above integer minimal principle solution as a starting point for phase searching in the *CRUNCH* crystallographic system (de Gelder *et al.*, 1993) is more advantageous than using a randomly generated starting set of phases. However, model M2 comes with two shortcomings:

(i) M2 does not account for phase shifts that are present in non-symmorphic crystallographic space groups;

(ii) M2 does not enforce atomicity constraints on electron density that have been found necessary in order to obtain meaningful sets of phases (Miller *et al.*, 1993).

As a result, a solution vector of M2 is unlikely to provide a correct structure, unless it is further processed by another phasing algorithm. These shortcomings are addressed in this manuscript.

2. Phase shifts

The set of symmetry elements associated with any space group defines relations among the atoms that make up the crystal. These symmetry elements can always be described in terms of a single rotational, **R**, and/or translational, **v**, operation. For example, in a crystal with rotational symmetry **R**, the electron densities, $\rho(\mathbf{x})$ and $\rho(\mathbf{Rx})$, at positions **x** and **Rx**, respectively, are identical:

$$\rho(\mathbf{x}) = \rho(\mathbf{Rx}). \tag{10}$$

Just as symmetry relations require certain equalities in direct space, an analogous situation exists in reciprocal space. Using the same symmetry operation in reciprocal space, we have

$$F_{\mathbf{H}} = \sum_j f_j \exp(2\pi i\mathbf{H} \cdot \mathbf{x}_j) \tag{11}$$

$$= \sum_j f_j \exp(2\pi i\mathbf{H} \cdot \mathbf{Rx}_j) \tag{12}$$

$$= F_{\mathbf{HR}}, \tag{13}$$

where $\mathbf{x}_j$ and f_j denote the position and atomic scattering factor of atom *j*, while $F_{\mathbf{H}}$ is the structure factor corresponding to reflection **H**. Clearly, then, the phase of **H** is equal to the phase of **HR** for the above example with only rotational symmetry. For the case of a symmetry operation with a rotational element, **R**, and translational element, **v**, we have

$$F_{\mathbf{H}} = \sum_j f_j \exp(2\pi i\mathbf{H} \cdot \mathbf{x}_j) \tag{14}$$

$$= \sum_j f_j \exp[2\pi i\mathbf{H} \cdot (\mathbf{Rx}_j + \mathbf{v})] \tag{15}$$

$$= \exp(2\pi i\mathbf{H} \cdot \mathbf{v}) \sum_j f_j \exp(2\pi i\mathbf{H} \cdot \mathbf{Rx}_j) \tag{16}$$

$$= \exp(i\sigma)F_{\mathbf{HR}}, \tag{17}$$

where σ is a phase shift. For centrosymmetric crystals, this phase shift is restricted to values of 0 or π.

The original integer minimal principle formulation M2 does not include any additional symmetry requirements in the constraint set. Lack of these extra constraints allows for a simple solution approach to the original formulation (Vaia & Sahinidis, 2005). Briefly, the objective function of the original formulation is minimized when all β_t equal 0, reducing the problem to that of solving a system of homogeneous equations defined by the constraints. The system is homogeneous, so it will always have a solution. Further, based on origin selection for a particular space group, a certain number of degrees of freedom will always exist in the homogeneous solution. In $P2_1/c$, for instance, regardless of how many triplet relations are used, there will always be a minimum of three degrees of freedom. Since these degrees of freedom can take values of either 0 or π, a non-trivial solution to the original minimal principle will always exist. Unfortunately, most such solutions produce phases with inconsistent symmetry relations and, thus, do not correspond to a meaningful structure.

The proposed new integer minimal principle ensures that phases maintain their symmetry relations. Since the integer minimal principle requires that the crystal be centrosymmetric, symmetry-related phases are either equal or differ by a factor of π. For a reflection with Miller index **H**, let *S* be the set of symmetry reflections related to the base reflection by a shift of π. Similarly, let *U* be the set of reflections related to the base reflection by a shift of 0. Each symmetry reflection *s* or *u* is related to its corresponding base reflection by a rotation $\mathbf{R}_s$ or $\mathbf{R}_u$ in equal order

$$\phi_{\mathbf{H}} + \phi_{\mathbf{HR}_s} = 1 \quad s \in S \tag{18}$$

$$\phi_{\mathbf{H}} = \phi_{\mathbf{HR}_u} \quad u \in U, \tag{19}$$

where it is clear that the Miller index of the symmetry-related reflection is $\mathbf{HR}_s$ in the shifted case and $\mathbf{HR}_u$ in the unshifted case. Additionally, the cardinalities of *S* and *U* can be related to the total number of symmetry operations in a given space group υ as follows:

$$\operatorname{card}(S) + \operatorname{card}(U) = \upsilon. \tag{20}$$

In other words, the set of all $\mathbf{HR}_s$ and $\mathbf{HR}_u$ for every reflection is the full sphere of data for the experiment. Adding the constraints defined by symmetry relations (18) and (19) to the original integer minimal principle M2 yields the new integer minimal principle.

