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 and so is less effective than the random-omit method described in the next section.

16.1.5.3. Random omit maps

A third peak-picking strategy involves selecting approximately $1.3N_u$ 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 the hope of finding an improved position for the deleted fragment, this version of peak picking is described as *random omit*. This procedure helps to prevent the dual-space recycling from getting stuck in a local minimum and is thus an efficient search algorithm.

16.1.6. Fourier refinement

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

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

(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 ($B_{\rm iso}$) 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 particularly useful when the requirement of atomic resolution is barely fulfilled, and it makes it easier to interpret the resulting maps by classical macromolecular methods.

16.1.7. Resolution enhancement: the 'free lunch' algorithm

Direct methods take a set of phases, refine them and also determine new ones. There is no reason, however, why they cannot be used to predict new amplitudes as well. If density modification of a real-space map is performed, then any process of real-space density modification will, following a Fourier transformation, give structure-factor amplitudes for reflections that were not used to generate it, and these can be outside the resolution limit. If direct methods require atomic resolution data, can we use these techniques to extrapolate structure factors (*i.e.* predict not only phases, but also amplitudes for missing data) and extend data resolution? The idea is not new, but it has been quite extensively studied in recent years. Sheldrick has termed such algorithms 'free lunch', with reference to the saying: 'There is no such thing as a free lunch'! In one example (Usón *et al.*, 2007), weak SIRAS starting phase information followed by density modification led to an $|F_o|$ weighted mean phase error (MPE) of 54° at 1.98 Å resolution, but when the density modification was performed with amplitude extrapolation to 1.0 Å, the MPE fell to 17°. Caliandro *et al.* (2005*a*,*b*) used Patterson or direct methods to obtain trial phases that are submitted to various density-modification methods. Following this, extrapolated phases were generated. This too transformed uninterpretable maps into a solution amenable to automatic tracing. Palatinus *et al.* (2007) used maximum entropy (ME) methods for amplitude extrapolation. In some ways these should be ideal for this purpose, and it is worth noting that ME maps have, *de facto*, optimal resolution enhancement built in, although they can be difficult to generate for large structures.

Why does this work, and why is it sometimes so spectacular? The answer probably lies with the fact that maps are much more sensitive to phases than amplitudes and, if the model bias of predicting new amplitudes is not too great, then using a nonzero value is better than zero, which is the default. Fourier-truncation errors may also be reduced, resulting in less spurious map detail.

16.1.8. Utilizing Pattersons for better starts

When slightly heavier atoms such as sulfur are present, it is possible to start recycling procedures from a set of atomic positions that are consistent with the Patterson function. For large structures, the vectors between such atoms will correspond to Patterson densities around or even below the noise level, so classical methods of locating the positions of these atoms unambiguously from the Patterson are unlikely to succeed. Nevertheless, the Patterson function can still be used to filter sets of starting atoms. This filter is currently implemented as follows in *SHELXD*. First, a sharpened Patterson function (Sheldrick *et al.*, 1993) is calculated, and the top 200 (for example) non-Harker peaks further than a given minimum distance from the origin are selected, in turn, as two-atom translation-search fragments, one such fragment being employed per solution attempt. For each of a large number of random translations, all unique Patterson vectors involving the two atoms and their symmetry equivalents are found and sorted in order of increasing Patterson density. The sum of the smallest third of these values is used as a figure of merit (PMF). Tests showed that although the globally highest PMF for a given two-atom search fragment may not correspond to correct atomic positions, nevertheless, by limiting the number of trials, some correct solutions may still be found. The two-atom vectors are chosen by biased random sampling that favours the vectors corresponding to higher Patterson values. The two atoms

Table 16.1.8.1
Overall success rates for full structure solution for hirustasin using different two-atom search vectors chosen from the Patterson peak list

Resolution (Å)	Two-atom search fragments	Solutions per 1000 attempts
1.2	Top 100 general Patterson peaks	86
1.2	Top 300 general Patterson peaks	38
1.2	One vector, error = 0.08 Å	14
1.2	One vector, error = 0.38 Å	41
1.2	One vector, error = 0.40 Å	219
1.2	One vector, error = 1.69 Å	51
1.4	Top 100 general Patterson peaks	10
1.5	Top 100 general Patterson peaks	4
1.5	One vector, error = 0.29 Å	61

may be used to generate further atoms using a full Patterson superposition minimum function or a weighted difference synthesis.

In the case of the small protein BPTI (Schneider, 1998), 15 300 attempts based on 100 different search vectors led to four final solutions with mean phase error less than 18°, although none of the globally highest PMF values for any of the search vectors corresponded to correct solutions. Table 16.1.8.1 shows the effect of using different two-atom search fragments for hirustasin, a previously unsolved 55-amino-acid protein containing five disulfide bridges first solved using *SHELXD* (Usón *et al.*, 1999). It is not clear why some search fragments perform so much better than others; surprisingly, one of the more effective search vectors deviates considerably (1.69 Å) from the nearest true S–S vector.

