8 Direct Methods and Refinement

8.1 Introduction

In this chapter, we consider direct methods, also known as phase probability methods, of solving the phase problem, together with Patterson search techniques, least-squares refinement, and other important procedures that are involved in the overall investigation of crystal and molecular structure.

8.2 Direct Methods of Phase Determination

Direct methods of solving the phase problem are an important technique, particularly in their ability to yield good phase information for structures containing no heavy atoms. One feature common to the structure-determining methods that we have encountered so far is that values for phases of X-ray reflections are derived initially by structure factor calculations, albeit on only part of the structure. Since the data from which the best phases are ultimately derived are the F_o values, we may imagine that the phases are encoded somehow in these quantities, even though their actual values are not recorded experimentally. This philosophy led to the search for analytical methods of phase determination, which are independent of trial structures, and initiated the development of direct methods.

8.2.1 Normalized Structure Factors

An important stage in direct phase-determining formulae results by first replacing $F_o(hkl)$ by the corresponding normalized structure factor $|\mathrm{E}(hkl)|$, which is given by the equation

$$|\mathrm{E}(hkl)|^2 = \frac{K^2 F_o(hkl)^2}{\varepsilon \sum_{j=1}^{N} g_j^2} \tag{8.1}$$

The summation in the denominator of (8.1), which is θ-dependent through g_j, may be obtained through a K-curve, similar to that in Fig. 7.24c(ii). The data must be on an absolute scale, Sect. 4.2.1, and ε is incorporated in setting up the K-curve; see also Sects. 13.4.6 and 13.4.10.

The $|\mathrm{E}|$ values have properties similar to those of the sharpened F_o values derived for a point-atom model, Sect. 7.4.4: they are compensated for the fall-off in $\boldsymbol{f}$ with sin θ. High-order reflections with

M. Ladd and R. Palmer, *Structure Determination by X-ray Crystallography: Analysis by X-rays and Neutrons*, DOI 10.1007/978-1-4614-3954-7_8,

© Springer Science+Business Media New York 2013

Table 8.1 Some theoretical and experimental values related to |E| values statistics

	Theoretical values		Experimental values and conclusions	
Mean values	$P\bar{1}(C)$	$P1(A)$	Crystal 1	Crystal 2
$\lvert E\rvert^2$	1.00	1.00	0.99	0.98
$\lvert E\rvert$	0.80	0.89	0.85 *A*/*C*	0.84 *A*/*C*
$\lvert\lvert E\rvert^2 - 1\rvert$	0.97	0.74	0.91 *C*	0.82 *A*
Distribution	%	%	%	%
$\lvert E\rvert > 3.0$	0.30	0.01	0.20 *C*	0.05 *A*
$\lvert E\rvert > 2.5$	1.24	0.19	0.90 *C*	0.98 *C*
$\lvert E\rvert > 2.0$	4.60	1.80	2.70 *A*/*C*	2.84 *A*/*C*
$\lvert E\rvert > 1.75$	8.00	4.71	7.14 *C*	6.21 *A*/*C*
$\lvert E\rvert > 1.5$	13.4	10.5	12.9 *C*	10.5 *A*
$\lvert E\rvert > 1.0$	32.0	36.8	33.7 *C*	37.1 *A*

comparatively small |F| values can have relatively large |E| values, an important fact in the application of direct methods. We may note in passing that $|E|^2$ values, or |E| |F|, can be used as coefficients in a sharpened Patterson function; also, since $|E|^2 = 1$, Table 8.1 and Sect. 4.2.5, the coefficients ($|E|^2 - 1$) produce a sharpened Patterson function with the origin peak removed. This technique is useful because, in addition to the general sharpening effect, vectors of small magnitude that are often swamped by the origin peak may be revealed.

Epsilon ($\boldsymbol{\varepsilon}$) Factor

Because of the importance of individual reflections in direct phasing methods, care must be taken to obtain the best possible |E| values. The factor ε in the denominator of (8.1) takes account of the fact that reflections in certain reciprocal lattice zones or rows may have average intensity greater than that for general reflections. The ε-factor depends upon the crystal class, and its values for all crystal classes are listed in Table 4.2. Detailed considerations of the ε factor and of the statistics of |E| values have been presented in Sect. 4.2.3.

Distributions of |E| Values

The distribution of |E| values holds useful information about the space group of a crystal. Theoretical quantities derived for equal-atom structures in space groups $P1$ and $P\bar{1}$ are indicated in Table 8.1, together with the experimental results for two crystals.

Crystal 1 is pyridoxal phosphate oxime dihydrate, $C_8H_{11}N_2O_6P{\cdot}2H_2O$, which is triclinic. The values in Table 8.1 favor the centric distribution *C*, and the structure analysis [1] confirmed the assignment of space group $P\bar{1}$. Crystal 2 is a penta-uloside sugar; the results correspond, on the whole, to an acentric distribution *A*, as expected for a crystal of space group $P2_12_12_1$ [2]. It should be noted that the experimentally derived quantities do not always have a completely one-to-one correspondence with the theoretical values, and care should be exercised in using these statistics to select a space group.

8.2.2 Structure Invariants and Origin-Fixing Reflections

The formulae used in direct phasing require, initially, the use of a few reflections with phases known, either uniquely or symbolically; we consider first centrosymmetric primitive space groups.

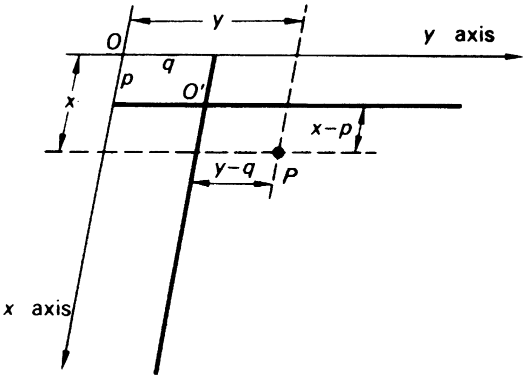


Fig. 8.1 Transformation of the point $P(x, y)$, with respect to two-dimensional axes, by moving the origin from O to O' (p, q); the transformed coordinates of P are $(x - p, y - q)$

In this group of crystals, the origin is usually taken on the center of symmetry in the unit cell at 0, 0, 0 and we speak of the sign $s(hkl)$ of the reflection; $s(hkl)$ is $\mathrm{F}(hkl)/|\mathrm{F}(hkl)|$ and is either + or −. We shall show that, in any primitive centrosymmetic space group, arbitrary signs can be allocated to three reflections, chosen according to certain rules, in order to specify the origin at one of the eight centers of symmetry in the unit cell. These signs form a basic set, or "fountainhead," from which more signed reflections emerge as the analysis proceeds. From (3.69), remembering that we are dealing with a centrosymmetric crystal, we write

$$\mathrm{F}(hkl)_{0,0,0} = \sum_{j=1}^{N} g_j \cos 2\pi(hx_j + ky_j + lz_j) \tag{8.2}$$

where $\mathrm{F}(hkl)_{0,0,0}$ indicates an origin of coordinates at the point 0, 0, 0. If this origin is moved to a center of symmetry at $\frac{1}{2}, \frac{1}{2}, 0$, the point that was originally x_j, y_j, z_j, now becomes $x_j - \frac{1}{2}, y_j - \frac{1}{2}, z_j$, Fig. 8.1, with $p = q = \frac{1}{2}$. The structure factor equation may now be written as

$$\mathrm{F}(hkl)_{1/2,1/2,0} = \sum_{j=1}^{N} g_j \cos 2\pi[(hx_j + ky_j + lz_j) - (h+k)/2] \tag{8.3}$$

Expanding the cosine term (see Web Appendix WA5), and remembering that $\sin[2\pi(h + k)/2] = \sin[n\pi] = 0$, and $\cos[2\pi(h + k)/2] = \cos[n\pi] = (-1)^n$, we obtain

$$\mathrm{F}(hkl)_{1/2,1/2,0} = (-1)^{h+k}\mathrm{F}(hkl)_{0,0,0} \tag{8.4}$$

Equation (8.4) demonstrates that the amplitude $|\mathrm{F}(hkl)|$ is invariant under change of origin, as would be expected, but that a change of sign may occur, depending on the parity of the indices hkl. The complete results are presented in Table 8.2, the use of which is illustrated by the following examples.[1]

[1] See also Bibliography, Ladd and Palmer (1980).

Table 8.2 Effect of a change of the origin coordinates, among centers of symmetry, on the sign of a structure factor

	Parity group							
	1	2	3	4	5	6	7	8
Centers of symmetry	*h* even *k* even *l* even	*h* odd *k* even *l* even	*h* even *k* odd *l* even	*h* even *k* even *l* odd	*h* even *k* odd *l* odd	*h* odd *k* even *l* odd	*h* odd *k* odd *l* even	*h* odd *k* odd *l* odd
0, 0, 0	+	+	+	+	+	+	+	+
$\frac{1}{2}$, 0, 0	+	−	+	+	+	−	−	−
0, $\frac{1}{2}$, 0	+	+	−	+	−	+	−	−
0, 0, $\frac{1}{2}$	+	+	+	−	−	−	+	−
0, $\frac{1}{2}$, $\frac{1}{2}$	+	+	−	−	+	−	−	+
$\frac{1}{2}$, 0, $\frac{1}{2}$	+	−	+	−	−	+	−	+
$\frac{1}{2}$, $\frac{1}{2}$, 0	+	−	−	+	−	−	+	+
$\frac{1}{2}$, $\frac{1}{2}$, $\frac{1}{2}$	+	−	−	−	+	+	+	−

For example, the reflection 312 belongs to the odd-odd-even (ooe) parity group 7. If *s*(312) is given a plus sign, the origin could be regarded as being restricted to one from the following list:

$$0, 0, 0; \quad 0, 0, \tfrac{1}{2}; \quad \tfrac{1}{2}, \tfrac{1}{2}, 0; \quad \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}$$

Similarly, if *s*(322), parity group 2(oee), is also given a plus sign, the possible origins are

$$0, 0, 0; \quad 0, \tfrac{1}{2}, 0; \quad 0, 0, \tfrac{1}{2}; \quad 0, \tfrac{1}{2}, \tfrac{1}{2}$$

Combining these two sign allocations, the common origins are

$$0, 0, 0; \quad 0, 0, \tfrac{1}{2}$$

In order to fix the origin uniquely at, say, 0, 0, 0, we select another reflection with a plus sign with respect to 0, 0, 0. Reference to Table 8.2 shows that parity groups 4, 5, 6, and 8 each meet this requirement. For the following reasons, parity groups 1 and 3 are excluded from the choice as the third origin-specifying reflection. Group 1 is a special case because signs of eee reflections do not change sign among the permitted ($\bar{1}$) origins, and Group 3 (eoe) is related to groups 2 and 7 through an addition of indices:

$$312 + 322 \rightarrow 634 \tag{8.5}$$

or, more generally,

$$\text{ooe} + \text{oee} \rightarrow \text{eoe} \tag{8.6}$$

since o + o = e + e = e, and e + o = o. Parity groups 2, 3, and 7 are said to be linearly related, and cannot be used together in defining the choice of origin. As stated above, structure factors belonging to parity group 1 do not change sign on change of origin, as is evident from both the development of (8.4) and Table 8.2. Reflections in this group are called *structure seminvariants*; their actual signs depend on the structure among the permitted origins, and cannot be used to restrict the origin. Hence, the origin will be fixed by choosing a third reflection from one of the groups 4, 5, 6, or 8; the reflections so chosen, three in this case, form a *starting set*.