Table 1
Crystal information for test structures.

Structure No.	Chemical formula	Atoms (ASU)	Space group	Resolution (Å)	Reference
1	$C_{30}H_{32}N_2O_6$	19	$P2_1/c$	0.84	Sun *et al.*, (2002)
2	$C_{44}H_{38}O_4$	24	$P\bar{1}$	0.84	Vande Velde *et al.* (2002)
3	$C_{36}H_{62}$	36	$P2_1/c$	0.75	Bragg *et al.* (2002)
4	$C_{30}H_{22}O_6S$	37	$P2_1/c$	0.82	Krishnakumar *et al.* (2002)
5	$C_{34}H_{26}N_2O$	37	$P2_1/c$	0.84	Zhuang *et al.* (2002)
6	$C_5H_{12}NO^+ \cdot C_{28}H_{37}B_6N_2O_{10}^- \cdot 0.5C_4H_{10}O$	55.5	$P\bar{1}$	0.76	Kliegel *et al.* (2002)
7	$C_{41}H_{78}O_{11}Si_8$	60	$P\bar{1}$	0.81	Arnold & Blake (2001)
8	$C_{50}H_{66}O_6 \cdot C_3H_7NO$	61	$P2_1/c$	0.84	Bryan & Levitskaia (2002)
9	$3C_{40}H_{32}O_2 \cdot 4C_6H_6$	75	$P\bar{1}$	0.84	Ohba *et al.* (2002)

Model M3

$$\min R_{\min} = \sum_t A_t[4\beta_t + 1 - 2\omega_t + \omega_t^2] \Big/ \Big(\sum_t A_t\Big) \tag{21}$$

$$\text{s.t.}\ \phi_{\mathbf{H}_t} + \phi_{\mathbf{K}_t} + \phi_{-\mathbf{H}_t-\mathbf{K}_t} = 2\alpha_t + \beta_t, \quad \forall t \in T, \tag{22}$$

$$\phi_{\mathbf{H}_m} + \phi_{\mathbf{H}_m\mathbf{R}_s} = 1, \quad \forall m \in M, \forall s \in S_m, \tag{23}$$

$$\phi_{\mathbf{H}_m} = \phi_{\mathbf{H}_m\mathbf{R}_u}, \quad \forall m \in M, \forall u \in U_m, \tag{24}$$

$$\alpha_t, \beta_t \in \{0, 1\}, \quad \forall t \in T, \tag{25}$$

$$\phi_{\mathbf{H}_m} \in \{0, 1\}, \quad \forall m \in M, \tag{26}$$

where M, as before, denotes the total number of reflections from an X-ray diffraction experiment after all the symmetry-equivalent reflections have been removed, S_m is the set of shifted phases related to $\mathbf{H}_m$ by rotational symmetry $\mathbf{R}_s$ and U_m is the set of unshifted phases related to $\mathbf{H}_m$ by rotational symmetry $\mathbf{R}_u$. In the absence of (23) and (24), Vaia & Sahinidis (2003) and Vaia & Sahinidis (2005) found that setting all β_t to 0 solves M2 and that the global solution of the model can be obtained by linear algebra on the homogeneous system. In the presence of shifts, however, the constraint set of M3 can have no solution after all β_t are set to 0. The latter case requires the use of integer programming software, such as *CPLEX* (ILOG, 2003), to solve the above integer minimal principle model. This, in general, could be a difficult task. In addition, the model in its present form does not account for atomicity constraints. In the next section, we propose a way to deal with atomicity constraints. Moreover, the proposed approach avoids the need for an integer programming software. Instead, we rely on the simple and efficient linear algebra techniques: sparse Gauss elimination with a Markowitz pivot selection.

3. Atomicity constraints characterization *via* even triplets

A phase search done entirely in reciprocal space, which yields a globally optimal $R_{\min}$ value, may converge to so-called *false minima*, minima that do not correspond to a true set of phases (Xu *et al.*, 2000). These false minima are characterized by the presence of a large 'uranium' peak on the corresponding Fourier map and do not correspond to a meaningful structure. For symmorphic space groups, an obvious example of a false-minimum solution for the minimal principle model is that of all phases set to zero. Unfortunately, there are additional, less trivial and more difficult to characterize, false minima for all space groups.