16.1.9. *Shake-and-Bake*: an analysis of a dual-space method in action

The *Shake-and-Bake* algorithm generated the *SnB* program written in Buffalo at the Hauptman–Woodward Institute, principally by Charles Weeks and Russ Miller (Miller *et al.*, 1994; Weeks & Miller, 1999*a*). *SHELXD* (Usón & Sheldrick, 1999; Schneider & Sheldrick, 2002) and later *HySS* (Grosse-Kunstleve & Adams, 2003) were both inspired by *SnB* and employ the *Shake-and-Bake* philosophy with various modifications, in particular involving the use of the Patterson function to obtain starting phases. It is instructive to see how such software works in detail.

16.1.9.1. Flowchart and program comparison

A flowchart for the generic *Shake-and-Bake* algorithm, which provides the foundation for these programs, is presented in Fig. 16.1.9.1. It contains two refinement loops embedded in the trial-structure loop. The first of these loops (steps 5–9) is a dual-space phase-improvement loop entered by all trial structures, and the second (steps 11–14) is a real-space Fourier-refinement loop entered only by those trial structures that are currently judged to be the best on the basis of some figure of merit. These loops have been called the internal and external loops, respectively, in previous descriptions of the *SHELXD* program (*e.g.* Sheldrick & Gould, 1995; Sheldrick, 1997, 1998). Currently, the major algorithmic differences between the programs are the following:

(*a*) During the reciprocal-space segment of the dual-space loop (Fig. 16.1.9.1, step 5), *SnB* can perform tangent refinement or use parameter shift to reduce the minimal function [equation (16.1.4.2)] or an exponential variant of the minimal function (Hauptman *et al.*, 1999). *SHELXD* performs Karle-type tangent expansion (Karle, 1968). During tangent or parameter-shift refinement with *SnB*, all phases computed in the preceding structure-factor calculation (steps 4 or 9) are refined. During tangent expansion in *SHELXD*, the phases of (typically) the 40% highest calculated E magnitudes are held fixed, and the phases of the remaining 60% are determined by using the tangent formula. If there is a tendency for *SHELXD* to produce uranium-atom solutions, more phases should be held fixed in the tangent phase expansion.

(*b*) In real space, *SnB* uses simple peak picking, varying the number of peaks selected on the basis of structure size and composition. *SHELXD* contains provisions for all the forms of peak picking described above.

(*c*) *SnB* relies primarily on the minimal function [equation (16.1.4.2)] as a figure of merit whereas *SHELXD* uses the correlation coefficient [equation (16.1.6.1)], calculated using all data, after the final dual-space (internal) cycle and in the real-space (external) loop. In addition, *SHELXD* calculates a further correlation coefficient, CC_{weak}, calculated in the same

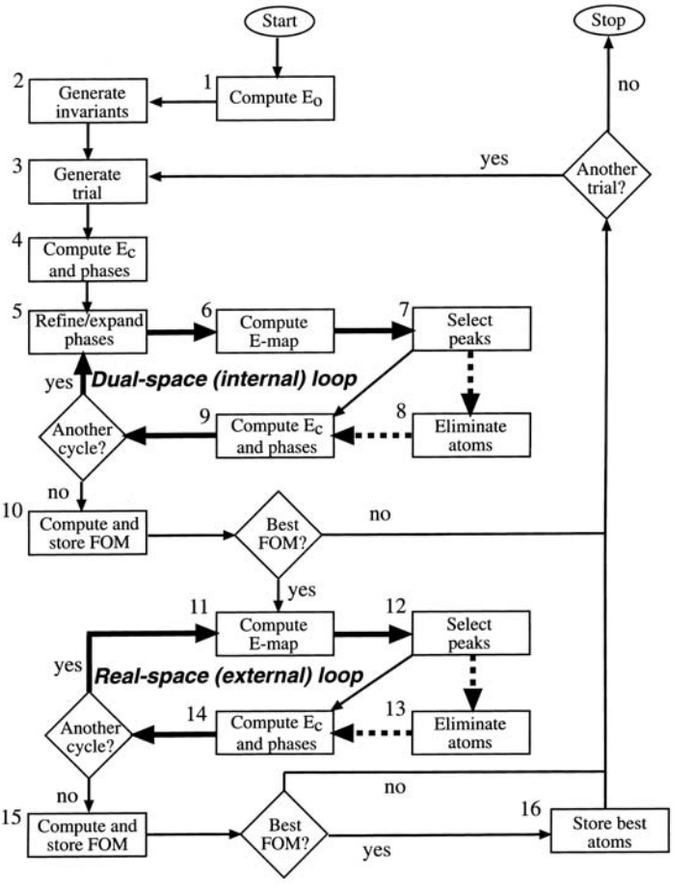

Figure 16.1.9.1
A flowchart for the *Shake-and-Bake* procedure, which is implemented in both *SnB* and *SHELXD*. The essence of the method is the dual-space approach of refining trial structures as they shuttle between real and reciprocal space. In the general case, steps 7 and 12 are any density-modification procedure, and steps 9 and 14 are inverse Fourier transforms rather than structure-factor calculations. The optional steps 8 and 13 take the form of *iterative peaklist optimization* or *random omit maps* in *SHELXD*. Any suitable starting model can be used in step 3, and *SHELXD* attempts to improve on random models (when possible) by utilizing Patterson-based information. Step 4 is bypassed if phase sets (random or otherwise) provide the starting point for the dual-space loop. *SHELXD* enters the real-space loop if the FOM (correlation coefficient) is within a specified threshold (1–5%) of the best value so far.