Acta Crystallographica Section A
**Foundations of Crystallography**
ISSN 0108-7673

Received 8 April 2003
Accepted 11 June 2003



# An integer programming approach to the phase problem for centrosymmetric structures

**Anastasia Vaia and Nikolaos V. Sahinidis***

Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, 600 South Mathews Avenue, Urbana, IL 61801, USA. Correspondence e-mail: nikos@uiuc.edu

The problem addressed in this paper is the determination of three-dimensional structures of centrosymmetric crystals from X-ray diffraction measurements. The 'minimal principle' that a certain quantity is minimized only by the crystal structure is employed to solve the phase problem. The mathematical formulation of the minimal principle is a nonconvex nonlinear optimization problem. To date, local optimization techniques and advanced computer architectures have been used to solve this problem, which may have a very large number of local optima. In this paper, the minimal principle model is reformulated for the case of centrosymmetric structures into an integer programming problem in terms of the missing phases. This formulation is solvable by well established combinatorial optimization techniques that are guaranteed to provide the global optimum in a finite number of steps without explicit enumeration of all possible combinations of phases. Computational experience with the proposed method on a number of structures of moderate complexity is provided and demonstrates that the approach yields a fast and reliable method that resolves the crystallographic phase problem for the case of centrosymmetric structures.

## 1. Introduction

Since the mid-1900's, analysis of X-ray diffraction data of crystals has been used extensively for the determination of molecular structure and properties. While the method is employed almost on a routine basis worldwide, it is often a major challenge to identify the three-dimensional structure that best fits the diffraction data. A key obstacle, in particular, is the identification of the phases of the diffracted rays from measurements of intensities alone.

Methods developed for the phase problem have included the tangent formula (Karle & Hauptman, 1956), the maximum entropy (Bricogne, 1984), the minimal principle (Debaerdemaeker & Woolfson, 1983), and variants of the above (Germain & Woolfson, 1968; Germain *et al.*, 1971; Olthof & Schenk, 1982; Gull *et al.*, 1987; Hauptman, 1988; Sheldrick, 1990; Altomare *et al.*, 1993; Miller *et al.*, 1993; Gilmore, 1996; Sheldrick, 1997; Chang *et al.*, 1997; Giacovazzo, 1998; Hauptman *et al.*, 1999).

Most of the methods for the phase problem make use of a merit function to score potential structures based on how well they match the experimental data. The complexity of the resulting phase-estimation problem is significant because of the existence of multiple local optima in the underlying optimization formulations. To this date, crystallographers have resorted to combinations of local optimization and stochastic global optimization techniques to solve these models. For example, the *Shake-and-Bake* approach (Miller *et al.*, 1993) is based on alternating phase refinement in reciprocal space with a peak-picking technique in real space and terminates once a prespecified number of iterations has been reached.

For centrosymmetric structures, it has long been observed that the phases can only take values of 0 or $\pi$. While this, effectively, makes the phase problem a discrete optimization problem, no current solution strategy exploits the mathematical properties of the problem to effectively resolve the phase problem. Yet a very large number of crystal structures are centrosymmetric. For instance, nearly 76% of the over a quarter of a million crystal structures in the Cambridge Structural Database are centrosymmetric (Allen, 2002).

In this paper, we address the problem of using X-ray measurements to determine structures with a center of symmetry. Our starting point is the minimal principle model. In the general case, this model requires the solution of a highly nonlinear nonconvex optimization problem with trigonometric terms in its objective function. When the structure is centrosymmetric, we show that the underlying optimization problem can be reformulated in a way that avoids the trigonometric terms. Through the introduction of a suitable set of binary variables, the objective function is rendered linear and the model is reduced to a mathematically equivalent integer linear optimization problem. Off-the-shelf optimization software of the branch-and-bound type can be utilized to solve the integer model. These algorithms require no starting point and

are guaranteed to terminate finitely with a global optimum. Comparisons with other algorithms for structure determination are provided in this paper to illustrate the relative effectiveness of our method in terms of solution time and structure quality.

## 2. Minimal principle

Consider an X-ray experiment that provides the normalized structure-factor amplitudes, $|E_m|$, for $m = 1, \ldots, M$ reflections, each of which corresponds to a reciprocal-lattice vector $\mathbf{h}_m$ and phase $\phi_m$. Hauptman & Karle (1953) introduced certain linear combinations of the phases, the *structure invariants*, whose values are independent of the choice of origin. The most important of these invariants are the triplets:

$$\omega_t = \phi_{m_t} + \phi_{m_t'} + \phi_{m_t''}, \quad t = 1, \ldots, T,$$

where $m_t$, $m_t'$ and $m_t''$ are indices corresponding to reciprocal-lattice vectors $\mathbf{h}_{m_t}$, $\mathbf{h}_{m_t'}$ and $\mathbf{h}_{m_t''}$, respectively, such that $\mathbf{h}_{m_t} + \mathbf{h}_{m_t'} + \mathbf{h}_{m_t''} = \mathbf{0}$ for all $T$ triplet invariants. Under the assumption that all $n$ atoms in the unit cell are identical, the conditional probability distribution of the triplet $\omega_t$ is given by (Cochran, 1955)

$$P(\omega_t \mid |E_{m_t}|, |E_{m_t'}|, |E_{m_t''}|) = \frac{1}{2\pi I_0(A_t)} \exp(A_t \cos\omega_t), \quad t = 1, \ldots, T, \tag{1}$$

where $A_t = (2/n^{1/2})|E_{m_t}||E_{m_t'}||E_{m_t''}|$ and $I_0$ is the modified Bessel function of order zero.