It is useful to define some terminology with regard to a true phase solution for a given structure. We define a triplet for which

$$\Phi = \phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}} = 0 \,(\operatorname{mod} 2) \tag{27}$$

as an *even* triplet and a triplet for which

$$\Phi = \phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{-\mathbf{H}-\mathbf{K}} = 1 \,(\operatorname{mod} 2) \tag{28}$$

as an *odd* triplet. This classification is useful since odd triplets are always present in true phase sets (Xu & Hauptman, 2004). Further, the odd triplets characterize the gap between the unconstrained $R_{\min}$ and the $R_{\min}$ of the true structural solution.

Inclusion of odd triplets in a traditional direct-method technique can have a detrimental impact on solution quality. Thus, characterization of odd triplets has been extensively explored in relevant direct-methods literature. Traditionally, odd triplets have been identified using quintet extension (*cf.* Giacovazzo, 1976, 1977). Previous attempts have focused on filtering triplet lists to remove odd triplets before phasing (Gilmore & Hauptman, 1985).

It has been observed that there is a correlation between the even triplets and their associated A values: those triplets having large A values have a high probability of being even (Xu & Hauptman, 2004). This observation coincides with the conclusion derived from the Sayre equation, which explores the relationship among the normalized structure factors in a crystal:

$$E_{\mathbf{H}} = \theta_{\mathbf{H}} \sum_{\mathbf{K}} E_{\mathbf{K}} E_{\mathbf{H}-\mathbf{K}}. \tag{29}$$

This equation can also be written in terms of amplitude and phase information:

$$|E_{\mathbf{H}}| \exp(i\phi_{\mathbf{H}}) = \theta_{\mathbf{H}} \sum_{\mathbf{K}} |E_{\mathbf{K}}||E_{\mathbf{H}-\mathbf{K}}| \exp[i(\phi_{\mathbf{K}} + \phi_{\mathbf{H}-\mathbf{K}})]. \tag{30}$$

Arguments with a very large amplitude on the right-hand side of the last equation will have an angular component similar to that of the left-hand side. Hence, the following approximation is particularly valid:

$$\phi_{\mathbf{H}} \approx \phi_{\mathbf{K}} + \phi_{\mathbf{H}-\mathbf{K}}. \tag{31}$$

Table 2
Triplet information for test structures.

Structure	Even triplets (%)	Odd triplets (%)	First odd	Total triplets
1	89.3	10.7	210	3800
2	93.5	6.5	555	4800
3	88.6	11.4	125	13780
4	93.5	6.5	96	3700
5	86.8	13.2	108	4200
6	89.6	10.4	267	3700
7	98.9	1.1	1371	4600
8	86.6	13.4	61	6100
9	89.1	10.9	142	5500

Since it is the product of $E_{\mathbf{H}}$, $E_{\mathbf{K}}$ and $E_{\mathbf{H}-\mathbf{K}}$, which defines a triplet A value, a triplet with a very large A value has a high probability of being even.

4. The triplet sieve method

A set of test structures along with their important crystallographic information, including chemical formula, non-H atoms in the asymmetric unit, space group and reference, are listed in Table 1. In *SnB* (Weeks & Miller, 1999), triplets and their associated A values are typically generated from $M = 10N$ reflections having the largest $|E|$ values, where N is the number of non-H atoms in the asymmetric unit. All triplets are sorted in descending order of A values and the strongest $T = 100N$ triplets are selected to construct the input data for the integer minimal principle M3.

From the deposited atomic coordinates in CIF files, and based on the assumption that atomic scattering factors equal the atomic number of the elements, we calculate the 'true' phases of the structures and, from these, the percentages of even and odd triplets listed in Table 2. For each structure, the table also lists the occurrence of the first odd triplet when triplets are sorted in descending order of A values. From Table 2, one should immediately notice that the percentage of odd triplets is significantly smaller than the percentage of even triplets. In addition, the position of the first odd triplet in the sorted list of triplets, ranging from 61st for structure 8 to 1371st for structure 7, suggests that, for each structure, there exists a number $T' < T$ such that $\Phi_t = \phi_{\mathbf{H}_t} + \phi_{\mathbf{K}_t} + \phi_{-\mathbf{H}_t-\mathbf{K}_t}$, $1 \le t \le T'$, are even triplets. When sufficient degrees of freedom are present, the global minimum of M3 can be found by solving a set of linear equations. This set of T' triplets can always be chosen such that the system is underdetermined. Thus, the phases involved in these T' triplets can be calculated by setting all β_t to 0. It then remains to solve the following system for the phases.