8.2.3 Sign Determination: Centrosymmetric Crystals

Many equations have been proposed that are capable of providing sign information for centrosymmetric crystals. Two of these expressions have proved to be outstandingly useful, and it is to them that we first turn our attention.

Triple-Product Sign Relationship

In 1952, Sayre [3] derived a general formula for hypothetical structures containing identical resolved atoms. For centrosymmetric crystals, it was given in the form

$$s(hkl)\,s(h'k'l')\,s(h-h',k-k',l-l') \approx +1 \tag{8.7}$$

where ≈ here means "is probably equal to." The vectors associated with these reflections, $d^*(hkl)$, $d^*(h'k'l')$, and $d^*(h-h', k-k', l-l')$ form a closed triangle, or vector triplet, in reciprocal space. In practice, it may be possible to form several such vector triplets for a given hkl; Fig. 8.2a shows two triplets for the vector 300. If any two of the signs in (8.7) are known, the third can be deduced, and we can then extend the sign information beyond that given in the starting set.

A physical meaning can be given to (8.7) by drawing the traces, in real space, of the three planes that form a vector triplet in reciprocal space (Fig. 8.2a, b). For a centrosymmetric crystal, we may write

$$\rho(xyz) = \frac{2}{V_{\mathrm{c}}}\sum_h\sum_k\sum_l \pm|\mathrm{F}(hkl)|\cos 2\pi(hx+ky+lz) \tag{8.8}$$

The $|\mathrm{F}(hkl)|$ terms in this equation take a *positive* sign if the traces of the corresponding planes pass through the origin, like the full lines in Fig. 8.2b, and a *negative* sign if they lie midway between these positions, like the dashed lines in the figure. The combined contributions from the three planes in question will thus have maxima at the points of their mutual intersections, which are therefore potential atomic sites, and correspond to regions of high electron density.

This argument is particularly strong if the three planes have large |E| values, because in |E| the damping effect of f has been significantly reduced, leaving a term that is governed by the structure itself. It may be seen from the diagram that triple intersections occur only at points where either three full lines (+ + +) meet, or two dashed lines and one full line meet (some combination of + − −). This result is in direct agreement with (8.7). It is interesting to recall that the structure of hexamethylbenzene was solved in 1929 by Lonsdale [4] through drawing the traces of three high-order, high-intensity reflection planes, $7\bar{3}0$, 340, and $4\bar{7}0$, and placing atoms at their intersections. These planes form a vector triplet, and this structure determination contained, therefore, the first, but apparently inadvertent, use of direct methods.

Σ_2−Formula

Hauptman and Karle [5] have given the more general form of (8.7) as

$$s[\mathrm{E}(hkl)] \approx s\left[\sum_{h'k'l'} E(h'k'l')\,E(h-h',k-k',l-l')\right] \tag{8.9}$$

where $E(\mathbf{h})$ is the value of $|\mathrm{E}(\mathbf{h})|$ with its sign, and the summation is taken over all vector pairs with known signs which form a triplet with hkl. The probability associated with (8.9) is given by [6]

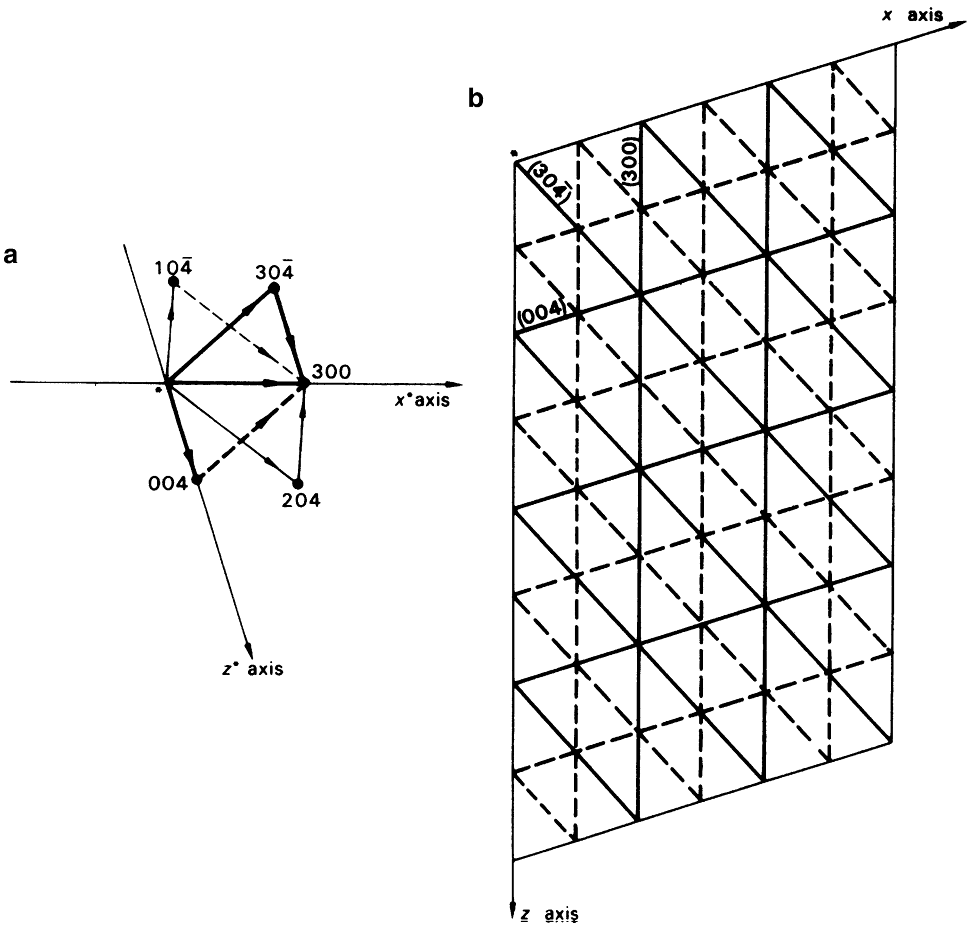


Fig. 8.2 (**a**) Vector triplets 300, 204, $10\bar{4}$ and 300, $30\bar{4}$, 004. (**b**) Physical interpretation of (8.7): the lines are traces of the families (300), (004), and $(30\bar{4})$, and the points of triple intersection correspond with (8.7)

$$P_+(hkl) = \tfrac{1}{2} + \tfrac{1}{2}\tanh[(\sigma_3/\sigma_2^{3/2})\alpha'] \tag{8.10}$$

where α' and σ_n are given by

$$\alpha' = |\mathrm{E}(hkl)| \sum_{h'k'l'} E(h'k'l')\, E(h-h', k-k', l-l') \tag{8.11}$$

$$\sigma_n = \sum_j Z_j^n \tag{8.12}$$

and Z_j is the atomic number of the jth atom. For a structure containing N identical atoms, $\sigma_3/\sigma_2^{3/2}$ is equal to $N^{-1/2}$. From (8.11), the probability is strongly dependent upon the magnitudes of the |E| values. Furthermore, unless glide-plane or screw-axis symmetry is present, or there exist some other means of generating negative signs, (8.9) will produce only positive signs for all $|\mathrm{E}(hkl)|$ used. Such a situation would correspond to a structure with a very heavy atom at the origin and would, in general, indicate an incorrect solution.

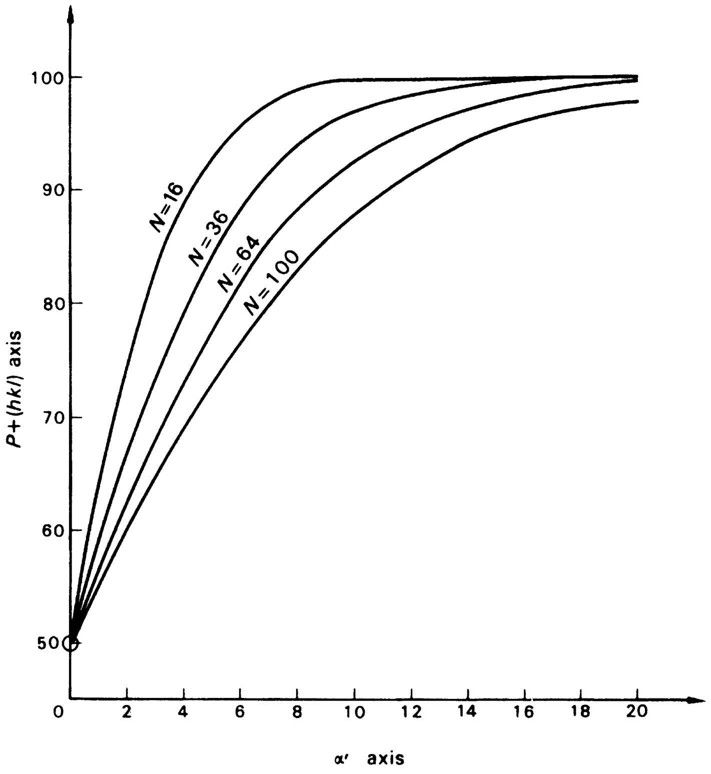


Fig. 8.3 Percentage probability of a single triple-product (Σ_2) sign relationship as a function of α' for different numbers N of atoms in a unit cell, according to (8.10)

If the combination of signs under the summation in (8.11) produces a large and negative value for α', the corresponding value of $P_+(hkl)$ may tend to zero. This result indicates that $s(hkl)$ is negative, with a probability that tends to unity. Probability curves for different numbers N of atoms in the unit cell as a function of α' are shown in Fig. 8.3. Since the most reliable signs from (8.9) are associated with large |E| values, we can now add to the origin-specifying criteria the requirements of both large |E| magnitudes and a large number of Σ_2 interactions for each reflection in the starting set. In this way, strong and reliable sign propagation is encouraged.

In order to illustrate the operation of the Σ_2 relationship, we shall consider the two vector triplets in Fig. 8.2. The sign to be determined is $s(300)$, the others are assumed to be known. It may be noted that sometimes we speak of a sign as + or − and at other times as +1 or −1. The latter formulation is clearly more appropriate to computational methods. The data are tabulated as follows:

hkl	$\lvert E(hkl)\rvert$
300	2.40

$h'k'l'$	$E(h'k'l')$	$h-h'$, $k-k'$, $l-l'$	$E(h-h'$, $k-k'$, $l-l')$	α'	$s(hkl)$	$P_+(hkl)$ (%)
$10\bar{4}$	+2.03	204	−2.22	−19.3	−1	0.8
004	−1.95	$30\bar{4}$	+1.81		−1	

Assuming an N of 64, the indication given is that $s(300)$ is negative with a probability of 99.2%.

8.2.4 Amplitude Symmetry and Phase Symmetry

In space group $P\bar{1}$, the only symmetry-related structure factors are F(hkl) and $\mathrm{F}(\bar{h}\,\bar{k}\,\bar{l})$. According to Friedel's law the intensities and, hence the amplitudes, of these structure factors are equal, and in centrosymmetric space groups $s(hkl) = s(\bar{h}\,\bar{k}\,\bar{l})$. Thus, the amplitude symmetry and the phase symmetry follow the same law, but this will not necessarily be true in other space groups.