From the basic probability distribution in (1), it is readily found that the expected value of $\omega_t$ is zero. Thus, an estimate for the triplet $\omega_t$ is

$$\omega_t = \phi_{m_t} + \phi_{m_t'} + \phi_{m_t''} \approx 0, \quad t = 1, \ldots, T, \tag{2}$$

and is valid only for large values of $A_t$. Equation (2) is a milestone in traditional direct methods. However, as $n$ increases, the value of $A_t$ decreases and the estimate $\omega_t \approx 0$ is not accurate. Therefore, this estimate is not appropriate for molecules consisting of many atoms. This limitation has motivated Debaerdemaeker & Woolfson (1983) to suggest a least-squares minimal principle involving the cosine of the invariants instead of the invariants themselves. For noncentrosymmetric structures, the conditional expected value of the cosine of a triplet invariant is (Germain *et al.*, 1970)

$$\langle \cos\omega_t \rangle_t = \frac{I_1(A_t)}{I_0(A_t)} > 0, \quad t = 1, \ldots, T, \tag{3}$$

where $I_1$ and $I_0$ are the modified Bessel functions of order one and zero, respectively. When the structure possesses a center of symmetry, the conditional expected value of the cosine of the triplet invariant is (Woolfson, 1954)

$$\langle \cos\omega_t \rangle_t = \tanh(A_t/2) > 0, \quad t = 1, \ldots, T. \tag{4}$$

The minimal principle approach estimates the phases by solving a least-squares optimization problem that requires the triplet invariants to be as close as possible to the theoretical prediction in (3) or (4). The optimization problem with respect to the triplet invariants and phases can be cast as follows (Debaerdemaeker & Woolfson, 1983; Hauptman, 1988; Miller *et al.*, 1993; DeTitta *et al.*, 1994):

*Indices*

$m$ index used for reflections ($m = 1, \ldots, M$).

$t$ index used for triplet invariants ($t = 1, \ldots, T$).

*Variables*

$\phi_m$ phase of the $m$th reflection.

$\omega_t$ triplet invariant defined by $\omega_t = \phi_{m_t} + \phi_{m_t'} + \phi_{m_t''}$, where $\mathbf{h}_{m_t} + \mathbf{h}_{m_t'} + \mathbf{h}_{m_t''} = \mathbf{0}$.

*Parameters*

$M$ number of reflections.

$n$ number of atoms in the unit cell.

$T$ number of invariants.

$|E_m|$ normalized structure-factor amplitude associated with reflection $\mathbf{h}_m$.

$A_t$ constant equal to $(2/n^{1/2})|E_{m_t}||E_{m_t'}||E_{m_t''}|$.

$\overline{\omega}_t$ conditional expected value of the cosine of the triplet invariant from the right-hand side of (3) or (4).

*Model M1*

$$\min \quad f(\omega) = \frac{\sum_{t=1}^{T} A_t(\cos\omega_t - \overline{\omega}_t)^2}{\sum_{t=1}^{T} A_t} \tag{5}$$

$$\text{s.t.} \quad \omega_t = \phi_{m_t} + \phi_{m_t'} + \phi_{m_t''}, \quad t = 1, \ldots, T, \tag{6}$$
$$\phi_m \in [0, 2\pi], \quad m = 1, \ldots, M,$$
$$\omega_t \in [0, 6\pi], \quad t = 1, \ldots, T.$$

Note that M1 is a constrained optimization problem in which $f(\omega)$, the objective function in (5), is minimized subject to (s.t.) satisfying the relationships between phases and triplet invariants (6).

In order to solve M1, Bashir *et al.* (1990) have used a simulated-annealing method that did not always converge owing to the nature of the objective function. A greedy local optimization technique employed by the same authors was computationally unattractive. The parallel genetic algorithm of Chang *et al.* (1994) was successful for small structures but very time consuming.

The *Shake-and-Bake* approach (Miller *et al.*, 1993) begins with random atomic positions that result in non-negative electron density and atoms no closer than 1.2 Å. From the initial atomic positions, the values of the corresponding phases are calculated. In each cycle of the algorithm, a phase is perturbed by a prespecified amount and the function $f(\omega)$ is calculated. Then, for the set of phases corresponding to the smallest $f$ value in this cycle, a Fourier transformation, an $E$-map interpretation and a Fourier back inversion is performed to further refine the phases. The algorithm terminates when a predetermined number of trial structures has been tested and a predetermined number of phase perturbations has been completed. Convergence of this algorithm to the solution of the problem highly depends on several parameters, including accurate knowledge of the number of independent atoms in the unit cell and how the phases are perturbed (Miller *et al.*, 1994). Nonetheless, this algorithm has successfully determined many structures that involve from

**Table 1**
Model sensitivity to $N : M$ and $M : T$ ratios.

| | | Structure 4 | | Structure 5 | | Structure 6 | |
|---|---|---|---|---|---|---|---|
| $N : M$ | $M : T$ | $R$ | $t$ | $R$ | $t$ | $R$ | $t$ |
| 1 : 10 | 1 : 10 | 0.043 | 254 | 0.056 | 127 | 0.132 | 112 |
| 1 : 7 | 1 : 10 | 0.047* | 106 | 0.064* | 133 | 0.152* | 26 |
| 1 : 5 | 1 : 10 | 0.047* | 29 | 0.064* | 18 | 0.152* | 3 |
| 1 : 10 | 1 : 10 | 0.043 | 254 | 0.056 | 127 | 0.133 | 112 |
| 1 : 10 | 1 : 7 | 0.047* | 0.4 | 0.080* | 29 | 0.153 | 45 |
| 1 : 10 | 1 : 5 | 0.051* | 1 | 0.089* | 37 | – | 39 |

* Atoms wrongly positioned. – Structure not identified.

tens to more than a thousand atoms, thus demonstrating the usefulness of the minimal principle model.

