Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Recent advances in direct phasing methods for heavy-atom substructure determination

Hongliang Xu* and Herbert A. Hauptman

Hauptman–Woodward Medical Research Institute and Department of Structural Biology, School of Medicine and Biomedical Sciences, State University of New York at Buffalo, 700 Ellicott Street, Buffalo, NY 14203, USA

Correspondence e-mail: xu@hwi.buffalo.edu

Received 26 January 2006
Accepted 7 April 2006

Macromolecular crystal structure determination has typically been a two-step process. When diffraction data from multiple chemically isomorphous or anomalously scattering crystals are available, the positions of heavy atoms from amplitude differences arising from native–derivative crystal pairs or an anomalously scattering crystal are first located and phasing of the whole protein structure is then completed using the heavy-atom substructure as a bootstrap. *Shake-and-Bake*, a direct-methods-based dual-space refinement procedure, provides heavy-atom substructure solutions by finding the constrained global minimum of a probabilistically defined minimal function. This minimal function relies on probabilistic estimates of the cosines of the structure invariants. A novel statistically defined minimal function that utilizes the statistical properties of the structure invariants has recently been proposed and tested. Applications of the statistical *Shake-and-Bake* procedure show that statistical direct methods provide a simple, reliable and efficient method of heavy-atom substructure determination.

1. Introduction

The phase problem of X-ray crystallography is defined as the problem of determining the phases, φ, of the structure factors, $F = |F|\exp(i\varphi)$, from measurements of intensities alone. The phase information, which is lost in the diffraction experiment, is in fact recoverable from the measurable intensities. The methods devised to achieve this goal are known as direct methods, a class of *ab initio* methods in which probabilistic phase relations are used to derive reflection phases. Examples of successful direct methods include the tangent formula (Karle & Hauptman, 1956), the minimal principle (Debaerdemaeker & Woolfson, 1983; DeTitta *et al.*, 1994) and the maximum-entropy (Bricogne, 1984) and the minimal charge (Elser, 1999) methods.

The beginnings of the mathematical direct-methods theory of crystallographic phasing were the discoveries of three fundamental relationships: (i) the discovery by Harker & Kasper (1948) that owing to the non-negativity of the electron-density distribution in a crystal there are inequality relationships among X-ray structure factors of centrosymmetric crystals, (ii) the discovery by Karle & Hauptman (1950) that the necessary and sufficient condition for non-negative electron density is a determinantal structure-factor relationship from which all the inequalities derive, including those of Harker and Kasper, and (iii) the discovery by Sayre (1952) of the Fourier transform convolution relationship among triplets of structure factors, F_H, F_{-K} and F_{K-H}, for crystals composed of resolved equal atoms. Tremendous progress in the field of direct methods has been made over the last half century.

Direct methods, as implemented in widely used highly automated computer programs such as *MULTAN* (Main *et al.*, 1980), *SAYTAN* (Debaerdemaeker *et al.*, 1985), *SIR* (Burla *et al.*, 1989) and *SHELXS* (Sheldrick, 1990), provide computationally efficient solutions for structures containing fewer than 100 independent non-H atoms.

The development of a dual-space recycling procedure known as *Shake-and-Bake* (Miller *et al.*, 1993; DeTitta *et al.*, 1994; Weeks *et al.*, 1994) has dramatically increased the size of structures solvable by direct methods. *Shake-and-Bake,* the first algorithm to find the constrained global minimum of a probabilistically defined minimal function, alternates phase refinement in reciprocal space with density modification in real space to impose constraints through a physically meaningful interpretation of the electron-density function. Benchmark achievements of *Shake-and-Bake* include *ab initio* phasing of atomic resolution X-ray data from triclinic crystals of hen egg-white lysozyme, a protein composed of 1001 independent non-H atoms with no atom heavier than sulfur (Deacon *et al.*, 1998), and *ab initio* determination from single-wavelength anomalous dispersion X-ray data of the 160-atom selenium substructure in SeMet *Escherichia coli* ketopantoate hydroxymethyl transferase crystals, which contain two decamers of 26 kDa monomers in their asymmetric unit (von Delft *et al.*, 2003). These two benchmarks demonstrate that direct methods are useful for larger molecules (more than 250 independent non-H atoms) and unique to the macromolecular field when combined with anomalous dispersion measurements or multiple diffraction patterns that include single isomorphous replacement (SIR), single-wavelength anomalous scattering (SAS) and multi-wavelength anomalous dispersion (MAD). The widely used *Shake-and-Bake* algorithm has been implemented in *SnB* (Miller *et al.*, 1994; Weeks & Miller, 1999) and *BnP* (Weeks *et al.*, 2002) software and adapted in the *SHELXC*/*D*/*E* (Sheldrick, 2006) and *PHENIX* (Adams, 2006) software.