From the geometric structure factor for space group $P2_1/c$, see (3.87),

$$|\mathrm{F}(hkl)| = |\mathrm{F}(\bar{h}\,\bar{k}\,\bar{l})| = |\mathrm{F}(h\bar{k}l)| = |\mathrm{F}(\bar{h}k\bar{l})| \tag{8.13}$$

and for the signs there are two possibilities to consider:

$$k + l = 2n: \quad s(hkl) = s(\bar{h}\,\bar{k}\,\bar{l}) = s(h\bar{k}l) = s(\bar{h}k\bar{l}) \tag{8.14}$$

$$k + l = 2n + 1: \quad s(hkl) = s(\bar{h}\,\bar{k}\,\bar{l}) = -s(h\bar{k}l) = -s(\bar{h}k\bar{l}) \tag{8.15}$$

These relationships provide enhanced opportunities for Σ_2 relationships to be developed, including negative values for signs, and in this way space-group symmetry can improve the chances of a successful phase determination. When considering phase relationships in a non-centrosymmetric space group, we need to take note of the change in sign and magnitude of both the A and the B components of the geometrical structure factor, as h, k, or l take a negative sign. Consider, for example, space group Pc (b axis unique; origin on c). We can separate (3.85) into two parts, according to whether l is odd or even:

$l = 2n$

$$\begin{aligned} A &= 2\cos 2\pi(hx + lz)\,\cos 2\pi ky \\ B &= 2\sin 2\pi(hx + lz)\,\cos 2\pi ky \end{aligned}$$

$l = 2n + 1$

$$\begin{aligned} A &= -2\sin 2\pi(hx + lz)\,\sin 2\pi ky \\ B &= 2\cos 2\pi(hx + lz)\,\sin 2\pi ky \end{aligned}$$

It is clear that, in all cases, $\phi(hkl) = -\phi(\bar{h}\,\bar{k}\,\bar{l}) \neq \phi(\bar{h}kl)$. However, for the two parities of l, the following expressions hold.

$$\begin{aligned} l = 2n&: \quad \phi(hkl) = \phi(h\bar{k}l) \quad \text{and} \quad \phi(\bar{h}kl) = -\phi(hk\bar{l}) \\ l = 2n + 1&: \quad \phi(hkl) = \pi + \phi(h\bar{k}l) \quad \text{and} \quad \phi(\bar{h}kl) = \pi - \phi(hk\bar{l}) \end{aligned}$$

The amplitude symmetry and phase symmetry for all space groups are contained in the *International Tables for X-ray Crystallography*, Volume A (or Volume 1) [7].

8.2.5 Σ_2-Listing

Because of both the increased probability in relationships developed for reflections with high |E| values and the existence of many vector triplets in a complete set of data, the initial application of direct methods is limited to reflections with large |E| values, say, greater than 1.5.

A Σ_2 listing is prepared by considering each value of $|\mathrm{E}(hkl)|$ greater than the preset limit in order of decreasing magnitude, as a basic hkl vector, and searching the data for all possible interactions with $h'k'l'$

Table 8.3 Starting set for the symbolic-addition procedure

hkl	|E|	Sign
$9\bar{1}\bar{4}$	2.97	+
$8\bar{1}\bar{5}$	3.00	+
$\bar{1}40$	2.38	+
020	4.50	*A*
253	2.24	*B*
822	2.71	*C*
303	2.69	*D*
023	2.28	*E*

and $h - h'$, $k - k'$, $l - l'$. Some reflections will enter into many such interactions, while others will produce only a small number.

8.2.6 Symbolic-Addition Procedure: Example

Karle and Karle [8] described a technique for the systematic application of the Σ_2 formula for building up a self-consistent sign set. The various steps involved are outlined below, using results obtained with pyridoxal phosphate oxime dihydrate [9].

Crystal Data
Formula: $C_8H_{11}N_2O_6P{\cdot}2H_2O$
System: triclinic
Unit-cell dimensions: $a = 10.94$ Å, $b = 8.06$ Å, $c = 9.44$ Å; $\alpha = 57.18°$, $\beta = 107.68°$, $\gamma = 116.53°$
V_c: 623.75 Å^3
D_m: 1.57 g cm^{-3}
M_r: 298.19
Z: 1.98 or 2 to the nearest integer
Absent spectra: none
Possible space groups: $P1$ or $P\bar{1}$; $P\bar{1}$ was chosen on the basis of intensity statistics
All atoms are in general positions.

Sign Determination

1. A total of 163 reflections for which $|E| \geq 1.5$ was arranged in descending order of magnitude, and a Σ_2 listing was obtained, Table 8.4.
2. From a study of the Σ_2 listing, three reflections were allocated + signs, Table 8.3; they are the origin-fixing reflections, selected according to the procedures already discussed.
3. Equation (8.9) was used by searching, initially between members of the origin-fixing set and other reflections. In order to maintain a high probability, only the highest |E| values were used. For example, $9\bar{5}\bar{5}$ with $|E| = 2.31$ is generated by the combination of $8\bar{1}\bar{5}$ and $\bar{1}40$

$$s(9\bar{5}\bar{5}) \approx s(8\bar{1}\bar{5})\,s(\bar{1}40) = (+1)(+1) = +1 \tag{8.16}$$

From (8.11), α' is 16.5, and Fig. 8.3, with $N = 38$ and excluding hydrogen, shows that the probability of this indication is about 99.7%. The new sign was accepted and used to generate more signs. This process was continued until no new signs could be developed with high probability.

Table 8.4 Σ_2 listing for reflection $9\bar{8}\,\bar{6}$ of pyridoxal oxime phosphate with symbols added[a]

$h'k'l'$	$s(h'k'l')$	$\lvert E_2\rvert$: $\lvert E(h'k'l')\rvert$	$h-h'$, $k-k'$, $l-l'$	$s(h-h'$, $k-k'$, $l-l')$	$\lvert E_3\rvert$: $\lvert E(h-h'$, $k-k'$, $l-l')\rvert$	$\lvert E_1\rvert\lvert E_2\rvert\lvert E_3\rvert$	$s(9\bar{8}\,\bar{6})$	$P_+(9\bar{8}\,\bar{6})$ (%)
$1\,\bar{5}\,0$	*BD*	2.16	$8, \bar{3}, \bar{6}$	*A*	1.63	6.40	*ABD*	90
$10, \bar{2}\,\bar{2}$	*AB*	2.04	1, 6, 4	*D*	1.88	6.97	*ABD*	91
$10, \bar{7}\,\bar{1}$	*D*	1.87	1, 1, 5	*AB*	1.63	5.54	*ABD*	87
$4\,\bar{8}\,\bar{3}$	*D*	1.83	$5, 0, \bar{3}$	*ECD*	1.58	5.25	*EC*	85
$3\,\bar{9}\,\bar{4}$		1.76	$6, 1, \bar{2}$		1.58	5.03		
$3\,\bar{5}\,\bar{6}$		1.70	$6, \bar{3}, 0$		1.51	4.66		
$6\,\bar{7}\,\bar{2}$		1.68	$3, \bar{1}, \bar{4}$		1.63	4.98		
$10, \bar{4}\,\bar{2}$	*B*	1.62	1, 4, 4	*AD*	1.67	4.93	*ABD*	84
$0\,\bar{2}\,0$	$-A$	4.50	$9, \bar{6}, 6$		1.73	14.08		
$0\,\bar{8}\,0$	+	2.48	$9, 0, \bar{6}$		1.85	8.30		

[a] $\lvert E(9\bar{8}\,\bar{6})\rvert = \lvert E_1\rvert = 1.89$. We can use both **h** − **k** and **h** + **k** in these triple products

4. At this stage, it is often found that the number of signs developed with confidence is small. This situation arose with pyridoxal phosphate oxime dihydrate, and the Σ_2 formula was then applied to reflections with symbolic signs. In this technique, a reflection was selected, again by virtue of its high |E| value and long Σ_2 listing, and allocated a letter symbol, as shown in Table 8.3; the letter symbols *A*–*E* represent a + or − sign. Generally, less than five symbolic phases are sufficient, and there are no necessary restrictions on the parities of these reflections. However, it is desirable that there are no redundancies in the complete starting set, that is, no three reflections in the set should themselves be related by a triple-product relationship.

   As a symbol became involved in a sign of a reflection, it was written into the Σ_2 listing. The example in Table 8.4 shows a Σ_2 entry for reflection $9\bar{8}\,\bar{6}$. Reading across the table, sign combinations are seen to be generated by multiplying $s(h'k'l')$ by $s(h-h', k-k', l-l')$, which are then written as $s(9\bar{8}\,\bar{6})$ in the penultimate column. Recurring symbol combinations, such as *ABD*, gave rise to consistent indications. If the probability that $s(9\bar{8}\,\bar{6}) = s(ABD)$ is sufficiently large, this sign value is entered for $s(9\bar{8}\,\bar{6})$ wherever these indices occur. In the final column of the table, the probability of each sign indication is listed. Although they are small individually, the combined probability from (8.10) that $s(9\bar{8}\,\bar{6})$ was *ABD* is 100%.
5. When this process had been exhausted, the results were examined for agreement among sign relationships. For example, in Table 8.4 there is a weak indication that $ABD = EC$. The most significant relationships found overall were $AC = E$, $C = EB$, $B = ED$, $AD = E$, and $AB = CD$. Multiplying the first by the second, and the first by the fourth, and remembering that products such as A^2 equal +1 reduces this list to $A = B$, $C = D$, and $E = AC$. The five symbols were reduced, effectively, to two, *A* and *C*. The sign determination was rewritten in terms of signs and the symbols *A* and *C*; reflections with either uncertain or undetermined signs were rejected from the first electron density calculation.

8.2.7 Calculation of E Maps

The result of the above analysis meant that four possible sign sets could be generated by the substitutions $A = \pm 1$, $C = \pm 1$. The set with $A = C = +1$ was rejected immediately because this phase assignment implies a very heavy atom at the origin of the unit cell. The three other sign

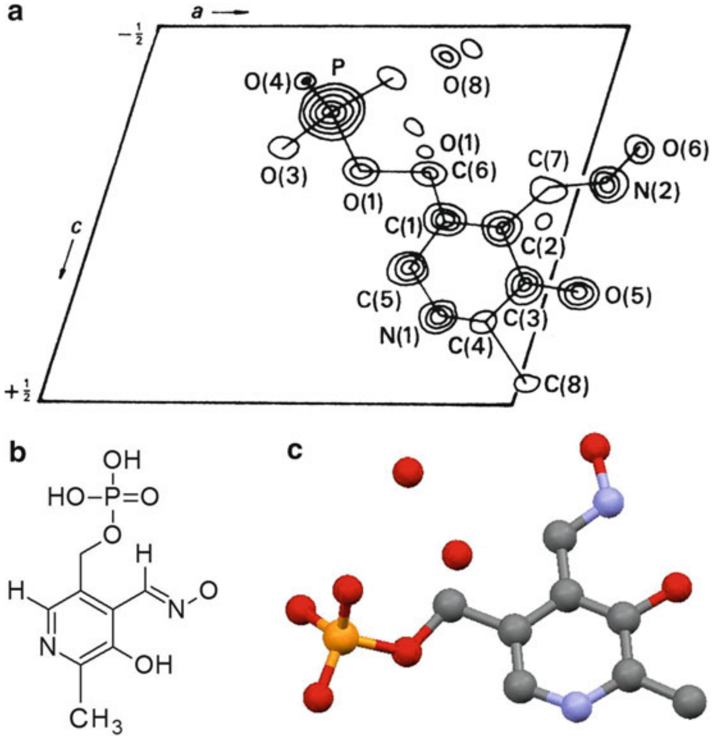


Fig. 8.4 Pyridoxal phosphate oxime dihydrate. (**a**) Composite three-dimensional E map as seen along *b*. (**b**) Chemical formula. (**c**) Molecular conformation derived from (**a**) excluding H atoms (drawn with RASMOL)

combinations were used to calculate E maps. These maps were obtained by Fourier syntheses, using (8.8), but with |E| replacing |F| as the coefficients. The sharp nature of |E| implicit in (8.1) is advantageous when using a limited number of data to resolve atomic peaks in the electron density map, although normally about eight reflections per atom in the asymmetric unit are desirable. Spurious peaks can arise, however, like that in the vicinity of O(8), because of the limited number of coefficients in the Fourier series for an E map. The sign combination for pyridoxal phosphate oxime dihydrate that led to an interpretable E map was $A = C = -1$. The atomic positions from this map, Fig. 8.4a, were used in a successful refinement of the structure, and Fig. 8.4b shows the chemical formula and Fig. 8.4c shows the molecular conformation determined in this study.