Model M4

$$\phi_{\mathbf{H}_t} + \phi_{\mathbf{K}_t} + \phi_{-\mathbf{H}_t-\mathbf{K}_t} = 0, \quad \forall t \in T' \tag{32}$$

$$\phi_{\mathbf{H}_m} + \phi_{\mathbf{H}_m\mathbf{R}_s} = 1, \quad \forall m \in M, \ \forall s \in S_m \tag{33}$$

$$\phi_{\mathbf{H}_m} = \phi_{\mathbf{H}_m\mathbf{R}_u}, \quad \forall m \in M, \ \forall u \in U_m \tag{34}$$

$$\phi \in \{0, 1\}. \tag{35}$$

System M4 can be solved in polynomial time using sparse Gaussian elimination in a binary field as was done by Vaia & Sahinidis (2005) for model M2, which does not involve equations (33) and (34). Of course, system M4 has multiple solutions because the number of phases is greater than the number of equations. The exact number of solutions to system M4 depends on the number of degrees of freedom in M4. This number is the subject of the next subsection.

4.1. Degrees of freedom

It is clear from Table 2 that a small set of the strongest triplets will contain, if any, a very small number of odd triplets. Unfortunately, such a system of equations may also have a very large number of degrees of freedom. For a system with δ degrees of freedom, one would have to consider 2^{δ} possible phase solutions. To curtail computational effort, it is thus important to develop ways for reducing the number of degrees of freedom. On the other hand, a very small number of degrees of freedom may also be unacceptable since the number of reflections that can be phased is determined solely by their participation in the truncated triplet set. There are two possible approaches to address this trade-off between

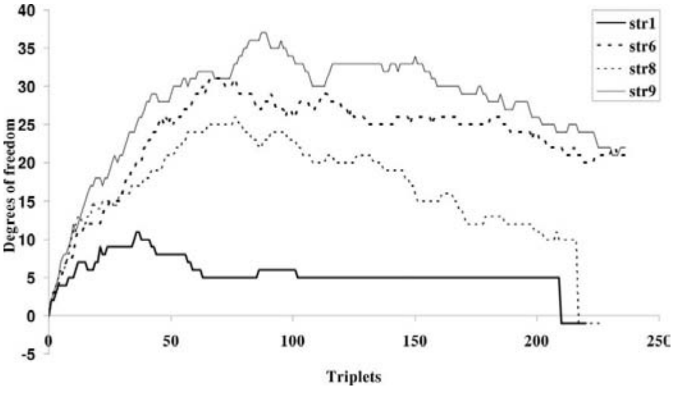

Figure 1
Degrees of freedom for truncated triplet sets: Strongest triplets generated from 10N reflections.

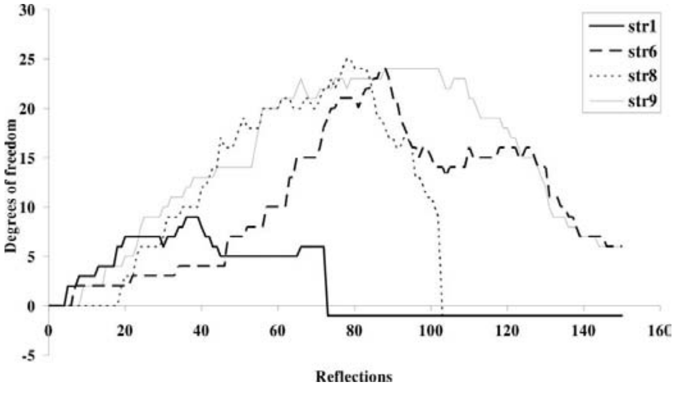

Figure 2
Degrees of freedom for truncated reflection sets: all triplets generated from strongest truncated reflection sets

reducing the computational effort and obtaining a meaningful solution.

(i) Triplet generation can be done based exclusively on the use of strongest triplets. Fig. 1 illustrates how the number of degrees of freedom of system M4 increases as the triplet list (generated from $M = 10N$ reflections) is progressively truncated for a selection of structures. A value of -1 for the degrees of freedom indicates that system M4 is overdetermined. As seen in this figure, in all cases, truncation that is sufficient to ensure that no odd triplets are present may also raise the number of degrees of freedom to an unacceptable level from the computational point of view. For instance, for structure 9 (top line in the figure), with over 30 degrees of freedom, one would be required to consider over 1 billion phase solution sets. Yet the figure indicates that, for all structures, the curves are approximately unimodal, thus suggesting that the identification of a cut-off T' on the increasing part of the curves may suffice to solve the problem.

(ii) Triplet generation can also be done based exclusively on a subset of the strongest reflections. For a given set of strong reflections, one can simply generate all possible triplet invariants, thus generating a subset of the constraints of system M4. Fig. 2 illustrates the number of degrees of freedom as a function of the number of reflections for the same structures considered in Fig. 1. It is important to note that triplets constructed from strong reflections will also have large A values. Thus, the triplets generated from a truncated reflection set will, in practice, have a small number of odd triplets while maintaining a reasonably small number of degrees of freedom.