The main challenge associated with solving M1 comes from the objective function (5), which may cause M1 to have a very large number of local optima. To illustrate this difficulty, 1000 local searches for M1 were performed with the commercial nonlinear optimization software *MINOS* (Murtagh & Saunders, 1995) from different randomly generated starting points for $C_{10}H_{19}ClO$, a structure with only 12 non-H atoms. This structure is centrosymmetric in the $P2_1/c$ space group. Fig. 1 presents the distribution of solutions of M1 found by *MINOS*. It is important to note that the global solution of the problem was not found and that most of the identified local solutions correspond to large objective function values. Thus, a local search algorithm will most probably fail to find a global optimum of the problem even for small structures. In the next section, we propose an approach to overcome this difficulty.

## 3. Integer programming approach

An important simplification of M1 can be achieved by observing that the phases must obtain a 0 or $\pi$ value for centrosymmetric structures. Therefore, the triplet invariants, $\omega_t$ in (6), obtain values from the set $\{0, \pi, 2\pi, 3\pi\}$. As a result, the cosines of the triplet invariants can only take values from the set $\{-1, 1\}$. To model this discrete set of acceptable solutions, we use the transformation

$$\cos\omega_t = 1 - 2\beta_t, \quad t = 1, \ldots, T,$$

where $\beta_t$ is a binary variable allowed to take values of 0 or 1. This transformation forces the cosines of the triplet invariants to take values from the set $\{-1, 1\}$ and turns the quadratic objective function into a linear one in terms of the binary variables $\beta_t$:

$$(\cos\omega_t - \overline{\omega}_t)^2 = [(1 - 2\beta_t) - \overline{\omega}_t]^2 = 4\beta_t\overline{\omega}_t + (1 + \overline{\omega}_t^2 - 2\overline{\omega}_t).$$

Simultaneously with this transformation, we will replace the phases by their values modulo $\pi$:

$$\varphi_{m_t} = \phi_{m_t}/\pi, \quad \varphi_{m'_t} = \phi_{m'_t}/\pi, \quad \varphi_{m''_t} = \phi_{m''_t}/\pi.$$

The normalized phases $\varphi_{m_t}$, $\varphi_{m'_t}$ and $\varphi_{m''_t}$ must then be binary variables since the original phases $\phi_{m_t}$, $\phi_{m'_t}$, $\phi_{m''_t}$ can only attain a 0 or $\pi$ value. We can then reformulate M1 as follows:

*Model M2*

$$\min \quad f(\beta) = \frac{\sum_{t=1}^{T} A_t[4\beta_t\overline{\omega}_t + (1 + \overline{\omega}_t^2 - 2\overline{\omega}_t)]}{\sum_{t=1}^{T} A_t} \tag{7}$$

$$\text{s.t.} \quad \varphi_{m_t} + \varphi_{m'_t} + \varphi_{m''_t} = 2\alpha_t + \beta_t, \quad t = 1, \ldots, T, \tag{8}$$

$$\varphi_m \in \{0, 1\}, \quad m = 1, \ldots, M, \tag{9}$$

$$\alpha_t, \beta_t \in \{0, 1\}, \quad t = 1, \ldots, T. \tag{10}$$

Since $\beta_t$ can only receive a 0 or 1 value, the additional binary variables $\alpha_t$ were introduced above in order to enforce the requirement that the sum of the phases in the left-hand side of (8) receives a value from the set $\{0, 1, 2, 3\}$.

A nontrivial solution of M2, *i.e.* a solution that does not have all phases equal to zero, provides the solution to the phase problem. Model M2 is a constrained linear integer programming problem. It is a much easier optimization problem to solve to global optimality than M1. In order to solve M2, we use a branch-and-bound global optimization algorithm that is based on the 'divide-and-conquer' concept. Note that any binary combination of the variables satisfying (8) provides an upper bound for the optimal value of M2. The algorithm begins by relaxing the integrality requirements (9) and (10). In particular, we allow all variables to take values over the continuous interval [0, 1]. As this relaxation enlarges the feasible space, solving the corresponding linear programming problem (LP) provides a lower bound for the optimal value of M2. If all variables were binary in the LP solution, the