If there are n symbolic signs in the final centrosymmetric phase solution, there will be 2^n combinations, each of which can give rise to an E map, and it is desirable to set up criteria that will seek the most probable set. We shall consider such criteria during our discussion of the non-centrosymmetric case, where they are of even greater importance.

8.2.8 Phase Determination: Non-centrosymmetric Crystals

The non-centrosymmetric case is more difficult, both in theory and in practice. Much of this difficulty stems from the fact that the phase angle can take on any value between 0 and 2π, with a consequent imprecision in its determination. Nevertheless, direct methods are used regularly to solve structures, the limiting factor being the number of atoms N in the unit cell, which is space-group dependent. Structures where N is 400 or more can be expected to be solved without great difficulty using current versions of the various programs now available.

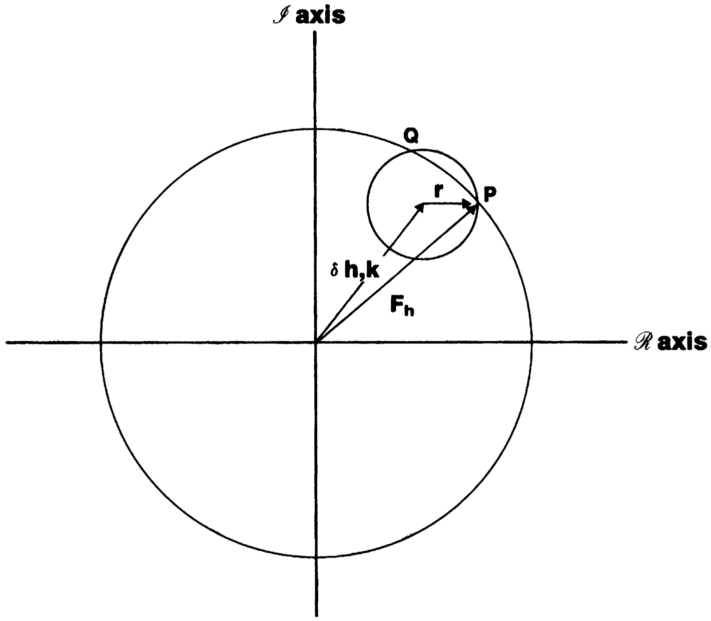


Fig. 8.5 Illustration of (8.19) at equality

Equations for a non-centrosymmetric crystal are, not surprisingly, more generalized expressions of those such as (8.9), which relates to centrosymmetric crystals. Using the fact that the electron density distribution is a nonnegative function, Karle and Hauptman derived a set of inequality relationships [10]. The first three inequalities may be written as nonnegative functions.

$$\mathrm{F}_{000} \geq 0 \tag{8.17}$$

$$|\mathrm{F}_{\mathbf{h}}| \leq \mathrm{F}_{000} \tag{8.18}$$

$$\mathrm{F}_{\mathbf{h}} - \delta_{\mathbf{h},\mathbf{k}} \leq |\mathbf{r}| \tag{8.19}$$

where

$$\delta_{\mathbf{h},\mathbf{k}} = \mathrm{F}_{\mathbf{h}-\mathbf{k}}\mathrm{F}^{\mathrm{a}}_{\mathbf{k}}/\mathrm{F}_{000} \tag{8.20}$$

and

$$|\mathbf{r}| = \frac{\begin{vmatrix} \mathrm{F}_{000} & \mathrm{F}^{*}_{\mathbf{h}-\mathbf{k}} \\ \mathrm{F}_{\mathbf{h}-\mathbf{k}} & \mathrm{F}_{000} \end{vmatrix} \begin{vmatrix} \mathrm{F}_{000} & \mathrm{F}^{*}_{\mathbf{k}} \\ \mathrm{F}_{\mathbf{k}} & \mathrm{F}_{000} \end{vmatrix}}{\mathrm{F}_{000}} \tag{8.21}$$

We use here the convenient notation **h** for *hkl*, **k** for *h′k′l′*, **h** − **k** for a third reflection that forms a vector triplet with **h** and **k**, and, in order to avoid excessive parentheses, $\mathrm{F}_{\mathbf{h}}$ for F(**h**). Equations (8.17) and (8.18) are immediately acceptable in terms of earlier discussions in this book; $\mathrm{F}_{\mathbf{k}}\mathrm{F}_{\mathbf{h}-\mathbf{k}}$ is a multiplication of two structure factors and may be interpreted on an Argand diagram as $|\mathrm{F}_{\mathbf{k}}||\mathrm{F}_{\mathbf{h}-\mathbf{k}}| \exp\{\mathrm{i}(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})\}$.

The structure factor $\mathrm{F}_{\mathbf{h}}$, given only $|\mathrm{F}_{\mathbf{h}}|$, must lie on a circle of that radius on an Argand diagram, Fig. 8.5. Equation (8.19) then indicates that $\mathrm{F}_{\mathbf{h}}$ lies within a circle, center $\delta_{\mathbf{h},\mathbf{k}}$ and radius $|\mathbf{r}|$, between the points P and Q. Expansions of the determinants in (8.21), remembering that $\mathrm{FF}^* = |\mathrm{F}|^2$, shows that the larger the values of $|\mathrm{F}_{\mathbf{k}}|^2$ and $|\mathrm{F}_{\mathbf{h-k}}|^2$, the closer $\mathrm{F}_{\mathbf{h}}$ approaches $\delta_{\mathbf{h,k}}$. For a given **h**, as **k** is varied, $\mathrm{F}_{\mathbf{h}}$ is proportional to the average over **k**:

$$\mathrm{F}_{\mathbf{h}} \propto \langle \mathrm{F}_{\mathbf{k}}\mathrm{F}_{\mathbf{h-k}} \rangle_{\mathbf{k}} \tag{8.22}$$

the proportionality constant being F(000). We can see how (8.22) can give rise to (8.9) or (8.7) for a single interaction.

Using the general relation

$$\mathrm{F}_{\mathbf{h}} = |\mathrm{F}_{\mathbf{h}}| \exp(\mathrm{i}\phi_{\mathbf{h}}) \tag{8.23}$$

we obtain the phase addition formula

$$\phi_{\mathbf{h}} \approx \phi_{\mathbf{k}} + \phi_{\mathbf{h-k}} \tag{8.24}$$

The sign ≈ indicates an approximation which is better the larger the values of the corresponding structure factors. Where several triplets are involved with a given **h**, (8.24) becomes

$$\phi_{\mathbf{h}} \approx \langle \phi_{\mathbf{k}} + \phi_{\mathbf{h-k}} \rangle_{\mathbf{k}} \tag{8.25}$$

where $\langle\ \rangle_{\mathbf{k}}$ implies an average, taken over a number of triple product relationships (TPRs) common to **h**.

The F_{o} data derived experimentally are converted to |E| values. Again, we commence phase determination with |E| values greater than about 1.5 in order to maintain acceptable probability limits. Equation (8.25) is illustrated by an Argand diagram in Fig. 8.6 for four values of κ; $\phi_{\mathbf{h}}$ is the estimated phase angle associated with the resultant $\mathrm{R}_{\mathbf{h}}$. Each direction labeled κ depends on a product $|\mathrm{E}_{\mathbf{k}}||\mathrm{E}_{\mathbf{h-k}}|$ and may be resolved into components A and B along the real and imaginary axes, respectively, such that

$$A = |\mathrm{E}_{\mathbf{k}}||\mathrm{E}_{\mathbf{h-k}}| \cos(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}}) \tag{8.26}$$

and

$$B = |\mathrm{E}_{\mathbf{k}}||\mathrm{E}_{\mathbf{h-k}}| \sin(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}}) \tag{8.27}$$

It follows from (8.25) and (8.26), including a frequency enhancement weight $w_{\mathbf{h}}$, that

$$\tan\phi_{\mathbf{h}} \approx \frac{\sum_{\mathbf{h}} w_{\mathbf{h}} |\mathrm{E}_{\mathbf{k}}||\mathrm{E}_{\mathbf{h-k}}| \sin(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}})}{\sum_{\mathbf{h}} w_{\mathbf{h}} |\mathrm{E}_{\mathbf{k}}||\mathrm{E}_{\mathbf{h-k}}| \cos(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}})} \tag{8.28}$$

Equation (8.28) is a weighted tangent formula. Weights may be used as defined below under MULTAN. Current phase-determining procedures are based largely on (8.28), and the reliability of (8.28) can be measured by the variance $V(\phi_{\mathbf{h}})$. Figure 8.7 shows $V(\phi_{\mathbf{h}})$ as a function of $\alpha_{\mathbf{h}}$, where

$$\alpha_{\mathbf{h}}^2 = \left[\sum_{\mathbf{h}} \kappa_{\mathbf{hk}} \cos(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}})\right]^2 + \left[\sum_{\mathbf{k}} \kappa_{\mathbf{hk}} \sin(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}})\right]^2 \tag{8.29}$$

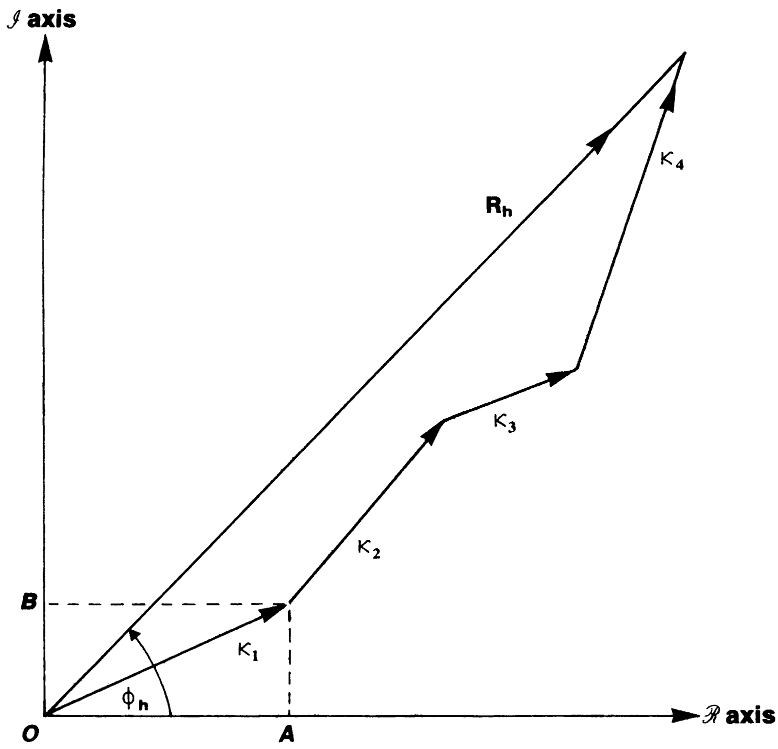


Fig. 8.6 Summation of four "vectors" κ_1–κ_4 on an Argand diagram; the resultant is R_h, with a phase angle ϕ_h