For structures in a non-symmorphic space group, there is a limit on the total number of triplets that can be used since, as the number of triplet relations increases, the system will eventually have no solution. Further, Figs. 1 and 2 suggest that identifying a suitable subset of triplets based on a subset of strongest reflections is likely to result in fewer degrees of freedom. This method also has the advantage of solving for the strongest reflections, which will then result in a more accurate electron-density map. Hence, we opt to use this generation technique.

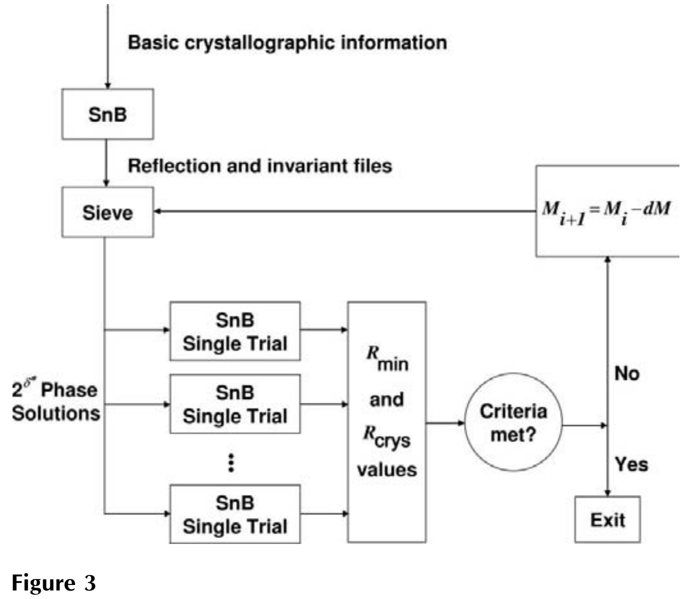
Basic crystallographic information
SnB
Reflection and invariant files
Sieve
$M_{i+1} = M_i - dM$
2^{δ^*} Phase Solutions
SnB Single Trial
SnB Single Trial
SnB Single Trial
$R_{\rm min}$ and $R_{\rm crys}$ values
Criteria met?
No
Yes
Exit

Figure 3
Incorporating the triplet sieve into *SnB*

No matter whether generation of triplets is done based on strongest triplets or reflections, the number of degrees of freedom can be further reduced *via* the following two approaches.

(i) The origin of a crystal structure has to be fixed in order to assign individual phase values. In the space groups explored in this paper, three linearly independent reflections can be identified and assigned phase values arbitrarily, thus reducing the number of degrees of freedom by $\delta_{\rm origin} = 3$.

(ii) In some triplet systems, it is possible for a reflection to participate in only one triplet equation. In these situations, wrong assignment of the phase value for this reflection, termed isolated reflection, will only affect solution quality marginally. Hence, isolated reflections do not represent critical degrees of freedom in the system and their values can also be assigned arbitrarily. We denote the number of these degrees of freedom by $\delta_{\rm isolated}$.

We can now define the number of reduced degrees of freedom, δ^*, as $\delta^* = \delta - \delta_{\rm isolated} - \delta_{\rm origin}$. Once the number of reduced degrees of freedom has been identified, the corresponding sets of 2^{δ^*} phase solutions will need to be explored, as opposed to 2^{δ} solutions if the degrees of freedom are not reduced. As an example, the system of triplets generated from the top 102 reflections of structure 8 has $\delta = 9$ and $\delta^* = 2$. Thus, elimination of isolated and origin-related phases reduces the number of phase solutions for this structure from $2^9 = 512$ to $2^2 = 4$, while one out of the latter 4 sets of solutions still provides a true set of phases.

Once the phases corresponding to strong reflections are found, one can proceed to construct an electron-density map, generate atomic positions after peak picking, and use these atomic positions to generate a complete set of phases.

5. Incorporating the triplet sieve method into *SnB*

A progressively smaller set of reflections will effectively remove odd triplets and, consequently, improve the likelihood of converging one of the 2^{δ^*} solutions to a true structural solution. Once the number δ^* of reduced degrees of freedom has been identified, the corresponding 2^{δ^*} solutions can then be expanded to complete sets of phases *via* a crystallographic computing program such as *SnB* (Weeks & Miller, 1999). Our implementation of the triplet sieve procedure within *SnB* is outlined in Fig. 3.

First, a set of traditional inputs is provided to the *SnB* software package. This includes unit-cell parameters, diffraction wavelength, cell contents, a space-group specification and *hkl* data. This information is then used by *SnB* to produce formatted reflection and triplet invariant files. These inputs are then passed to the triplet sieve, along with a specification for how many reflections, dM, should be removed after each iteration.