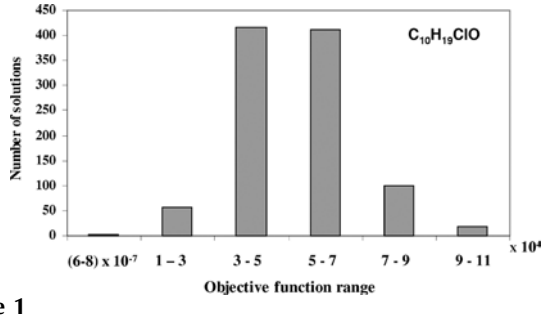


**Figure 1**
Distribution of solutions from 1000 applications of *MINOS* to M1 for $C_{10}H_{19}ClO$

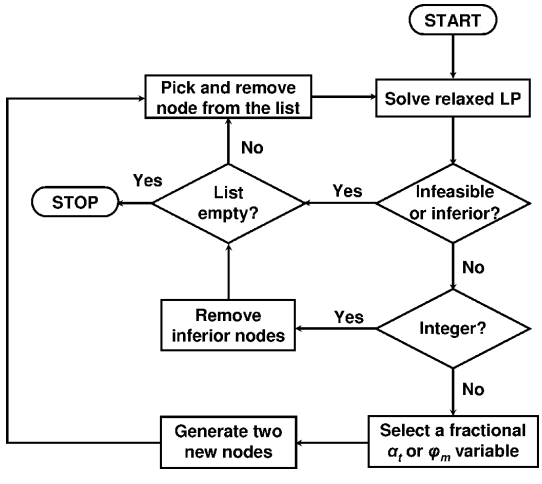


**Figure 2**
Flowchart of solution algorithm for model M2

**Table 2**
Test data sets.

| Structure | Chemical formula | $N$ | $N_I$ | $Z$ | Space group | Reference |
|---|---|---|---|---|---|---|
| 1 | $C_{50}H_{66}O_6 \cdot C_3H_7NO$ | 61 | 61 | 4 | $P2_1/c$ | Bryan & Levitskaia (2002) |
| 2 | $C_{30}H_{22}O_6S$ | 37 | 37 | 4 | $P2_1/c$ | Krishnakumar *et al.* (2002) |
| 3 | $C_{30}H_{32}N_2O_6$ | 38 | 19 | 2 | $P\bar{2}_1/c$ | Sun *et al.* (2002) |
| 4 | $C_{44}H_{38}O_4$ | 48 | 24 | 1 | $P\bar{1}$ | Vande Velde *et al.* (2002) |
| 5 | $C_{34}H_{42}B_2N_2O_4$ | 42 | 21 | 2 | $P2_1/n$ | Kliegel, Amt *et al.* (2002) |
| 6 | $C_{34}H_{26}N_2O$ | 37 | 37 | 4 | $P\bar{2}_1/c$ | Zhuang *et al.* (2002) |
| 7 | $C_5H_{12}NO^{1+} \cdot C_{28}H_{37}B_6O_{10}{}^{1-} \cdot 0.5C_4H_{10}O$ | 111 | 56 | 2 | $P\bar{1}$ | Kliegel, Drückler *et al.* (2002) |
| 8 | $3C_{40}H_{32}O_2 \cdot 4C_6H_6$ | 150 | 75 | 1 | $P\bar{1}$ | Ohba *et al.* (2002) |
| 9 | $C_{42}H_{56}N_2O_2$ | 46 | 23 | 2 | $P2_1/n$ | Lynch (2002) |
| 10 | $C_{36}H_{62}$ | 36 | 36 | 4 | $P2_1/c$ | Bragg *et al.* (2002) |
| 11 | $C_{17}H_{19}N_3O_2$ | 22 | 22 | 4 | $P2_1/n$ | Wilson (2002) |
| 12 | $C_{10}H_{19}ClO$ | 12 | 12 | 4 | $P2_1/c$ | Wilson (2002) |
| 13 | $C_{18}H_{15}NO_3$ | 22 | 22 | 4 | $P2_1/c$ | Wilson (2002) |
| 14 | $C_{13}H_{14}N_2O_3$ | 18 | 18 | 4 | $Pc$ | Wilson (2002) |
| 15 | $C_{41}H_{78}O_{11}Si_8$ | 60 | 60 | 2 | $P\bar{1}$ | Arnold & Blake (2001) |
| 16 | $C_{44}H_{52}N_4 \cdot C_2H_6O$ | 51 | 27 | 4 | $C2/c$ | Camiolo *et al.* (2001) |
| 17 | $C_{12}H_{10}O_3$ | 15 | 15 | 4 | $P2_1/n$ | Howie *et al.* (2001) |
| 18 | $C_{24}H_{12}N_6 \cdot 4CHCl_3$ | 46 | 46 | 4 | $P2_1/n$ | Alfonso & Stoeckli-Evans (2001) |

corresponding upper bound would have the same value as the lower bound and the algorithm would terminate with an optimality proof. In case some of the variables assume fractional values, one of these variables is selected and two new problems are constructed: in one of them, the selected variable is fixed to 0, while in the other this variable is fixed to 1. The search space is thus broken into smaller subsets, each of which is bound from below by solving its corresponding LP relaxation. Upper bounds are obtained when the LPs provide solutions that satisfy the integrality requirements. Lower and upper bounds are recorded and subsets are eliminated when their lower bounds are no lower than the best available upper bound. The process is repeated on all unresolved subsets and the algorithm terminates when all subsets are eliminated. As well established LP techniques are used in the context of this algorithm, this results in a fairly efficient and robust solution approach.

The entire algorithm is outlined in Fig. 2. The procedure yields a search tree with each subproblem corresponding to a node of the tree. Since the coefficients of the $\beta_t$ variables in (7) are positive, the LP relaxation will attempt to minimize all of these variables. Hence, once all $\alpha_t$ and $\varphi_m$ variables are fixed to binary values, all $\beta_t$ will naturally be 0 or 1. Thus, it suffices during the branch-and-bound algorithm to generate subsets by considering fractional values of $\alpha_t$ and $\varphi_m$ variables only.

By mere virtue of the finiteness of the search space of this integer program, branch-and-bound is a finite algorithm that *implicitly* enumerates all integer combinations of the normalized phases in order to identify a globally optimal combination. We refer the reader to Nemhauser & Wolsey (1988) for a more detailed discussion on algorithms for solving integer programming problems, including recent advances in convexification and decomposition.

### 3.1. Implementation

In our implementation, we begin by using the *LEVY* and *EVAL* programs (Blessing, 1989) to obtain the normalized structure-factor amplitudes $|E_m|, m = 1, \ldots, M$. Next, a global solution of the integer programming problem is obtained through *CPLEX7.0* (ILOG, 2000), which employs a branch-and-bound optimization strategy. This step provides the values of the phases. Then, a modified version of the *CRUNCH* system (de Gelder *et al.*, 1993) is used to calculate an $E$ map, perform the peak-picking procedure and calculate the atomic coordinates corresponding to the phases found from solving M2. All runs reported in the sequel were performed on a 1.5 GHz Dell Xeon workstation with 1 Gbyte memory.