and

$$\kappa_{\mathbf{hk}} = 2\sigma_3\sigma_2^{-3/2}|E_{\mathbf{h}}||E_{\mathbf{k}}||E_{\mathbf{h-k}}| \tag{8.30}$$

with

$$\sigma_n = \sum_{j=1}^{N} Z_j^n \tag{8.31}$$

as before, and the sum is taken over the N atoms in the unit cell. The parameter $\alpha_{\mathbf{h}}$ gives a measure of the reliability with which $\phi_{\mathbf{h}}$ is determined by the tangent formula. When (8.29) contains only one term, as it may in the initial stages of phase determination, then $\alpha_{\mathbf{h}} = \kappa_{\mathbf{hk}}$ and is strongly dependent on the product $|E_{\mathbf{h}}||E_{\mathbf{k}}||E_{\mathbf{h-k}}|$. Figure 8.7 shows clearly that $V(\phi_{\mathbf{h}})$ has acceptably small values when $\alpha_{\mathbf{h}}$ is greater than about 4, corresponding to $V^{1/2} < 30°$, but increases rapidly for $\alpha_{\mathbf{h}}$ decreasing below about 3, corresponding to $V^{1/2} > 40°$: $\alpha_{\mathbf{h}}$ depends also on $\sigma_3\sigma_2^{-3/2}$, which, in turn, depends on the number and types of atoms in the unit cell. This dependence may be illustrated by hypothetical structures containing different numbers N of identical atoms; $\alpha_{\mathbf{h}}$ $(=\kappa_{\mathbf{hk}})$ is then given by

$$\alpha_{\mathbf{h}} = \frac{2}{N^{1/2}}|E_{\mathbf{h}}||E_{\mathbf{k}}||E_{\mathbf{h-k}}| \tag{8.32}$$

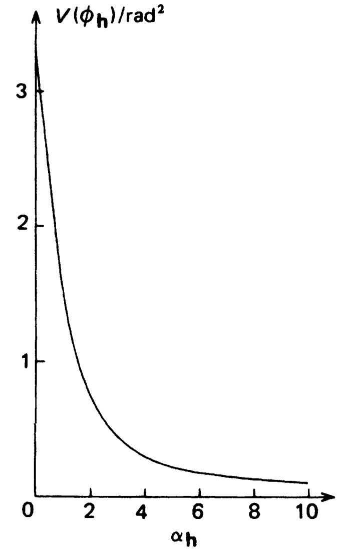


Fig. 8.7 Variance $(V\phi_h)$ as a function of $\alpha_{\mathbf{h}}$

Table 8.5 lists the values of $|\mathrm{E}_{\min}|$ needed to obtain $\alpha_{\mathbf{h}} = 3$ for selected values of N from 25 to 100. The table illustrates clearly an important limitation of direct methods: the required $|\mathrm{E}_{\min}|$ increases dramatically as a function of N, whereas, as indicated earlier, the distribution of $|\mathrm{E}|$ values is largely independent of structural complexity. Therefore, it becomes more and more difficult to form a good starting set as N becomes larger and larger. Calculation of $\alpha_{\mathbf{h}}$ from (8.29) is possible only when phases are available. In the initial stages of phase determination this is not practicable, and the following formula for the expectation value (α_{E}^2) of $\alpha_{\mathbf{h}}^2$, which uses only the values of $\kappa_{\mathbf{hk}}$, has been developed:

$$\alpha_{\mathrm{E}}^2 = \sum_{\mathbf{k}} \kappa_{\mathbf{hk}}^2 + \sum_{\mathbf{k}} \sum_{\substack{\mathbf{k}' \\ \mathbf{k} \neq \mathbf{k}'}} \kappa_{\mathbf{hk}} \kappa_{\mathbf{hk}'} \frac{I_1(\kappa_{\mathbf{hk}}) I_1(\kappa_{\mathbf{hk}'})}{I_0(\kappa_{\mathbf{hk}}) I_0(\kappa_{\mathbf{hk}'})} \tag{8.33}$$

where I_0 and I_1 are modified Bessel functions of the zero and first orders, respectively. The function $I_1(\kappa)/I_0(\kappa)$ has the form shown in Fig. 8.8 and may be expressed as the polynomial

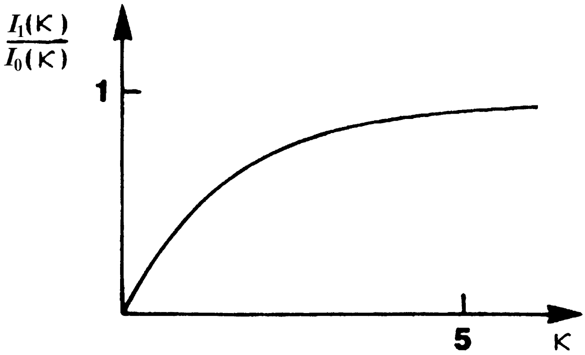
$$I_1(\kappa)/I_0(\kappa) \approx 0.5658\kappa - 0.1304\kappa^2 + 0.0106\kappa^3$$

in the range $0 \le \kappa \le 6$; for larger values of κ the function is essentially unity. These principles, used in conjunction with those discussed earlier for selecting the origin-determining reflections, may help a direct methods analysis to be established on a sound basis right from the beginning, and so lead to a number of sufficiently accurate phases to give an interpretable E map. Experience shows, however, that even with great care, the development of phases may not always be successful. In such an event the remedy is often to try again with a different starting set of reflections.

Table 8.5 Values of $|\mathrm{E}_{\min}|$ for $\alpha_{\mathbf{h}} = 3.0$ in structures containing N identical atoms per unit cell

N	$\lvert\mathrm{E}_{\min}\rvert$
25	1.96
36	2.08
49	2.19
64	2.29
81	2.38
100	2.47

Fig. 8.8 Variation of $I_1(\kappa)/I_0(\kappa)$ with κ



Σ_1-Relationships

In (8.24), let $-\mathbf{k} = \mathbf{h}$; then, since $-\phi_{-\mathbf{h}} = \phi_{\mathbf{h}}$,

$$\phi_{\mathbf{h}} \approx \phi_{-\mathbf{h}} + \phi_{2\mathbf{h}} \approx \phi_{2\mathbf{h}}/2 \tag{8.34}$$

If the structure is centrosymmetric then, in (8.7), $s_{\mathbf{h}}$ is ±1; since $s_{\mathbf{h}}\, s_{-\mathbf{h}} = 1$, it follows that

$$s_{2\mathbf{h}} \approx 1 \tag{8.35}$$

These two Σ_1 relationships, like (8.7) and (8.24) themselves, require large values of both $|\mathrm{E}_{\mathbf{h}}|$ and $|\mathrm{E}_{2\mathbf{h}}|$ for a high probability of their validity.

Multan System

We refer to equations of the MULTAN program system [11, 12] in order to define the weights required in (8.28). Multan employs a modified weighted tangent formula given by

$$\tan\phi_{\mathbf{h}} = \left[\sum_{\mathbf{k}} Q_{\mathbf{h}}, \mathbf{k}\sin(\phi_{\mathbf{k}} + \phi_{\mathbf{h}} - \mathbf{k})\right] \Big/ \left[\sum_{\mathbf{k}} Q_{\mathbf{h}}, \mathbf{k}\cos(\phi_{\mathbf{k}} + \phi_{\mathbf{h}} - \mathbf{k})\right] = T_{\mathbf{h}}/B_{\mathbf{h}}$$

where

$$Q_{\mathbf{h},\mathbf{k}} = \omega_{\mathbf{k}}\omega_{\mathbf{h}-\mathbf{k}}|\mathrm{E}_{\mathbf{k}}||\mathrm{E}_{\mathbf{h}-\mathbf{k}}|/(1 - |U_{\mathbf{h}}|^2)$$

where the unitary structure factor $|\mathrm{U}_{\mathbf{h}}|$ is defined by

$$|U_{\mathbf{h}}| = |\mathrm{F}_{\mathbf{h}}| \Big/ \left(\varepsilon^{1/2} \sum\nolimits_j g_j\right)$$

and j varies from 1 to N, the number of atoms in the unit cell; g_j is f_j corrected for temperature; $\omega_{\mathbf{k}}$ is given by

$$\omega_{\mathbf{h}} = \tan[\sigma_3\sigma_2^{-3/2}|\mathrm{E}_{\mathbf{h}}|(T_{\mathbf{h}}^2 + B_{\mathbf{h}}^2)]$$

and σ_n is defined by (8.31). Thus, each phase assignment carries a weight designed to ensure that poorly determined phases have but little effect on the generation of new phases, while inclusion of all phases ensures an efficient propagation of phase information throughout the data set.

8.2.9 Enantiomorph Selection

In non-centrosymmetric space groups, such as $P2_1$ and $P2_12_12_1$, that contain no inversion symmetry, but can exist in enantiomorphous forms, it is always possible to specify two enantiomorphic arrangements of the atoms in the structure that will lead to the same values of |F|. For example, in the structure in Fig. 1.8, which has two molecules per unit cell in space group $P2_1$, the two enantiomorphic arrangements would be related by inversion through the origin, and will be referred to as the structure (S) and its inverse (I). From the structure factor theory discussed earlier, we can write

$$\mathrm{F}(\mathbf{h})_S = A(\mathbf{h})_S + \mathrm{i}B(\mathbf{h})_S$$

for the structure, and

$$\mathrm{F}(\mathbf{h})_I = A(\mathbf{h})_I + \mathrm{i}B(\mathbf{h})_I \qquad (8.36)$$

for its inverse. From the inversion relationship, we know that $\mathrm{F}(\mathbf{h})_S$ and $\mathrm{F}(\mathbf{h})_I$ are complex conjugates; hence,

$$A(\mathbf{h})_S = A(\mathbf{h})_I$$

and

$$B(\mathbf{h})_S = -B(\mathbf{h})_I \qquad (8.37)$$

For either the structure or its inverse, we can choose $B(\mathbf{h})$ to be positive, so that the corresponding phase angle $\phi(\mathbf{h})$lies in the range $0 \le \phi(\mathbf{h}) \le \pi$. This procedure was followed in the structure analysis of tubercidin, Sect. 8.2.10, where the phase of symbolic reflection $a(13\bar{8})$ was restricted to a value between 0 and π, specifically $3\pi/4$.

In $P2_12_12_1$, a space group of frequent occurrence in practice, the zonal reflections $0kl$, $h0l$, and $hk0$ are centric, and may be given phases equal to $m\pi/2$, since no 2_1 axis passes through the origin. The value of m takes the same parity as the Miller index following zero, working in a cyclic manner. Thus, an origin and an enantiomorph could be specified in this space group by the selection

$$\begin{array}{ll} \left.\begin{array}{rrrr} 5 & 2 & 0 & +\pi/2 \\ 0 & 1 & 1 & +\pi/2 \\ 11 & 3 & 0 & +\pi/2 \end{array}\right\} & \text{Origin} \\ \begin{array}{rrrr} 11 & 0 & 0 & +\pi/2 \end{array} & \text{Enantiomorph} \end{array}$$

Table 8.6 Crystal data for tubercidin

Formula, M_r	$C_{11}H_{14}N_4O_4$, 266.26
Space group	$P2_1$
a (Å)	9.724(9)
b (Å)	9.346(11)
c (Å)	6.762(10)
β (°)	94.64(10)
V_c (Å^3)	612.52
D_m (g cm^{-3})	1.449
D_c (g cm^{-3})	1.444
Z	2
F(000)	280

A detailed practical treatment on the origin and enantiomorph for all space groups has been given by Rogers [13]. It is important not to confuse the specification of the enantiomorph with the selection of the absolute configuration of a structure: in both cases the same type of space group is involved. Selection of the enantiomorph is essential to a correct application of direct methods to a structure with an enantiomorphous space group. However, the derived solution of the structure may correspond to either the absolute configuration or its inverse. This dilemma has to be resolved by further tests, usually involving anomalous scattering, Sect. 7.6. We discuss enantiomorph selection and related topics in Appendix E.