The sieve first probes for a good starting number of reflections. This is done based on two main criteria: (*a*)

Table 3
Sieve performance on test structures.

Structure	M_0	$M_{\rm final}$	Deduction	δ^*	Expansion	$R_{\rm min}$	$R_{\rm crys}$	MPE (°)
1	72	52	3	2	10	0.313	0.24	0.0
2	100	90	2	4	11	0.170	0.21	0.0
3	109	99	2	7	37	0.790	0.27	6.9
4	65	55	2	3	16	0.206	0.30	1.5
5	95	85	2	4	10	0.434	0.22	1.0
6	146	126	3	7	41	0.370	0.18	1.3
7	119	119	1	5	7	0.067	0.24	2.1
8	102	102	1	2	4	0.587	0.23	3.3
9	132	132	1	4	15	0.354	0.21	0.5

sufficient number of degrees of freedom and (*b*) evidence of local decrease in the degrees of freedom with respect to decreasing number of reflections. The number of degrees of freedom for a successful sieve must be greater than the number of reflections required to specify an origin: this prevents solution of a homogeneous system for which the phases calculated are equivalent to the solution of all phases set to zero. This threshold can easily be converged upon by testing the degrees of freedom in a system of T triplets generated from M_i reflections using a binary search starting from $M_{\rm max}$. Here, $M_{\rm max}$ can be the total number of reflections recorded in the data file. However, $M_{\rm max}$ is typically set to the number of reflections *SnB* phases with, *i.e.* $10N$. Once the number of reflections to be used has been determined, it is still important to determine whether or not the number of degrees of freedom decreases for a small decrease in the number of reflections. This is particularly important for symmorphic space groups; in these situations, the set of triplets to sieve with can be arbitrarily large since the system of equations generated is always homogeneous. Sufficient descent is currently defined as when the number of degrees of freedom drops by at least three when the number of reflections is reduced by dM. No phase solutions are calculated during the probing phase; the probing process merely calculates the number of degrees of freedom of M4 for different subsets T' of T.

Next, execution proceeds into sieve deduction/expansion iterations. The sieve will generate a set of triplets based on M_i. This set of triplets is used to find solutions for all the reflections at M_i, in terms of a specific number of reduced degrees of freedom, δ^*. The phase solutions are generated by enumerating all possible degrees of freedom. Each of the thus generated 2^{δ^*} phase solutions is then passed to *SnB* for phase expansion *via* a dual-space cycling. The known phases from the triplet sieve method are fixed and unknown phases, with random initial values, are subject to change during the cycling. After a predetermined number of *SnB* cycles, the resultant minimal function values, $R_{\rm min}$, and crystal R values, $R_{\rm crys}$, are used for automatic structure solution detection. If an acceptable solution is identified, the sieve procedure terminates; otherwise, we set $M_{i+1} = M_i - {\rm d}M$, and the sieve procedure is repeated.

Currently, the method terminates when any one of the following criteria is met.

(i) When M_i falls below some lower bound. This criterion is based on an estimation of when, despite the accuracy of the phases provided, *SnB* is unable to converge in a predetermined number of *SnB* cycles to a proper solution.

(ii) When δ^* is too high for more than one consecutive iteration. By default, the sieve considers an iteration with $\delta^* \geq 7$ as having too many solutions to enumerate.

(iii) When the highly effective solution detection criterion of Xu *et al.* (2006) is satisfied. This criterion uses the standard deviation and average of $R_{\rm min}$ and $R_{\rm crys}$ from previous trials, and is applied once the number of phase expansions exceeds five.

6. Results and discussion

The sieve procedure was tested on each of the structures presented in Table 1. First, using the deposited CIF file, basic crystallographic information was input to *SnB*, including space group, cell parameters, chemical formula and radiation type. Then, using the *DREAR* package in *SnB*, E values were generated from the experimental *hkl* data provided by the authors listed in Table 1. As a final initialization step, *SnB* was used to generate reflection and invariant files, typically containing $10N$ reflections and $100N$ triplets. Then, a sieve run was executed using ${\rm d}M = 10$ for every structure. The results of running the sieve method on each of the nine sample structures are presented in Table 3. In this table, M_0 is the number of strongest reflections the probe terminated on, $M_{\rm final}$ is the number of strongest reflections the triplet sieve method used in the final iteration in order to solve the problem, deduction is the number of sieve iterations, *i.e.* the number of times that M was decremented by dM, δ^* is the number of reduced degrees of freedom in the final iteration, expansion is the total number of *SnB* jobs for phase expansion, $R_{\rm min}$ is the minimal function value for the identified solution, $R_{\rm crys}$ is the crystallographic merit function for the identified solution, and MPE is the mean phase error based on atomic coordinates provided in the CIF for each structure.