## 4. Computational results

The purpose of this section is threefold. First, to experiment with the proposed model and identify how computational requirements and solution quality change when varying the number of phases and invariants used in the model. Second, to present computational results on the solution of a number of structures. Third, to compare the proposed computational model with existing ones in the literature.

### 4.1. Parametric analysis

Table 1 illustrates the effect of the number of reflections and triplets on solution quality and computational requirements of M2. We use reflection data for structures 4, 5 and 6 of Table 2 with 48, 42 and 37 non-H atoms ($N$), respectively. Table 1 shows that there is a trade-off between the running time ($t$, in seconds) and the quality of the solution when more phases and invariants are included in M2. Solution quality was measured in terms of crystallographic $R$. Clearly, the CPU time decreased while $R$ deteriorated when $N : M$ was increased with a constant $M : T$ ratio equal to 1 : 10. Structure quality also deteriorated when $M : T$ was increased under a constant $N : M$ equal to 1 : 10. When either of the two ratios became too large, atoms were wrongly positioned and, in one of the cases, the structure was not identified. The results of Table 1 suggest that, if the dimension of M2 must be reduced, one should reduce the number of reflections $M$ and create $10M$ triplet invariants.

**Table 3**
Model dimensions and results with integer programming approach.

| Structure | $N$ | $M$ | $T$ | Variables | $f$ | CPU s | $f'$ | CPU s |
|---|---|---|---|---|---|---|---|---|
| 1 | 61 | 610 | 6100 | 12810 | 0.0625 | 196 | 0.0342 | 402 |
| 2 | 37 | 370 | 3700 | 7770 | 0.0351 | 83 | 0.0141 | 88 |
| 3 | 38 | 380 | 3800 | 7980 | 0.0862 | 89 | 0.0592 | 103 |
| 4 | 48 | 480 | 4800 | 10080 | 0.1347 | 254 | 0.0876 | 112 |
| 5 | 42 | 420 | 4200 | 8820 | 0.0965 | 127 | 0.0672 | 144 |
| 6 | 37 | 370 | 3700 | 7770 | 0.0586 | 112 | 0.0310 | 131 |
| 7 | 55.5 | 542 | 5500 | 11550 | 0.0316 | 238 | 0.0104 | 246 |
| 8 | 150 | 1378 | 13780 | 28938 | 0.3594 | 28 | 0.3495 | 29 |
| 9 | 46 | 460 | 4600 | 9660 | 0.0872 | 130 | 0.0588 | 80 |
| 10 | 36 | 360 | 3164 | 6688 | 0.0193 | 74 | 0.0027 | 95 |
| 11 | 22 | 220 | 2200 | 4620 | 0.0053 | 63 | 0.0002 | 57 |
| 12 | 12 | 150 | 1500 | 3150 | 0.0016 | 10 | 1.0E−8 | 11 |
| 13 | 22 | 220 | 2200 | 4620 | 0.0044 | 10 | 0.0052 | 23 |
| 14 | 18 | 220 | 1600 | 3420 | 0.0243 | 16 | 0.0054 | 19 |
| 15 | 60 | 590 | 6000 | 12590 | 0.0542 | 601 | 0.0261 | 350 |
| 16 | 51 | 510 | 5100 | 10710 | 0.1258 | 261 | 0.0982 | 234 |
| 17 | 15 | 200 | 2000 | 4200 | 0.0041 | 6 | 0.0001 | 133 |
| 18 | 52 | 520 | 5200 | 10920 | 0.0351 | 134 | 0.0155 | 22 |
| Avg | | | | | | 135 | | 127 |
| Std | | | | | | 143 | | 113 |

**Table 4**
*Shake-and-Bake* results.

| | SnB1 | | SnB10 | | SnB100 | | SnB1000 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Structure | $f$ | $t$ | $f$ | $t$ | $f$ | $t$ | $f$ | $t$ | $n^*$ | $T_e$ | $k$ |
| 1 | 0.869 | 3 | 0.596 | 22 | 0.406 | 540 | 0.406 | 1740 | 70 | 45 | 56/61 |
| 2 | 0.533 | 2 | 0.223 | 8 | 0.223 | 120 | 0.223 | 960 | 10 | 13 | 30/37 |
| 3 | 1.075 | 1 | 0.522 | 4 | 0.359 | 50 | 0.359 | 240 | 17 | 3 | 19/19 |
| 4 | 0.839 | 1 | 0.451 | 3 | 0.175 | 70 | 0.175 | 180 | 43 | 7 | 24/24 |
| 5 | 1.060 | 1 | 0.266 | 2 | 0.266 | 20 | 0.266 | 200 | 6 | 3 | 21/21 |
| 6 | 0.598 | 1 | 0.598 | 6 | 0.383 | 160 | 0.383 | 600 | 40 | 12 | 35/37 |
| 7 | 0.965 | 3 | 0.645 | 16 | 0.459 | 300 | 0.254 | 1380 | 149 | 173 | 53/56 |
| 8 | 1.286 | 4 | 0.583 | 9 | 0.373 | 420 | 0.373 | 3040 | 49 | 291 | 67/75 |
| 9 | 0.852 | 1 | 0.286 | 3 | 0.286 | 33 | 0.286 | 330 | 4 | 5 | 23/23 |
| 10 | 0.782 | 2 | 0.296 | 12 | 0.296 | 360 | 0.296 | 1020 | 3 | 38 | 34/36 |
| 11 | 0.806 | 1 | 0.119 | 2 | 0.102 | 22 | 0.096 | 220 | 23 | 2 | 15/22 |
| 12 | 0.314 | 1 | 0.155 | 4 | 0.155 | 40 | 0.155 | 120 | 8 | 3 | 4/12 |
| 13 | 0.589 | 1 | 0.123 | 14 | 0.123 | 110 | 0.123 | 360 | 2 | 9 | 22/22 |
| 14 | 0.369 | 1 | 0.234 | 4 | 0.234 | 60 | 0.225 | 180 | 575 | 6 | 5/18 |
| 15 | 0.66 | 4 | 0.419 | 21 | 0.298 | 540 | 0.083 | 1920 | 305 | 960 | 18/60 |
| 16 | 0.826 | 2 | 0.729 | 8 | 0.356 | 150 | 0.356 | 780 | 56 | 29 | 24/27 |
| 17 | 0.465 | 1 | 0.208 | 2 | 0.056 | 20 | 0.056 | 189 | 56 | 1 | 15/15 |
| 18 | 0.695 | 2 | 0.408 | 12 | 0.204 | 115 | 0.204 | 1150 | 65 | 58 | 43/46 |
| Avg | | 2 | | 8 | | 174 | | 812 | 82 | 92 | 28/34 |
| Std | | 1 | | 6 | | 178 | | 795 | 143 | 229 | 17/18 |