8.2.10 Phase Determination in Space Group $P2_1$

The method of symbolic addition can be used for phase determination in non-centrosymmetric crystals, but it can be somewhat laborious because of the general nature of $\phi(hkl)$. The structure of tubercidin was determined by Stroud [14] using this method: Table 8.6 lists the crystal data for this compound.

In space group $P2_1$, $|\mathrm{E}(hkl)|$ has the following symmetry equivalence:

$$|\mathrm{E}(hkl)| = |\mathrm{E}(\bar{h}k\bar{l})| = |\mathrm{E}(h\bar{k}l)| = |\mathrm{E}(\bar{h}\,\bar{k}\,\bar{l})| \tag{8.38}$$

The phases of the symmetry-related reflections in this space group are also linked, but in a different way, according to the parity of k:

$$\begin{aligned} k = 2n: \quad \phi(hkl) &= \phi(\bar{h}k\bar{l}) = -\phi(h\bar{k}l) \\ &= -\phi(\bar{h}\,\bar{k}\,\bar{l}) \end{aligned} \tag{8.39}$$

$$\begin{aligned} k = 2n+1: \quad \phi(hkl) &= \pi + \phi(\bar{h}k\bar{l}) = \pi - \phi(h\bar{k}l) \\ &= -\phi(\bar{h}\,\bar{k}\,\bar{l}) \end{aligned} \tag{8.40}$$

The $h0l$ zone of this space group is centric, with the origin transferred to the twofold rotation point at $x = z = 0$; hence, the phases are restricted to 0 or π, according to the arguments developed in Sect. 8.2.2, so that permitted origins here are 0, 0, 0; $\frac{1}{2}$, 0, 0; 0, 0, $\frac{1}{2}$; $\frac{1}{2}$, 0, $\frac{1}{2}$.

Again, the origin was specified by assigning phases to three reflections, following the rules discussed above, and shown in Table 8.7. Next, new phases were determined according

Table 8.7 Origin-specifying phases for tubercidin

hkl	$\|E_{\mathbf{h}}\|$	$\phi_{\mathbf{h}}$
$10\bar{6}$	1.95	0
$40\bar{1}$	2.09	0
$71\bar{4}$	2.45	0

Table 8.8 Course of the phase determination procedure for tubercidin

hkl	$\phi_{\mathbf{h}}$	$\|E_{\mathbf{h}}\|$	Number of numerical or symbolic phases
Origin set; Table 8.7			5
$13\bar{8}$	*a*	2.99	11
206	*b*	2.20	20
790	*c*	2.76	47

Table 8.9 Initial development of phases for tubercidin

h	$\|E_{\mathbf{h}}\|$	$\phi_{\mathbf{h}}$	$\|E_{\mathbf{h}}\|\|E_{\mathbf{k}}\|\|E_{\mathbf{h-k}}\|$
Origin set			
$40\bar{1}$	2.09	0	
$10\bar{6}$	1.95	0	
$71\bar{4}$	2.45	0	9.98
New phases (marked *)			
$40\bar{1}$	2.09	0	
$40\bar{1}$	2.09	0	
$80\bar{2}^*$	2.33	0	10.2
$\bar{1}06^a$	1.95	0	
$71\bar{4}$	2.45	0	
612^*	1.83	0	8.74
First symbol			
$13\bar{8}$	2.99	*a*	
612	1.83	0	
$74\bar{6}^*$	2.20	*a*	12.0
$7\bar{1}\,\bar{4}^b$	2.45	π	
$\bar{1}38^b$	2.99	$\pi + a$	
624^*	2.19	*a*	16.0

[a] Symmetry-related through (8.39)
[b] Symmetry-related through (8.40)

to (8.24) or (8.25). In order to maintain an expected variance $V(\phi_{\mathbf{h}})$, Fig. 8.7, of no more than 0.5 rad^2, the product $|E_{\mathbf{h}}||E_{\mathbf{k}}||E_{\mathbf{h-k}}|$ must be greater than 8.5 for this structure. Two new phases $\phi(80\bar{2})$ and $\phi(612)$ were thus determined from the origin set and added to it; further phases were determined in terms of the symbols allocated as shown in Table 8.8.

Eleven phases were generated in terms of the origin set and symbol *a*, 20 after adding symbol *b*, and 47 after adding the third symbolic phase *c*. Table 8.9 illustrates the initial stages of this process. The criteria for accepting a phase were as follows:

1. That $V(\phi_{\mathbf{h}})$, irrespective of the actual choice for phases *c* and *a* should be less than 0.5 rad^2, no matter how many contributors there were to the sum in (8.25); symbol *b*, in parity group eee, is a structure seminvariant with phase 0 or π with respect to the permitted origins.

Table 8.10 Relationships between letter symbols

Form of relationship	Number of indications
$c = \pi + 2a$	7
$c = \pi + 3a$	15
$c = \pi + 4a$	19
$c = 3a$	5
$c = 4a$	2
$c = -3a$	4 *or* 5
$a = 0$	2
$a = \pi$	2
$b = 0$	Many
$b = \pi$	None

2. That where there were two or more different indications for a phase, the phase would be accepted only when indications of one type predominated strongly. Manual phase determination using a Σ_2 listing indicated relationships such as

$$\phi(63\bar{3}) = c - 2a - b \tag{8.41}$$

and

$$b = 0 \tag{8.42}$$

which were strong because they both come from 6 and 3 multiple indications, respectively. By reiteration of the phase addition procedure above, the results in Table 8.10 indicate relationships between a and c. Bearing in mind that the objective is to obtain a self-consistent set of phases, it is well to consider how this might now be achieved. Refinement of phases could in principle be achieved by application of (8.28). However, this would be possible only if numerical values for a and c (taking b as zero) were available. Alternatively, if a working formula relating a and c could be found, (8.28) could be implemented by substitution of values for one symbol only. Table 8.10 shows that there were 41 indications that

$$c = \pi + pa \tag{8.43}$$

where the value for p was chosen as the weighted average of the first three indications for c in Table 8.10. Thus,

$$c = \pi + 3.29a \tag{8.44}$$

The symbol a was then limited to the range

$$0 < a < \pi \tag{8.45}$$

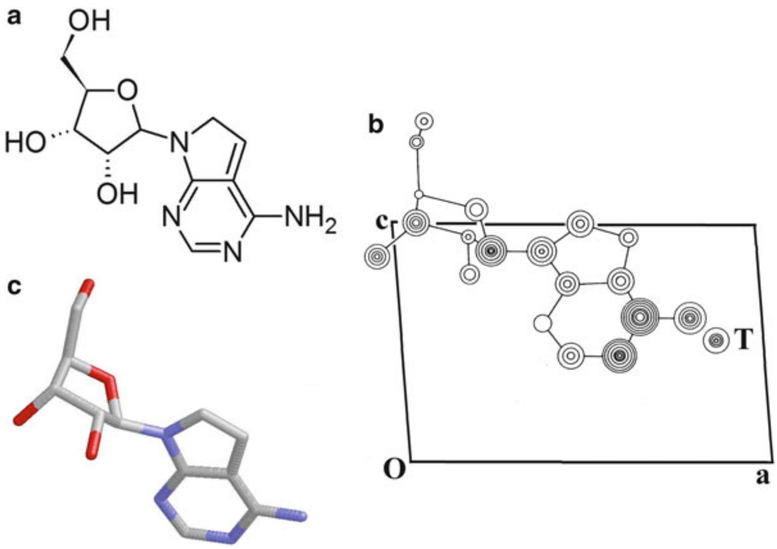


Fig. 8.9 Tubercidin, $C_{11}H_{14}N_4O_4$. (**a**) Structural formula in approximately the same orientation as in the E map. (**b**) Composite E map, with idealized contours drawn at arbitrary equal intervals. Some peaks are heavier than others because of the limited data set used; peak *T* was the only significant spurious peak. (**c**) Molecular structure derived from (**b**), and drawn with RASMOL

in order to fix the enantiomorph, Sect. 8.2.9. Values for a were chosen such that

$$a = n\pi/8 \quad (n = 1, 2, \ldots, 8) \tag{8.46}$$

and converted into phases by the above relationships. Each set was expanded and refined by (8.28), taking $w_{\mathbf{h}} = 1$, for up to 419 reflections with $|E_{\min}| \geq 1.0$. Some phases were rejected because of inconsistencies in their phase indications. An interpretable E map was obtained using the refined phase set with $a = 6\pi/8$; a composite diagram is given in Fig. 8.9.[2]

8.2.11 Advantages and Disadvantages of Symbolic Addition

Symbolic addition has several advantages and disadvantages, which we summarize as follows:

Advantages

1. The user is in control throughout the analysis, and has the responsibility of making sure that all formulae, including symmetry relationships, are applied correctly.
2. The user can make decisions regarding criteria of acceptance of phase indications, the number of |E| values to include, the number of symbolic phases, the choice of starting set, and so on.

[2] See also Bibliography, Ladd and Palmer (1980).

Disadvantages

1. The analysis can be carried out only by a specialist in crystallography.
2. The procedure is slow, requiring many hours of preparation before meaningful results emerge.
3. If a large number of symbols are required, many phase sets will be produced, each of which requires refinement by the tangent formula.

Not surprisingly, alternative rapid and more automatic methods of applying direct methods formulas were sought in the late 1960s, leading to development of the multi-solution methods starting with the program MULTAN, now superseded by other systems, particularly SHELX which we discuss in a subsequent section. Where automatic methods fail, and sometimes they do, it is important to recall how structures like tubercidin were solved by an intelligent manual approach.

8.2.12 Signs of Trouble, and Past Remedies When the Structure Failed to Solve

We summarize some of our own experiences, and record factors involved when a structure solution failed to emerge from direct phase-determining methods. Although these comments apply primarily to earlier program systems, similar considerations apply to other comparable procedures:

1. *Too few* Σ_2 *interactions are being used.* Increase the number of interactions by lowering $|E_{\mathrm{min}}|$.
2. *The origin-defining set is poor or incorrect*. Try another one, choosing it with the aid of the rules given.
3. *The* E *map contains one very large peak*. The phases are probably very inaccurate; the heavy peak may be located in the center of a closed ring. Start again; do not waste time trying to interpret the E map.
4. *The* E *map is not interpretable or chemically sensible*. The are incorrect. Try again.
5. *If heavy atoms are present in the structure, they alone may show up*. Proceed to Fourier methods using interpretable heavy-atom sites. A check against the Patterson function might prove useful here.
6. *Only a small molecular fragment is discernible on the* E *map*. Try recycling basing phases on the fragment found, or try to obtain more phases by increasing the initial data set.
7. *The program selects an incorrect or poor starting set, such as too few* Σ_2 *interactions*. Select your own starting set. If you suspect that a program may contain a fault, inform the author; do not attempt to correct it.
8. *The solution still fails to emerge*. Review the calculation of the $|E|$ values; perhaps omit reflections that appear to have a bad influence on the phase-determining pathway.
9. *All fails*. Go back to fundamentals. Check the space group, data collection, processing, and any other factor that might be at fault.

