16.1. *AB INITIO* PHASING

process is to sort the reflections in decreasing order according to their $|E|$ values and to choose the number of large $|E|$ reflections that are to be phased. The second step is to generate the possible invariants involving these intense reflections and then to sort them in decreasing order according to their A_{HK} values. Those invariants with the largest A_{HK} values are retained in sufficient number to achieve the desired overdetermination. *Ab initio* phase determination by direct methods requires not only a set of invariants, the average values of the cosines of which are presumed to be known, but also a set of starting phases. Therefore, the third step in the phasing process is the assignment of initial phase values. If enough pairs of phases, φ_K and φ_{-H-K}, are known, the structure invariants can then be used to generate further phases (φ_H) which, in turn, can be used to evaluate still more phases. Repeated iterations will permit most reflections with large $|E_H|$ to be phased.

Depending on the space group, a small number of phases can be assigned arbitrarily in order to fix the origin position and, in noncentrosymmetric space groups, the enantiomorph. However, except for the simplest structures, these reflections provide an inadequate foundation for further phase development. Consequently, a 'multisolution' or multi-trial approach (Germain & Woolfson, 1968) is normally taken in which other reflections are each assigned many different starting values in the hope that one or more of the resultant phase combinations will lead to a solution. Solutions, if they occur, must be identified on the basis of some suitable figure of merit. Although phases can be evaluated sequentially, the order determined by a so-called convergence map (Germain *et al.*, 1970), it has become standard in recent years to use a random-number generator to assign initial values to all available phases from the outset (Baggio *et al.*, 1978; Yao, 1981). A variant of this procedure is to use the random-number generator to assign initial coordinates to the atoms in the trial structures and then to obtain initial phases from a structure-factor calculation.

16.1.4. Reciprocal-space phase refinement or expansion (*shaking*)

Once a set of initial phases has been chosen, it must be refined against the set of structure invariants whose values are presumed known. In theory, any of a variety of optimization methods could be used to extract phase information in this way. However, so far only two (tangent refinement and parameter-shift optimization of the minimal function) have been shown to be of practical value.

16.1.4.1. *The tangent formula*

The *tangent formula*,

$$\tan(\varphi_H) = \frac{-\sum_K |E_K E_{-H-K}| \sin(\varphi_K + \varphi_{-H-K})}{\sum_K |E_K E_{-H-K}| \cos(\varphi_K + \varphi_{-H-K})}, \quad (16.1.4.1)$$

(Karle & Hauptman, 1956), is the relationship used in conventional direct-methods programs to compute φ_H given a sufficient number of pairs ($\varphi_K, \varphi_{-H-K}$) of known phases. It can also be used within the phase-refinement portion of the dual-space *Shake-and-Bake* procedure (Weeks, Hauptman *et al.*, 1994; Sheldrick & Gould, 1995). The variance associated with φ_H depends on $\sum_K E_H E_K E_{-H-K}/N^{1/2}$ and, in practice, the estimate is only reliable for $|E_H| \gg 1$ and for structures with a limited number of atoms (N). If equation (16.1.4.1) is used to redetermine previously known phases, the phasing process is referred to as *tangent-formula refinement*; if only new phases are determined, the phasing process is *tangent expansion*.

The tangent formula can be derived using the assumption of equal resolved atoms. Nevertheless, it suffers from the disadvantage that, in space groups without translational symmetry, it is perfectly

fulfilled by a false solution with all phases equal to zero, thereby giving rise to the so-called 'uranium-atom' solution with one dominant peak in the corresponding Fourier synthesis. In conventional direct-methods programs, the tangent formula is often modified in various ways to include (explicitly or implicitly) information from the so-called 'negative' quartet invariants (Schenk, 1974; Hauptman, 1974; Giacovazzo, 1976) that are dependent on the smallest as well as the largest E magnitudes. Such modified tangent formulas do indeed largely overcome the problem of pseudosymmetric solutions for small N, but because of the dependence of quartet-term probabilities on $1/N$, they are little more effective than the normal tangent formula for large N.