2. Probabilistic approach to the crystallographic phase problem

For a given reciprocal-lattice vector **H**, the normalized structure factor E_H is defined by

$$E_H = |E_H| \exp(i\varphi_H) = F_H / \langle |F_H|^2 \rangle^{1/2} = \sigma_2^{-1/2} \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{H} \cdot \mathbf{r}_j), \quad (1)$$

where N is the number of atoms in the unit cell, f_j and $\mathbf{r}_j$ are the scattering factor and the position vector of the jth atom and $\sigma_2 = \sum_{j=1}^{N} f_j^2$. Certain linear combinations of the phases, the structure invariants, are uniquely determined by the structure and are independent of the choice of the origin (Hauptman & Karle, 1953). The most important of these invariants are the triplets

$$\varphi_{HK} = \varphi_H + \varphi_K + \varphi_{-H-K}, \quad (2)$$

along with their associated parameters A_{HK}, defined in the equal-atom case by

$$A_{HK} = 2N^{-1/2} |E_H E_K E_{H+K}|. \quad (3)$$

The probabilistic approach to the phase problem assumes that the atomic position vectors **r** of the atoms in a crystal are random variables that are uniformly and independently distributed in the unit cell. Based on this simple assumption, the modern probabilistic theory provides the machinery to derive the conditional probability distribution of the structure invariants, given well defined sets of measured intensities,

$$P(\varphi | A_{HK}) = [2\pi I_0(A_{HK})]^{-1} \exp(A_{HK} \cos\varphi), \quad (4)$$

where I_0 is the modified Bessel function of zeroth order. From (4), one immediately obtains the estimate

$$\varphi_{HK} = \varphi_H + \varphi_K + \varphi_{-H-K} \simeq 0, \quad (5)$$

as long as the values of A_{HK} are large.

2.1. The tangent formula

The first application of the probabilistic approach to the phase problem is the tangent formula (Karle & Hauptman, 1956),

$$\tan(\varphi_H) = \frac{\sum_K W_{HK} \sin(\varphi_K + \varphi_{H-K})}{\sum_K W_{HK} \cos(\varphi_K + \varphi_{H-K})}, \quad (6)$$

where W_{HK} are appropriate weights (*e.g.* $W_{HK} = |E_K E_{H-K}|$ or A_{HK}). The tangent formula, together with its modified forms, represents the earliest development of the probabilistic approach to the phase problem and demonstrates the power of probabilistic methods on which the direct methods of phase determination are primarily based.

2.2. The minimal principle

From (4), one also obtains the conditional expected value of $\cos\varphi$ given A_{HK},

$$\langle \cos\varphi | A_{HK} \rangle = I_1(A_{HK}) / I_0(A_{HK}), \quad (7)$$

where I_1/I_0 is the ratio of the modified Bessel functions of the first and zeroth order (Cochran, 1955). The crystallographic phase problem can be formulated as a problem in constrained global minimization. The commonly used cosine minimal function (DeTitta *et al.*, 1994),

$$R(\varphi) = \left(\sum_{H,K} A_{HK} \right)^{-1} \sum_{H,K} A_{HK} \left[\cos(\varphi_{HK}) - \frac{I_1(A_{HK})}{I_0(A_{HK})} \right]^2, \quad (8)$$

measures the least-squares difference between the current values of the cosine structure invariants, $\cos\varphi_{HK}$, and their conditional expected values. It is expected that the minimal function (8) reaches its constrained global minimum when all the phases are equal to their true values for any choice of origin and enantiomorph (the minimal principle).

The successful application of the probabilistic approach to the crystallographic phase problem depends on the radius of convergence of the minimal function and requires a sufficient

Table 1
Percentage of structure invariants having non-negative $\cos(\varphi_{HK})$ values for the structures Iled and Trilys.

Data were truncated to various resolutions.