First and paramount, all of these solutions from the sieve method generated density maps with geometry and connectivity consistent with the published structural solutions. Second, the number of *SnB* cycles for phase expansion was reduced to 2 for the identified solutions, rather than the default value of N. The comparison of the sieve method *versus* *SnB*, based on the (average) number of dual-space cycles needed for producing a correct structure solution, is provided in Table 4, where the second and the third columns indicate the total number of dual-space cycles needed to produce the first solution from sieve and *SnB*, respectively, while the fourth column is the average number of cycles per *SnB* solution, calculated by a formula of $1000C/S$, based on the number of solutions (S) produced from 1000 *SnB* trials with a default number of cycles (C) per trial. This comparison shows that the sieve method is anywhere from two times to several thousand times faster in terms of cycles run than *SnB* for producing a correct structure solution.

Table 4
Comparison of sieve and *SnB*: number of dual-space cycles needed for producing a correct structure solution.

Structure	Sieve	*SnB*	⟨*SnB*⟩
1	20	60	167
2	22	672	632
3	74	144	419
4	32	180	257
5	20	90	367
6	82	616	2545
7	14	18300	30000
8	8	1380	612
9	30	629	2643
Average	34	2452	4182

Table 5
Comparison of sieve and *SnB*: computation time in seconds required for producing a correct structure solution.

Structure	Sieve	*SnB*	⟨*SnB*⟩
1	0.6	0.3	0.8
2	0.8	6.7	6.3
3	3.0	4.3	12.6
4	1.2	3.6	5.1
5	0.7	0.9	3.7
6	4.4	18.5	76.4
7	3.1	366	600
8	0.8	41.4	18.4
9	1.8	18.9	79.3
Average	1.8	51.2	89.2

The sieve does introduce some overhead associated with the linear system solution and enumeration of degrees of freedom. The total computation time exerted by the sieve *versus SnB* is shown in Table 5. The second column shows the sieve-modified *SnB* total time, the third column shows the time required for the unmodified *SnB* to find the first correct structural solution, and the fourth column presents the average time required for *SnB* to find a structural solution based on 1000 trials. Computational time was assessed on a 3.0 GHz P4 Linux workstation with 1 GB of RAM. It is clear from Table 5 that, with the exception of the smallest structure (structure 1), the sieve reduces the total time required to obtain a phase solution. Furthermore, the computational time requirements by the sieve do not seem to be affected significantly by the size of the structure, while those of the unmodified *SnB* seem to increase with structure size.

The effect of reduced data resolution was also explored. At 1 Å, structures 1 through 7 completed with a similar amount of computational effort. Structures 8 and 9, however, possessed too many odd triplets in acceptable degree-of-freedom truncation regions. Naturally, as expected, our reliance on direct methods results in typical observations with regard to solution limits. As structure size increases or the data resolution decreases, the distribution of odd triplets in small truncated sets increases. None of the structures produced solutions beyond resolution of 1.2 Å.

The purpose of the probe is to quickly identify a starting M_0, which provides a suitable number of degrees of freedom according to the criteria discussed above. By referencing Table 3, it becomes clear that the probe did indeed choose a starting M_0 for each structure which ensured that δ^* is greater than the number of reflections required to specify an origin, and on a point with appreciable slope in δ^* *versus* the number of reflections. In addition, the probe effectively chose an M_0 for which a solution existed in the inhomogeneous shifted systems. Considering all the structures, the number of sieve iterations to find a solution did not exceed three. This is an indication that, given dM, a solution was always identified in a range of 30 reflections from the starting point for these test structures. This result supports the effectiveness of the probe and demonstrates how quickly odd triplets are removed for low reflection sets.

Finally, despite ignoring isolated reflections, the MPE for each of the structures is low. In the case of structures 1 and 2, there were no isolated degrees of freedom so the iteration converged to the exact solution. This also illustrates that identification of an exact solution for a structure would simply require enumeration of the isolated reflections. Doing this would double the number of trial solutions required per isolated reflection enumerated. Finally, the unusually high MPE in the solution for structure 3 is actually due to an odd triplet that was present in the identified solution's system of equations. This odd triplet was assumed to be even, and hence produced a phase solution which is slightly in error. Despite the existence of this odd triplet, however, a solution which closely resembles the geometry of the published solution was determined. This illustrates that, in some cases, the sieve is not completely intolerant of odd triplets. Finally, an exact solution to this structure can be found by starting at a lower M_0.