16.1.4.2. *The minimal function*

Constrained minimization of an objective function like the *minimal function*,

$$R(\Phi) = \sum_{H, K} A_{HK} \{\cos \Phi_{HK} - [I_1(A_{HK})/I_0(A_{HK})]\}^2 \bigg/ \sum_{H, K} A_{HK}$$

$$(16.1.4.2)$$

(Debaerdemaeker & Woolfson, 1983; Hauptman, 1991; DeTitta *et al.*, 1994), provides an alternative approach to phase refinement or phase expansion. $R(\Phi)$ is a measure of the mean-square difference between the values of the triplets calculated using a particular set of phases and the expected values of the same triplets as given by the ratio of modified Bessel functions. The minimal function is expected to have a constrained global minimum when the phases are equal to their correct values for some choice of origin and enantiomorph (the minimal principle). Experimentation has thus far confirmed that, when the minimal function is used actively in the phasing process and solutions are produced, the final trial structure corresponding to the smallest value of $R(\Phi)$ is a solution provided that $R(\Phi)$ is calculated directly from the atomic positions before the phase-refinement step (Weeks, DeTitta *et al.*, 1994). Therefore, $R(\Phi)$ is also an extremely useful figure of merit. The minimal function can also include contributions from higher-order (*e.g.* quartet) invariants, although their use is not as imperative as with the tangent formula because the minimal function does not have a minimum when all phases are zero. In practice, quartets are rarely used in the minimal function because they increase the CPU time while adding little useful information for large structures. The cosine function in equation (16.1.4.2) can also be replaced by other functions of the phases giving rise to alternative minimal functions. In particular, an exponential expression has been found to give superior results for several *P*1 structures (Hauptman *et al.*, 1999).

16.1.4.3. *Parameter shift*

In principle, any minimization technique could be used to minimize $R(\Phi)$ by varying the phases. So far, a seemingly simple algorithm, known as parameter shift (Bhuiya & Stanley, 1963), has proven to be quite powerful and efficient as an optimization method when used within the *Shake-and-Bake* context to reduce the value of the minimal function. For example, a typical phase-refinement stage consists of three iterations or scans through the reflection list, with each phase being shifted a maximum of two times by 90° in either the positive or negative direction during each iteration. The refined value for each phase is selected, in turn, through a process which involves evaluating the minimal function using the original phase and each of its shifted values (Weeks, DeTitta *et al.*, 1994). The phase value that results in the lowest minimal-function value is chosen at each step. Refined phases are used immediately in the subsequent refinement of other phases. It should be noted that the parameter-shift routine is similar to that used in ψ-map refinement

335

(White & Woolfson, 1975) and *XMY* (Debaerdemaeker & Woolfson, 1989).

16.1.5. Real-space constraints (*baking*)

Peak picking is a simple but powerful way of imposing an atomicity constraint. The potential for real-space phase improvement in the context of small-molecule direct methods was recognized by Jerome Karle (1968). He found that even a relatively small, chemically sensible, fragment extracted by manual interpretation of an electron-density map could be expanded into a complete solution by transformation back to reciprocal space and then performing additional iterations of phase refinement with the tangent formula. Automatic real-space electron-density map interpretation in the *Shake-and-Bake* procedure consists of selecting an appropriate number of the largest peaks in each cycle to be used as an updated trial structure without regard to chemical constraints other than a minimum allowed distance between atoms. If markedly unequal atoms are present, appropriate numbers of peaks (atoms) can be weighted by the proper atomic numbers during transformation back to reciprocal space in a subsequent structure-factor calculation. Thus, *a priori* knowledge concerning the chemical composition of the crystal is utilized, but no knowledge of constitution is required or used during peak selection. It is useful to think of peak picking in this context as simply an extreme form of density modification appropriate when atomic resolution data are available. In theory, under appropriate conditions it should be possible to substitute alternative density-modification procedures such as low-density elimination (Shiono & Woolfson, 1992; Refaat & Woolfson, 1993) or solvent flattening (Wang, 1985), but no practical applications of such procedures have yet been made. The imposition of physical constraints counteracts the tendency of phase refinement to propagate errors or produce overly consistent phase sets. Several variants of peak picking, which are discussed below, have been successfully employed within the framework of *Shake-and-Bake*.