### 4.2. Solution of a collection of structures

We have used the integer programming approach to successfully determine the 18 structures of Table 2. For each structure, this table shows the number of atoms in the chemical formula ($N$), the number of independent atoms in the unit cell ($N_I$) and the number of molecules in the unit cell ($Z$). These structures crystallize in five different space groups. Six structures include moderately heavy atoms (S, Cl, Si). All structures are centrosymmetric with the exception of structure 14 ($C_{13}H_{14}N_2O_3$) that is noncentrosymmetric but was solved in $P2_1/c$ by our algorithm owing to the pseudosymmetry of the intensity data. Atomic resolution data sets were available for these structures in the references provided in this table. Although all of the structures were previously known, this information was not used in our solution strategy. In particular, no starting point was provided to the integer programming solver.

Table 3 shows the numbers of atoms in the chemical formula ($N$), phases ($M$) and invariants ($T$) used in the integer programming model. An $N : M : T$ ratio of 1 : 10 : 100 was used for all structures with five exceptions. For structure 8, there was an insufficient number of strong reflections. For structures 10, 12, 14 and 17, the available reflection data did not provide a sufficient number of triplet invariants that could be generated. Table 3 also provides the total number of $\alpha$, $\beta$ and $\varphi$ variables in the optimization model M2. There are $2T + M$ variables and exactly $T$ linear constraints (one constraint for each triplet invariant).

All 18 structures were correctly determined by the algorithm. The values of the minimal principle and the CPU time required for each structure are provided in Table 3. Our computational experience showed that M2 results in the same values for the phases irrespective of whether $\overline{\omega}_t$ is set equal to $\tanh(A_t/2)$ or $I_1(A_t)/I_0(A_t)$. In order to enable comparison of the objective function values with the *Shake-and-Bake* approach which uses $\overline{\omega}_t = I_1(A_t)/I_0(A_t)$, we report both objective functions, denoted as $f$ when $\overline{\omega}_t = I_1(A_t)/I_0(A_t)$ and $f'$ when $\overline{\omega}_t = \tanh(A_t/2)$ was used in the computations. The value of $f'$ is always smaller than the corresponding value of $f$ since (4) is more accurate than (3) for centrosymmetric structures. The value of $f$ or $f'$ depends on the number of triplet invariants used and, therefore, atoms in the structure, as well as on the quality of the experimental data.

As is apparent from Table 3, despite the large number of integer variables in the optimization problem, the running time is very short for all structures. The average (Avg) and standard deviation (Std) of the time required for solving the structures is about 2 min per structure while most of the structures were solved within a few seconds. We also note that the two different objectives do not significantly affect the CPU time required for solution.

### 4.3. Comparisons with other approaches

The structures of Table 2 were also solved with the *Shake-and-Bake* and *CRUNCH* systems. The former system applies a stochastic search algorithm to M2 with real-space refinement, while the latter system utilizes Karle–Hauptman matrices for the solution of the phase problem.

**Table 5**
Crystallographic *R* values for IP, *Shake-and-Bake* and *CRUNCH*.

| Structure | IP | SnB1 | SnB10 | SnB100 | SnB1000 | *CRUNCH* |
|---|---|---|---|---|---|---|
| 1 | 0.09 | 0.40 | 0.33 | 0.23 | 0.23 | – |
| 2 | 0.14 | 0.41 | 0.34 | 0.34 | 0.34* | 0.22 |
| 3 | 0.05 | 0.46 | 0.38 | 0.28 | 0.28 | 0.05 |
| 4 | 0.04 | 0.42 | 0.35 | 0.21 | 0.20 | 0.05 |
| 5 | 0.06 | 0.47 | 0.18 | 0.18 | 0.18 | 0.08 |
| 6 | 0.13 | 0.35 | 0.35 | 0.21 | 0.21 | 0.15 |
| 7 | 0.06 | 0.46 | 0.39 | 0.33 | 0.19 | 0.07 |
| 8 | 0.19 | 0.42 | 0.35 | 0.22 | 0.22 | 0.20 |
| 9 | 0.08 | 0.39 | 0.16 | 0.16 | 0.16 | 0.10 |
| 10 | 0.10 | 0.37 | 0.18 | 0.18 | 0.18 | 0.10 |
| 11 | 0.06 | 0.38 | 0.38 | 0.36 | 0.36* | 0.07 |
| 12 | 0.10 | 0.41 | 0.38 | 0.38 | 0.38* | 0.10 |
| 13 | 0.05 | 0.33 | 0.15 | 0.15 | 0.15 | 0.05 |
| 14 | 0.04 | 0.36 | 0.31 | 0.31 | 0.29* | 0.04 |
| 15 | 0.11 | 0.41 | 0.35 | 0.33 | 0.24* | 0.13 |
| 16 | 0.19 | 0.42 | 0.41 | 0.30 | 0.30 | – |
| 17 | 0.14 | 0.38 | 0.33 | 0.23 | 0.22 | 0.20 |
| 18 | 0.25 | 0.41 | 0.38 | 0.25 | 0.25 | – |