7. Conclusions

This paper has proposed a new integer minimal principle for phasing centrosymmetric structures, along with a triplet sieve method to determine a small number of systems of equations, the solution to one of which yields a correct set of phases. Incorporation of symmetry relations into the integer minimal principle model allows for proper solution of structures in non-symmorphic space groups. The computational results demonstrated that the triplet sieve is an effective means for producing a highly accurate partial phase solution by solving a system of equations generated from a small set of strongest reflections. When coupled with *SnB*, each trial solution can be expanded to a full phase set and produce a corresponding $R_{\rm min}$ and $R_{\rm crys}$. The statistics of these crystallographic figures of merit are then sufficient to identify when the sieve has produced a true structural solution. While we do observe the typical limitations of direct methods regarding structure size and resolution, a large increase in computational efficiency results for all of the tested structures. Similar structures are solved on a routine basis in a high-throughput manner. Hence, a tenfold increase in efficiency over a large number of crystal structures solved in a continuous fashion would represent a reduction in time from hours to minutes for a particular large-scale run.

An interesting future direction would be the study of the effect of adding quartets to the current model, as it seems that quartets could effectively lower the number of degrees of freedom even further, thus further improving the accuracy and speed of the proposed method. Additionally, instead of relying on progressive triplet set truncation, the availability of higher-order invariants could be used to assess the even or odd character of each triplet. Proper identification of included odd triplets would certainly reduce the required number of enumerated solutions and serve to push the current size and resolution limits.

The authors would like to express their appreciation to Charles M. Weeks and Herbert A. Hauptman of Hauptman–Woodward Medical Research Institute for very valuable discussions and suggestions that motivated this work. This work was supported in part by the Joint NSF/NIGMS Initiative to Support Research in the Area of Mathematical Biology under NIH award GM072023.

References

Arnold, P. L. & Blake, A. J. (2001). *Acta Cryst.* E**57**, o131–o133.

Bragg, S., Johnson, J. E. B., Graziano, G. M., Balaich, G. J. & Heimer, N. E. (2002). *Acta Cryst.* E**58**, o1010–o1012.

Bryan, J. C. & Levitskaia, T. G. (2002). *Acta Cryst.* E**58**, o240–o242.

Debaerdemaeker, T. & Woolfson, M. M. (1983). *Acta Cryst.* A**39**, 193–196.

Gelder, R. de, de Graaff, R. A. G. & Schenk, H. (1993). *Acta Cryst.* A**49**, 287–293.

Giacovazzo, C. (1976). *Acta Cryst.* A**32**, 967.

Giacovazzo, C. (1977). *Acta Cryst.* A**33**, 527–531.

Gilmore, C. J. & Hauptman, H. (1985). *Acta Cryst.* A**41**, 457–462.

ILOG (2003). *CPLEX 9.0 User's Manual.* ILOG CPLEX Division, Incline Village, NV, USA.

Kliegel, W., Druckler, K., Patrick, B. O., Rettig, S. J. & Trotter, J. (2002). *Acta Cryst.* E**58**, o393–o395.

Krishnakumar, R. V., Subha Nandhini, M., Renuga, S., Natarajan, S., Selvaraj, S. & Perumal, S. (2002). *Acta Cryst.* E**58**, o1174–o1176.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. (1993). *Science*, **259**, 1430–1433.

Ohba, S., Hiratsuka, T. & Tanaka, K. (2002). *Acta Cryst.* E**58**, o1013–o1015.

Sun, G. C., Li, Y. Z., He, Z. H., Li, Z. J., Qu, J. Q., Liu, C. R. & Wang, L. F. (2002). *Acta Cryst.* E**58**, o417–o418.

Vaia, A. & Sahinidis, N. V. (2003). *Acta Cryst.* A**59**, 452–458.

Vaia, A. & Sahinidis, N. V. (2005). *Acta Cryst.* A**61**, 445–452.

Vande Velde, C. M. L., Hoefnagels, R. & Geize, H. J. (2002). *Acta Cryst.* E**58**, o454–o455.

Weeks, C. M. & Miller, R. (1999). *J. Appl. Cryst.* **32**, 120–124.

Xu, H. & Hauptman, H. A. (2004). *Acta Cryst.* A**60**, 153–157.

Xu, H., Weeks, C. M., Deacon, A. M., Miller, R. & Hauptman, H. A. (2000). *Acta Cryst.* A**56**, 112–118.

Xu, H., Weeks, C. M. & Hauptman, H. A. (2006). Am. Crystallogr. Soc. Annual Meeting, Hawaii, HL, USA, Abstract W0513.

Zhuang, J.-P., Zheng, Y. & Zhang, W.-Q. (2002). *Acta Cryst.* E**58**, o720–o722.