16.1.5.1. *Simple peak picking*

In its simplest form, peak picking consists of simply selecting the top N_u E-map peaks where N_u is the number of unique non-H atoms in the asymmetric unit. This is adequate for true small-molecule structures. It has also been shown to work well for heavy-atom or anomalously scattering substructures where N_u is taken to be the number of expected substructure atoms (Smith *et al.*, 1998; Turner *et al.*, 1998). For larger structures ($N_u > 100$), it is likely to be better to select about $0.8N_u$ peaks, thereby taking into account the probable presence of some atoms that, owing to high thermal motion or disorder, will not be visible during the early stages of a structure determination. Furthermore, a recent study (Weeks & Miller, 1999*b*) has shown that structures in the 250–1000-atom range which contain a half dozen or more moderately heavy atoms (*i.e.*, S, Cl, Fe) are more easily solved if only $0.4N_u$ peaks are selected. The only chemical information used at this stage is a minimum inter-peak distance, generally taken to be 1.0 Å. For substructure applications, a larger minimum distance (*e.g.* 3 Å) is more appropriate.

16.1.5.2. *Iterative peaklist optimization*

An alternative approach to peak picking is to select approximately N_u peaks as potential atoms and then eliminate some of them, one by one, while maximizing a suitable figure of merit such as

$$P = \sum_{\mathbf{H}} |E_c^2|(|E_o^2| - 1). \tag{16.1.5.1}$$

The top N_u peaks are used as potential atoms to compute $|E_c|$. The atom that leaves the highest value of P is then eliminated. Typically, this procedure, which has been termed *iterative peaklist optimization* (Sheldrick & Gould, 1995), is repeated until only $2N_u/3$ atoms remain. Use of equation (16.1.5.1) may be regarded as a reciprocal-space method of maximizing the fit to the origin-removed sharpened Patterson function, and it is used for this purpose in molecular replacement (Beurskens, 1981). Subject to various approximations, maximum-likelihood considerations also indicate that it is an appropriate function to maximize (Bricogne, 1998). Iterative peaklist optimization provides a higher percentage of solutions than simple peak picking, but it suffers from the disadvantage of requiring much more CPU time.

16.1.5.3. *Random omit maps*

A third peak-picking strategy also involves selecting approximately N_u of the top peaks and eliminating some, but, in this case, the deleted peaks are chosen at random. Typically, one-third of the potential atoms are removed, and the remaining atoms are used to compute E_c. By analogy to the common practice in macromolecular crystallography of omitting part of a structure from a Fourier calculation in hopes of finding an improved position for the deleted fragment, this version of peak picking is described as making a *random omit map*. This procedure is a little faster than simply picking N_u atoms because fewer atoms are used in the structure-factor calculation. More important is the fact that, like iterative peaklist optimization, it has the potential for being a more efficient search algorithm.

16.1.6. Fourier refinement (*twice baking*)

E-map recycling, but without phase refinement (Sheldrick, 1982, 1990; Kinneging & de Graaff, 1984), has been frequently used in conventional direct-methods programs to improve the completeness of the solutions after phase refinement. It is important to apply Fourier refinement to *Shake-and-Bake* solutions also because such processing significantly increases the number of resolved atoms, thereby making the job of map interpretation much easier. Since phase refinement *via* either the tangent formula or the minimal function requires relatively accurate invariants that can only be generated using the larger *E* magnitudes, a limited number of reflections are phased during the actual dual-space cycles. Working with a limited amount of data has the added advantage that less CPU time is required. However, if the current trial structure is the 'best' so far based on a figure of merit (either the minimal function or a real-space criterion), then it makes sense to subject this structure to Fourier refinement using additional data, thereby reducing series-termination errors. The correlation coefficient

$$\begin{aligned}
\text{CC} = &\left[\left(\sum wE_o^2E_c^2 \sum w\right) - \left(\sum wE_o^2 \sum wE_c^2\right)\right] \\
&\times \left\{\left[\left(\sum wE_o^4 \sum w\right) - \left(\sum wE_o^2\right)^2\right]\right. \\
&\times \left.\left[\left(\sum wE_c^4 \sum w\right) - \left(\sum wE_c^2\right)^2\right]\right\}^{-1/2}
\end{aligned} \tag{16.1.6.1}$$

(Fujinaga & Read, 1987), where weights $w = 1/[0.04 + \sigma^2(E_o)]$, has been found to be an especially effective figure of merit when used with all the data and is, therefore, suited for identifying the most promising trial structure at the end of Fourier refinement. Either simple peak picking or iterative peaklist optimization can be employed during the Fourier-refinement cycles in conjunction with weighted *E* maps (Sim, 1959). The final model can be further improved by isotropic displacement parameter (B_{iso}) refinement for the individual atoms (Usón *et al.*, 1999) followed by calculation of the Sim (1959) or sigma-A (Read, 1986) weighted map. This is