Structure	Iled	Trilys
Atoms	84	1001
Reflections	840	10010
Invariants	8400	100100
1.0 Å	78.0%	67.1%
1.1 Å	74.3%	65.1%
1.2 Å	70.5%	62.4%
1.3 Å	68.0%	60.9%
1.4 Å	66.7%	59.7%
1.5 Å	65.0%	58.9%
1.6 Å	62.8%	58.0%

number of reliable estimates of the cosine values of the structure invariants, *i.e.* a sufficient number of triplets having large $A_{HK} = 2N^{-1/2}|E_H E_K E_{H+K}|$ values. Unfortunately, the average value of A_{HK} decreases as the number of atoms (N) increases or the data resolution is reduced (*i.e.* fewer reflections having $|E| > 1$). Table 1 shows the percentage of structure invariants having non-negative $\cos(\varphi_{HK})$ values for an 84-atom structure, Iled (Pletnev *et al.*, 1980), and a 1001-atom structure, Trilys (Deacon *et al.*, 1998), using data truncated from an original resolution of 0.94 Å for Iled and 0.85 Å for Trilys to various resolutions of 1.0, 1.1, ..., 1.6 Å. For each resolution limit, $10N$ reflections having the largest $|E|$ values are selected to generate $100N$ structure invariants having the largest A values. Table 1 shows that (i) the percentage drops monotonically from 78.0 to 62.8% for Iled and from 67.1 to 58.0% for Trilys when data resolution is reduced from 1.0 to 1.6 Å and (ii) the percentages for Trilys (large structure) are significantly lower than those for Iled (small structure). Owing to the fact that the conditional expected values of $\cos(\varphi_{HK})$ are always positive, the large error involved in the minimal function (8) may severely reduce the radius of convergence and increase the difficulty of reaching the constrained global minimum. When errors produced by unreliable estimates for $\cos(\varphi_{HK})$ reach a certain value, the minimal principle may no longer be valid, and even if it is, the radius of convergence of the *Shake-and-Bake* algorithm may be so small that the method fails in practice. This may explain why direct methods are limited with respect to structural size and data resolution.

3. Statistical approach to the crystallographic phase problem

We proposed a novel statistical approach to the crystallographic phase problem (Xu & Hauptman, 2004). In our approach, the statistical properties of the structure invariants were used to generate a statistically based minimal function. The statistical properties were observed from a variety of known structures consisting of centrosymmetric, non-centrosymmetric structures and Se-atom substructures with different sizes, resolutions and space groups. From distributions of triplets having true and random invariant values, we observed that the triplet distribution of the true invariant values was significantly higher than that of the random invariant values over a (statistical) interval $I = [-r, r]$, where $r < \pi$.

3.1. Statistical minimal function

These statistical properties motivated us to define a statistically based minimal function

$$m(\varphi) = 1 - \int_{-r}^{r} D(\varphi)\,\mathrm{d}\varphi = 1 - N_I/N_T, \qquad (9)$$

where $r < \pi$ is chosen arbitrarily, $D(\varphi)$ is the triplet distribution on $[-\pi, \pi]$, N_T is the total number of triplets and N_I is the number of triplets whose values lie within $I = [-r, r]$. Note that the value of N_I depends on the values of all selected phases. Thus, when an individual phase value changes, the values of all triplets associated with this phase will change and therefore the value of N_I will also change. It is obvious that the values of the statistical minimal function depend on the choice of the interval $I = [-r, r]$. It has been confirmed experimentally that with a proper choice of the statistical interval, the statistical minimal function reaches its constrained global minimum when all phases are equal to their true values for any choice of origin and enantiomorph (the statistical minimal principle).

3.2. Statistical *Shake-and-Bake*

It is one thing to formulate the phase problem as a problem of constrained global minimization; it is quite another to actually find the constrained global minimum. *Shake-and-Bake* (Miller *et al.*, 1993; DeTitta *et al.*, 1994; Weeks *et al.*, 1994), the most powerful direct-methods-based procedure yet devised, is the first algorithm to find the constrained global minimum of a probabilistically defined minimal function. The *Shake-and-Bake* procedure starts from random atomic structures and iterates cycles that alternate phase refinement in reciprocal space, based on the technique of parameter shift (Bhuiya & Stanley, 1963) to reduce the value of the minimal function, with density modification by atomic peak picking in real space. The *Shake-and-Bake* algorithm has been implemented in the computer program *SnB* (Weeks & Miller, 1999). Statistical *Shake-and-Bake* is a modification of *Shake-and-Bake* obtained by replacing the cosine minimal function (8) by the statistical minimal function (9). Statistical *Shake-and-Bake* has been implemented in the computer program *S-SnB*.

3.3. Applications for substructure determination

As successful direct-methods applications have utilized anomalous dispersion measurements or multiple diffraction patterns (SIR, SAS and MAD) to determine heavy-atom substructures, 19 known SeMet substructures ranging in size from five to 70 Se sites in the asymmetric unit were used as test structures (Xu *et al.*, 2005). Both *SnB* and *S-SnB* programs were executed using difference data from the test structures to obtain success rates (percentage of trial structures that converge to the solution). The resultant success rates were reported in the form $x \pm \sigma(x)$, where $\sigma(x)$ is the standard deviation of x, and were used for comparison and for optimization of the statistical interval.