If you exhaust these possibilities without achieving success, try another method for determining the structure. Or give it a rest and try again later—or study recent Bibliography on direct methods.

8.2.13 Triplets, Quartets, and the SHELX Program Strategy [15]

We discuss in Appendix E and Sect. 8.2.2 entities known as structure invariants and structure seminvariants. As we have shown, they may comprise one or more reflections, most often three or four in practice, and have important properties that are related to the choice of origin, or permitted origin when symmetry is present. In this section, we examine how these invariants may be used to enhance direct methods of phase determination. The structure factor equation has been expressed as

$$\mathrm{F}(hkl) = \sum_{j=1}^{N} g_j \exp[\mathrm{i}\,2\pi(hx_j + ky_j + lz_j)] = |\mathrm{F}(hkl)| \exp \mathrm{i}\,\phi(hkl) \tag{8.47}$$

where N is the number of atoms in one unit cell. In the following discussions it will generally be assumed, for simplicity, that we are dealing with an equal-atom structure.

Referring to Sect. 8.2.2 and Fig. 8.1, it can be seen that changing the origin to any point (Δx, Δy, Δz), changes each atom coordinate to ($x_j - \Delta x$, $y_j - \Delta y$, $z_j - \Delta z$), and that the structure factor in (8.47) will change to a new structure factor $\mathrm{F}'(hkl)$, given by

$$\mathrm{F}'(hkl) \sum_{j=1}^{N} \{g_j \exp[\mathrm{i}2\pi(hx_j + ky_j {+} lz_j)]\}\{\exp[-\mathrm{i}2\pi(h\Delta x + k\Delta y + l\Delta z)]\} \tag{8.48}$$

where $\{\exp[-\mathrm{i}2\pi(h\Delta x + k\Delta y + l\Delta z)]\}$ is a term external to the summation, so that

$$\mathrm{F}'(hkl) = \mathrm{F}(hkl)\{\exp[-\mathrm{i}2\pi(h\Delta x + k\Delta y + l\Delta z)]\} \tag{8.49}$$

Thus, changing the origin to the point (Δx, Δy, Δz) causes a change in $\varphi(hkl)$ given by

$$\Delta\phi(hkl) = -2\pi(h\Delta x + k\Delta y + l\Delta z) \tag{8.50}$$

$|\mathrm{F}(hkl)|$ is, of course, invariant, as is the intensity of reflection, $I(hkl)$. If we now consider the structure factors for two reflections where the two reflections have indices $h_1k_1l_1$ and $h_2k_2l_2$, we show in Appendix E that for a product of two structure factors, in the absence of any symmetry, if $\mathbf{h}_1 + \mathbf{h}_2 = 0$, then it follows that the phase sum $(\varphi_1 + \varphi_2)$ is a structure invariant. This result can be generalized for three and four (or more) reflections, thus:

$$(\phi_1 + \phi_2 + \phi_3) \text{ is a structure invariant if } \mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0 \tag{8.51}$$

$$(\phi_1 + \phi_2 + \phi_3 + \phi_4) \text{ is a structure invariant if } \mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 + \mathbf{h}_4 = 0 \tag{8.52}$$

The presence of symmetry normally requires that the origin be chosen on a symmetry element. Thus, in $P\bar{1}$, for example, we have:

$$(\phi_1 + \phi_2 + \phi_3) \text{ is a structure invariant if } \mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0 \tag{8.53}$$

but if $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$ modulo (2 2 2) the corresponding phase sum is a structure seminvariant. We discuss these invariants more fully in Appendix E.

Notice that the individual phases in these expressions are not themselves invariant, and that nothing has been said so far about the actual values of these phase sums. For (8.51), we can show that we have assumed tacitly, Sect. 8.2, that, for sufficiently large values of α, that is, of $2N^{-1/2}|\mathrm{E}_{\mathbf{h}_1}||\mathrm{E}_{\mathbf{h}_2}||\mathrm{E}_{\mathbf{h}_3}|$,

$$\phi_1 + \phi_2 + \phi_3 \approx 0 \tag{8.54}$$

Rearranging (8.54) gives:

$$\phi_1 \approx -\phi_2 - \phi_3 \tag{8.55}$$

We make a small change in notation, so as to accord with that used in the SHELX-97 program system and to focus on determining $\varphi_{\mathbf{h}}$. Let $\mathbf{h}_1 = \mathbf{h}$, $-\mathbf{h}_2 = \mathbf{k}$, $-\mathbf{h}_3 = \mathbf{h} - \mathbf{k}$; then if Friedel's law holds, that is $\phi(hkl) = -\phi(\bar{h}\,\bar{k}\,\bar{l})$, then (8.55) becomes

$$\phi_{\mathbf{h}} \approx \phi_{\mathbf{k}} + \phi_{\mathbf{h}-\mathbf{k}} \tag{8.56}$$

which is the phase addition formula (8.24) from which the tangent formula (8.28) was developed as a general means of exploiting TPRs for expanding and refining phases by direct methods. Equation (8.52) represents a four-phase *quartet*, and like three-phase triplets, may be a source of phase information, since it can be shown [16, 17] that for a sufficiently large value of the expression $2N^{-1}|\mathrm{E}_{\mathbf{h}_1}||\mathrm{E}_{\mathbf{h}_2}||\mathrm{E}_{\mathbf{h}_3}||\mathrm{E}_{\mathbf{h}_4}|$

$$\cos(\phi_1 + \phi_2 + \phi_3 + \phi_4) \approx +1 \tag{8.57}$$

provided that $|\mathrm{E}_{\mathbf{h}_1+\mathbf{h}_2}|$, $|\mathrm{E}_{\mathbf{h}_1+\mathbf{h}_3}|$, and $|\mathrm{E}_{\mathbf{h}_1+\mathbf{h}_4}|$ are all large. Alternatively,

$$\cos(\phi_1 + \phi_2 + \phi_3 + \phi_4) \approx -1 \tag{8.58}$$

if $|\mathrm{E}_{\mathbf{h}_1+\mathbf{h}_2}|$, $|\mathrm{E}_{\mathbf{h}_1+\mathbf{h}_3}|$, and $|\mathrm{E}_{\mathbf{h}_1+\mathbf{h}_4}|$ are all small. Equations (8.57) and (8.58) are *positive quartets* and *negative quartets* (NQRs), respectively. Thus, the sum of four phases is dependent not only on the intensities of the four corresponding reflections, the *primary* terms, but also on the intensities of three other index-related reflections, the *cross-terms*. As in the treatment for triplets, we need to cast the indices of the phase-quartet in terms of a target reflection **h**. The quartet phase sum then becomes

$$(\phi_{\mathbf{h}} + \phi_{-\mathbf{k}} + \phi_{-\mathbf{l}} + \phi_{-\mathbf{h}+\mathbf{k}+\mathbf{l}}) \tag{8.59}$$

where $\mathbf{h} = \mathbf{h}_1$, $-\mathbf{k} = \mathbf{h}_2$, $-\mathbf{l} = \mathbf{h}_3$, and $-\mathbf{h} + \mathbf{k} + \mathbf{l} = \mathbf{h}_4$. The primary terms are $|\mathrm{E}_{\mathbf{h}}|$, $|\mathrm{E}_{\mathbf{k}}|$, $|\mathrm{E}_{\mathbf{l}}|$, and $|\mathrm{E}_{-\mathbf{h}+\mathbf{k}+\mathbf{l}}|$, which should all be large; the cross terms are $|\mathrm{E}_{\mathbf{h}-\mathbf{k}}|$, $|\mathrm{E}_{\mathbf{h}-\mathbf{l}}|$, and $|\mathrm{E}_{\mathbf{k}+\mathbf{l}}|$, which should all be large for a positive quartet and small for an NQR.

We have seen in Sect. 8.2.11 that symbolic-addition techniques for phase determination employ a number of initial phase assignments for a relatively small number of reflections. Each phase set is then expanded via the tangent formula, which is based on the use of strong triplets. Only a very small number of such assignments can be expected to converge to produce a phase set that is somewhere near the correct one. To avoid the use of the time-consuming inspection of a large number of E maps, experience has shown that, with large structures, results based only on the consistency of triplet relationships may not discriminate the correct phase set. This failure has led to the use of NQRs, as well as triple-phase relations, for good phase determination.

8.2.14 The SHELX Computer Program System

The SHELX program suite, one of the most popular and widely used set of crystallographic programs, was developed originally from SHELX-76 which was specifically aimed at the determination and refinement of small-molecule structures [18, 19]. The latest version, SHELX-97 [20–22], is available free of charge to academic users by registering online [23]. It is recommended that the program be incorporated by the user into a specialized program system such as WinGX [22] which is downloadable [24]. SHELX and other programs have been integrated into WinGX to produce a comprehensive work-base for small-molecule crystallography. The implementation of the SHELX programs for both large and small molecules is fully covered in the user manual. When using WinGX, the crystallographer has access to the SHELX direct methods structure-solving programs SHELXS-86, SHELXS-97, SHELD, and PATSEE; see Sect. 8.3, which employs a Patterson Search method. The SHELXL programs are relatively easy to use and require only two input files:

- Atoms and instructions
- Reflection data

The instruction file includes initially only unit-cell dimensions, wavelength, atom types and their numbers, space group information, and the type of input data being used. The programs are written so as to handle all space groups in all settings, both standard and nonstandard, and are thus perfectly general. This has been achieved by requiring the coordinates of the equivalent positions in the unit cell to be specified together with a centrosymmetric/non-centrosymmetric flag, and therefore demands a good working knowledge of crystallography.

The reflection data file contains $h,k,l,\ F_o{}^2$ and $\sigma(F_o{}^2)$ as output by the data collection system. If using SHELX with WinGX, the user has access also to graphics and editing facilities and programs for monitoring the quality of the structure analysis prior to deposition and publication of the data. Users are advised to read the operation manuals of both SHELX-97 and WinGX very carefully and to keep an eye open for changes in either system.

8.2.15 The WinGX Program System

WinGX is an MS-Windows system of programs for solving, refining, and analyzing single crystal X-ray diffraction data for small molecules. It provides a consistent and user-friendly Graphical User Interface for some of the best publically available crystallographic programs, and has interfaces to other popular programs such as SHELX-97 and SirWare programs (SIR-97, SIR-2008). Users of WinGX must be registered users of the SHELX and SirWare programs. It is the responsibility of the user to register with the appropriate sources as per instructions; WinGX can be downloaded free of charge [23].

Installation of the system is fully described in the manual. The installation stage includes setting up of environment variables which may need the help of an expert. Once installed and called from the PC desktop shortcut, the menu shown in Fig. 8.10a is displayed. In this example, the project name is BW202W92(R) and the necessary files BW202W92(R).ins and BW202W92(R).hkl are stored in a folder called BW202W92(R) on the C:\ drive. Other menus in WinGX to which we shall refer are shown in Fig. 8.10b, c.