* Structure not identified by *Shake-and-Bake*. – Structure not identified by *CRUNCH*.

For the *Shake-and-Bake* (*SnB*) system, we used the default parameters as described in Miller *et al.* (1994) and in the software manual for centrosymmetric structures. We used the same number of triplet invariants as in the integer programming approach. Table 4 provides computational results with *SnB*. Four runs are reported for each problem: SnB1, SnB10, SnB100 and SnB1000 with 1, 10, 100 and 1000 *SnB* trials, respectively. For each run, we report the value ($f$) of the minimal principle objective obtained by *SnB* and the CPU seconds ($t$) taken by this algorithm. The entries under $n^*$ denote the iteration of this algorithm for the solution to be in the range of the best SnB1000 minimal principle objective value for the first time. Under $T_e$, we provide the CPU time of SnB1000 divided by the number of SnB1000 trials that led to the best solution identified by the algorithm. Since *SnB* relies on a stochastic search technique that depends on randomly generated starting points, $T_e$ can be thought of as the expected CPU time for this algorithm to reach its best solution under the software default options. The average CPU requirements (Avg) of the integer programming algorithm (Table 3) are about 30% higher than the average expected time in Table 4 and approximately 30% lower than the computational requirements of SnB100. Finally, the last column of this table presents the fraction ($k$) of the total number of independent atoms that were correctly identified by *SnB*. An average of 82% of the total number of atoms were correctly identified by this algorithm.

As Table 4 indicates, in three cases, over 100 trials were required for *SnB* to reach a plateau in the objective function value. For all cases, the *SnB* minimal principle value was higher than that of the integer programming model. Use of the integer programming model results in objective function values that are on average 13 times smaller than those of SnB1000. Obviously, *SnB* is reporting local minima of the minimal principle model for all of these 18 structures. It should be noted that the default 1000 trials may not always provide the best possible solution of the *SnB* algorithm. For example, when 10000 trials were run for structure 2, the best minimal principle objective function value found was $f = 0.193$ with a corresponding *R* value of 0.30, which represent improvements of 13 and 12%, respectively, with respect to the corresponding values of SnB1000. It is also interesting to note that other settings also affect this algorithm. For instance, for structure 11, when 1.5 times the default number of invariants were used, the *R* value decreased from 0.36 to 0.23, while 20 out of the 22 atoms were identified correctly.

Table 5 shows that the integer programming (IP) approach results in a much smaller crystallographic *R* than any of the *SnB* runs, with an average improvement of 55% with respect to SnB1000. A * next to the SnB1000 entries indicates those cases in which we were unable to identify a good match between peak positions found with SnB1000 and the published structures. This happened in five of the structures.

For the *CRUNCH* system, we used the default values for all the parameters as described in the software manual. At least ten trials for each structure were run. Fifteen out of the eighteen structures were identified by CRUNCH with at most two atoms missing. The structures of $C_{50}H_{66}O_6\cdot C_3H_7NO$, $C_{44}H_{52}N_4\cdot C_2H_6O$ and $C_{24}H_{12}N_6\cdot 4CHCl_3$ were not identified by *CRUNCH*. For all structures solved by *CRUNCH*, the integer programming method results in a smaller or equal crystallographic *R* value as shown in Table 5. Structures that were not identified by *CRUNCH* are denoted by a dash (–) in this table. For 11 of the 15 structures solved by *CRUNCH*, the IP solution had a strictly smaller *R*. An average improvement of 12% was achieved over these 15 structures. The running time of *CRUNCH* was an average of 2 min per structure and a total time of 35 min for all 18 structures, which is very similar to the CPU requirements of the integer programming approach.

## 5. Conclusions

This paper develops an integer programming reformulation of the minimal principle for structure determination for centrosymmetric structures and proposes a branch-and-bound algorithm for its solution. This integer programming method provides fast and accurate results for 18 structures to which it was applied. The approach improves the crystallographic *R* an average of 55 and 12% in comparison to the *Shake-and-Bake* and *CRUNCH* systems, respectively. Furthermore, structures that were not solved with the default parameters of the latter two systems were solved with the integer programming method.

An important feature of the proposed technique is that it comes with a theoretical guarantee that it provides the global optimum of the minimal principle in a finite number of steps. Furthermore, it solves the phase problem entirely in reciprocal space with no need for iterations in real space.

The results of this paper demonstrate that the global solution of the minimal principle corresponds to the correct structure for the cases solved. It would be interesting to investigate whether this is also the case for larger and

noncentrosymmetric structures. To address this question, we are currently working to develop fast specialized combinatorial optimization algorithms for the proposed model as well as to extend this approach to noncentrosymmetric structures.