Based on the test results from *S-SnB* (Xu *et al.*, 2005), we suggested two strategies for selecting the default statistical interval: (i) the conservative interval $I = [-90, 90^\circ]$ and (ii) the aggressive interval $I = [-r, r]$, with $r = 9.14\ln(N) + 55.3^\circ$, where N is the number of Se atoms in the asymmetric unit. The advantage of using a conservative interval is that one fixed interval can be used to determine substructures of any size with reasonably high success rates. The disadvantage is the loss of higher success rates for small and medium substructures (5–35 Se atoms). On the other hand, the advantage of using an aggressive interval is the potential for yielding maximal success rates for small structures. The results also suggested that a large interval, $I = [-r, r]$ with $r > 90^\circ$, may be needed to determine vary large substructures ($\geq$100).

When comparing two success rates (x and y) obtained from two different procedures, y is statistically higher than x if $y \geq x + 2\sigma(|y - x|)$, where $\sigma(|y - x|) = [\sigma^2(x) + \sigma^2(y)]^{1/2}$; y is statistically lower than x if $y \leq x - 2\sigma(|y - x|)$; otherwise y is statistically equivalent to x. When compared with *SnB* using 38 dispersive or anomalous difference data sets from 19 test structures, *S-SnB* with the conservative interval yielded 11 statistically higher and only one statistically lower success rates, while *S-SnB* with the aggresive interval yielded 19 statistically higher and only one statistically lower success rates.

4. Conclusion

The results described above confirm that statistical *Shake-and-Bake* is more powerful than traditional *Shake-and-Bake* for the determination of Se-atom substructures. Consequently, the statistical *Shake-and-Bake* procedure has been implemented as the default method in the latest versions of the computer programs *SnB* and *BnP*. These programs can be downloaded from the websites http://www.hwi.buffalo.edu/SnB/ or http://www.hwi.buffalo.edu/BnP/, respectively.

Research support from NIH grants EB002057 and GM-72023 is gratefully acknowledged.

References

Adams, P. D. (2006). *PHENIX*. http://phenix-online.org/.
Bhuiya, A. K. & Stanley, E. (1963). *Acta Cryst.* **16**, 981–984.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Burla, M. C., Camalli, M., Cascarano, G., Giacovazzo, C., Polidori, G., Spagna, R. & Viterbo, D. (1989). *J. Appl. Cryst.* **22**, 389–393.
Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.
Deacon, A. M., Weeks, C. M., Miller, R. & Ealick, S. E. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 9284–9289.
Debaerdemaeker, T., Tate, C. & Woolfson, M. M. (1985). *Acta Cryst.* A**41**, 286–290.
Debaerdemaeker, T. & Woolfson, M. M. (1983). *Acta Cryst.* A**39**, 193–196.
Delft, F. von, Inoue, T., Saldanha, S. A., Ottenhof, H. H., Schmitzberger, F., Birch, L. M., Dhanaraj, V., Witty, M., Smith, A. G., Blundell, T. L. & Abell, C. (2003). *Structure*, **11**, 985–996.
DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.
Elser, V. (1999). *Acta Cryst.* A**55**, 489–499.
Harker, D. & Kasper, J. S. (1948). *Acta Cryst.* **1**, 70–75.
Hauptman, H. A. & Karle, J. (1953). *Solution of the Phase Problem. I. The Centrosymmetric Crystal*. ACA Monograph 3. Michigan: American Crystallographic Association.
Karle, J. & Hauptman, H. A. (1950). *Acta Cryst.* **3**, 181–187.
Karle, J. & Hauptman, H. A. (1956). *Acta Cryst.* **9**, 635–651.
Main, P., Fiske, S. J., Hull, S. E., Lessinger, L., Germain, G., Declercq, J.-P. & Woolfson, M. M. (1980). *MULTAN*80. Universities of York, England and Louvain, Belgium.
Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.
Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *J. Appl. Cryst.* **27**, 613–621.
Pletnev, V. Z., Galitskii, N. M., Smith, G. D., Weeks, C. M. & Duax, W. L. (1980). *Biopolymers*, **19**, 1517–1534.
Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.
Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Sheldrick, G. M. (2006). *SHELX*. http://shelx.uni-ac.gwdg.de/SHELX/index.html.
Weeks, C. M., Blessing, R. H., Miller, R., Mungee, R., Potter, S. A., Rappleye, J., Smith, G. D., Xu, H. & Furey, W. (2002). *Z. Kristallogr.* **217**, 686–693.
Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210-220.
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. & Hauptman, H. A. (2005). *Acta Cryst.* D**61**, 976–981.