The authors thank Professor R. A. G. de Graaff for assisting in customizing the software *CRUNCH* to incorporate the phasing technique developed in this paper. We also thank Dr S. R. Wilson and the School of Chemical Sciences X-ray facility at the University of Illinois at Urbana-Champaign for supplying diffraction data. Finally, we thank C. M. L. Vande Velde, Professors R. A. G. de Graaff, G. M. Sheldrick, C. Weeks and H. Xu, as well as two anonymous referees for useful comments that helped us improve the quality of this manuscript. Partial financial support from the ExxonMobil Upstream Research Company, and the National Science Foundation under awards BES 98-73586, ECS 00-98770, and CTS 01-24751 is gratefully acknowledged.

## References

Alfonso, M. & Stoeckli-Evans, H. (2001). *Acta Cryst.* E**57**, o242–o244.
Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Altomare, A., Cascarano, G., Giacovazzo, C. & Guagliardi, A. (1993). *J. Appl. Cryst.* **26**, 343–350.
Arnold, P. L. & Blake, A. J. (2001). *Acta Cryst.* E**57**, o131–o133.
Bashir, N., Crovella, M., DeTitta, G., Han, F., Hauptman, H., Horvath, J., King, H., Langs, D., Miller, R., Sabin, T., Thuman, P. & Velmurugan, D. (1990). IEEE Proceedings of the Fifth Distributed Memory Computing Conference, pp. 513–521.
Blessing, R. H. (1989). *J. Appl. Cryst.* **22**, 396–397.
Bragg, S., Johnson, J. E. B., Graziano, G. M., Balaich, G. J. & Heimer, N. E. (2002). *Acta Cryst.* E**58**, o1010–o1012.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Bryan, J. C. & Levitskaia, T. G. (2002). *Acta Cryst.* E**58**, o240–o242.
Camiolo, S., Coles, S. J., Gale, P. A., Hursthouse, M. B. & Paver, M. A. (2001). *Acta Cryst.* E**57**, o258–o260.
Chang, C.-S., DeTitta, G., Miller, R. & Weeks, C. M. (1994). IEEE Proceedings of the Scalable High-Performance Computing Conference, pp. 796–802.
Chang, C.-S., Weeks, C. M., Miller, R. & Hauptman, H. (1997). *Acta Cryst.* A**53**, 436–444.
Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.
Debaerdemaeker, T. & Woolfson, M. M. (1983). *Acta Cryst.* A**39**, 193–196.
DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.
Gelder, R. de, de Graaff, R. A. G. & Schenk, H. (1993). *Acta Cryst.* A**49**, 287–293.
Germain, G., Main, P. & Woolfson, M. M. (1970). *Acta Cryst.* B**26**, 274–285.
Germain, G., Main, P. & Woolfson, M. M. (1971). *Acta Cryst.* A**27**, 368–376.
Germain, G. & Woolfson, M. M. (1968). *Acta Cryst.* B**24**, 91–96.
Giacovazzo, C. (1998). *Direct Phasing in Crystallography. Fundamentals and Applications.* Oxford University Press.
Gilmore, C. J. (1996). *Acta Cryst.* A**52**, 561–589.
Gull, S. F., Livesey, A. K. & Sivia, D. S. (1987). *Acta Cryst.* A**43**, 112–117.
Hauptman, H. A. (1988). Proc. Am. Crystallogr. Assoc. Meet. Abstract R4.
Hauptman, H. A. & Karle, J. (1953). *Am. Crystallogr. Soc.* Monograph 3. *Solution of the Phase Problem. I. The Centrosymmetric Crystal.* Michigan: American Crystallographic Association.
Hauptman, H. A., Xu, H., Weeks, C. M. & Miller, R. (1999). *Acta Cryst.* A**55**, 891–900.
Howie, R. A., Skakle, J. M. S. & Wardell, S. M. S. V. (2001). *Acta Cryst.* E**57**, o72–o74.
ILOG (2000). *CPLEX 7.0 User's Manual.* ILOG CPLEX Division, Incline Village, NV, USA.
Karle, J. & Hauptman, H. (1956). *Acta Cryst.* **9**, 635–651.
Kliegel, W., Amt, H., Patrick, B. O., Rettig, S. J. & Trotter, J. (2002). *Acta Cryst.* E**58**, o473–o475.
Kliegel, W., Drückler, 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.
Lynch, D. E. (2002). *Acta Cryst.* E**58**, o1025–o1027.
Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. (1993). *Science*, **259**, 1430–1433.
Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *J. Appl. Cryst.* **27**, 613–621.
Murtagh, B. A. & Saunders, M. A. (1995). *MINOS 5.5 User's Guide.* Tech. Rep. Sol 83-20R. Systems Optimization Laboratory, Department of Operations Research, Stanford University, CA, USA.
Nemhauser, G. L. & Wolsey, L. A. (1988). *Integer and Combinatorial Optimization.* Series in Discrete Mathematics and Optimization. New York: Wiley Interscience.
Ohba, S., Hiratsuka, T. & Tanaka, K. (2002). *Acta Cryst.* E**58**, o1013–o1015.
Olthof, G. J. & Schenk, H. (1982). *Acta Cryst.* A**38**, 117–122.
Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Sheldrick, G. M. (1997). *Methods Enzymol.* **276**, 319–343.
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.
Vande Velde, C. M. L., Hoefnagels, R. & Geise, H. J. (2002). *Acta Cryst.* E**58**, o454–o455.
Wilson, S. R. (2002). Personal communication.
Woolfson, M. M. (1954). *Acta Cryst.* **7**, 61–64.
Zhuang, J.-P., Zheng, Y. & Zhang, W.-Q. (2002). *Acta Cryst.* E**58**, o720–o722.