8.2.16 Direct Methods in the Program SHELX-97 for Small Molecules

We are concerned at this point with methods used to determine small-molecule crystal structures using the SHELX-97 system, but hasten to add that this system is in fact now capable of handling both small-molecule and protein structure analyses. For these purposes it contains a number of executable programs, such as those that solve structures by either Patterson or direct methods, carry out detailed least-squares refinement, and locate water molecules of crystallization in proteins. The heavy-atom

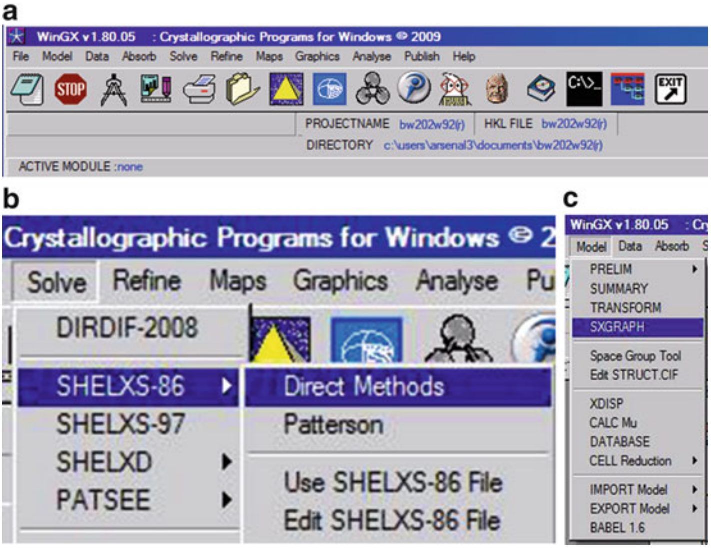


Fig. 8.10 The WinGX program system. (**a**) WinGX Menu showing the *Basic Functions* available. (**b**) Selection of SHELXS-86 direct methods program from the WinGX menu. (**c**) Selection of the graphics facility SXGRAPH in WinGX is activated by the option "Model" on the main menu (**a**); see also Fig. 8.11b, c

method can be invoked, and the Patterson search method used is effectively PATSEE, which we describe in Sect. 8.3.6. The least-squares procedures encompass inter alia dispersion, absorption, and extinction corrections, together with a wide range of constraints and restraints, and routines for fixing and refining hydrogen-atom positions. The direct methods routine is based on a random start multi-solution strategy, or more accurately a multi-permutation single-solution procedure, since it endeavors to identify the correct solution and then to improve on it by E map partial structure extension, or successive refinement. Geometry routines calculate bond lengths, bond angles, and torsion angles and identify possible hydrogen bonds, all of the results being tabulated for publication purposes. The system is strongly recommended to the serious structure analyst.

The triple-phase and NQR relationships discussed above are employed in the SHELX-97 to generate phases by a modified tangent formula. We write

$$\boldsymbol{\alpha} = 2N^{-1/2}|\mathrm{E}_{\mathbf{h}}|\mathrm{E}_{\mathbf{k}}\mathrm{E}_{\mathbf{h-k}} \tag{8.60}$$

$$\boldsymbol{\eta} = gN^{-1}|\mathrm{E}_{\mathbf{h}}|\mathrm{E}_{\mathbf{k}}\mathrm{E}_{\mathbf{l}}\mathrm{E}_{\mathbf{h-k-l}} \tag{8.61}$$

where $\mathrm{E}_{\mathbf{j}} = |\mathrm{E}_{\mathbf{j}}|(\cos\varphi_{\mathbf{j}} + \mathrm{i}\sin\varphi_{\mathbf{j}})$, $(\mathbf{j} = \mathbf{k}, \mathbf{l})$, and g is a positive constant set by the program to account for the cross-term $|\mathrm{E}|$ values; (8.60) and (8.61) are subject to the same conditions as (8.59). It has been found that computer time is optimized for cases where the number of NQRs is restricted to between 1000 and 8000. Only the most reliable relationships are retained in this process, strictly where *all three* cross-terms have actually been measured and found to be weak. However, all interconnecting triple-phase relationships are used, except for Σ_1 terms, Sect. 8.2.8, and those which all involve restricted phases that prevent the resultant phase from being zero [25].

A process of *phase annealing*, based on a principle similar to that of simulated annealing used in the refinement of macromolecular structures, Sect. 10.9.1, is employed in the next stage of phase refinement. The results from these two stages are then applied to a full tangent formula refinement for the best retained reflections. The total number of different attempts using these procedures can be set by the user and may be as many as 5000 for really difficult structures.

Initial Stages

In the initial stages of applying SHELX-97 to a structure determination, a number of cycles of weighted tangent formula (8.28) are performed, starting with a selected number of randomly generated phases. The best phase sets, as judged by NQR and triplet consistency, are retained and the process repeated to give a number of parallel-generated phase sets. After each iteration, the total number of phase sets processed in parallel is reduced until only 25% of the original number of phase sets is generated. Typically 8 or 16 best phase sets are retained from each cycle for a total run of 128 parallel permutations. The best reflections and strongest TPRs are retained and passed to the next stage of the procedure. The program uses a TPR figure of merit (FOM) based on the indicator NQUAL (a negative quantity), where the best phase set has the smallest value:

$$\mathrm{NQUAL} = \Sigma|\boldsymbol{\alpha} \cdot \boldsymbol{\eta}|/\Sigma|\boldsymbol{\alpha}||\boldsymbol{\eta}| \tag{8.62}$$

Phase Annealing

This method is used to refine the phase sets retained after the initial stages. Phase annealing supplies a correction to the phase produced by the tangent formula. As only a limited number of reflections are involved, the process uses computer time efficiently, employing only the strongest triplets and quartets.

Final Stages in Phase Determination

The total number of direct methods attempts (the variable *np* on the TREF instruction in SHELX-97) to be employed may have to be a very high value for difficult structures. The routines in SHELX-97 are written with efficiency in mind and the program requires only a few minutes on a PC for calculations on quite large structures, even when TREF is as high as 5000. The program will select the number n_E of reflections to be involved at this stage if not otherwise stipulated, but n_E may have to be reset to a higher value if the program fails to produce an interpretable E map. An example of the use of this program in solving a difficult structure (cyclosporin H) is given in Sect. 11.10.

8.2.17 Example of a SHELX-97 Structure Solution: Crystal Code Name BW202W92(R)

The SHELXS-97 direct determination described here was undertaken as part of the study of the mesylate salts of a novel voltage-gated sodium channel-binding ligand *R*-(−)-BW202W92 and its much less active *S*-(+)-enantiomer (BW203W92) (Fig. 8.11a) in order to determine the absolute configuration of each, a factor which is of vital importance to a full understanding of the biological activity of the drug. In addition each enantiomer exists as two distinct *atropisomeric* forms in the solid state, i.e., each has two independent molecules per asymmetric unit with slightly different configurational properties [26]. The solution for the *R*-form is described below.

Crystals of BW202W92(R) mesylate salt, $C_{11}H_9Cl_3FN_4{\cdot}CH_3SO_3$ Fig. 8.11a (I), are monoclinic in space group $P2_1$, with $a = 8.384(2)$ Å, $b = 16.984(3)$ Å, $c = 12.480(3)$ Å, $\beta = 104.14(2)°$, and with two independent molecules of each type per asymmetric unit. The X-ray data collected with Cu $K\alpha$ radiation had 3192 independent *hkl* data to a resolution of 0.800 Å ($\theta_{max} = 74.42°$). The files BW202W92(R).ins and BW202W92(R).hkl were stored in the folder BW202W92(R) which was read into WinGX as described in the previous section, Fig. 8.10a. The program SHELXS-86 was selected by using the "Solve" option on the WinGX menu, Fig. 8.10b. The resulting E map from this calculation, obtained using model/SXGRAPH, Fig. 8.10c, from the WinGX menu, is shown in Fig. 8.11b.

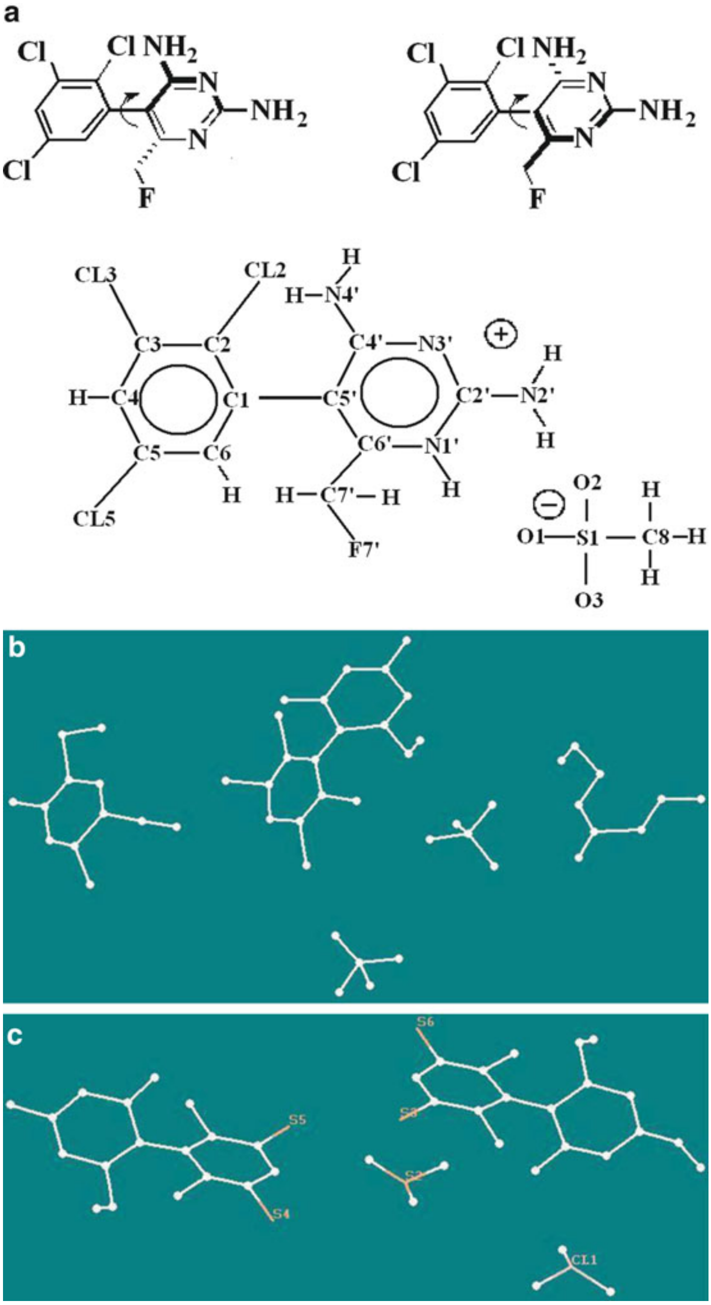


Fig. 8.11 (**a**) *Top*: schematic chemical structure for the sterically hindered rotamers of I and II, respectively. *Bottom*: general numbering scheme used for the enantiomeric compounds (*R*-form = BW202W92, *S*-form = BW203W92). Note that the pyrimidine ring is N1′-protonated in each of the mesylate acid addition salts. (**b**) E map of BW202W92(R) from SHELXS-86. The structure is essentially correct but the heavy atoms Cl and S are not identified and one molecule is fragmented, that is, the required symmetry operation to join the two halves up has not been applied. (**c**) E map of BW202W92(R) from SHELXS-97. Heavy atoms are recognized as such but apart from one S atom are incorrectly atom-typed. The structure is essentially correct apart from this fault which was easily rectified

The program SHELXS-97 was then implemented in a similar manner and resulted in the map shown in Fig. 8.11c. The input data for the SHELXS-97 run is shown in Table 8.11a and the output data in Table 8.11b, both with explanatory notes.

In all, 256 phase sets were refined; the best fit corresponds to the value 0.0503, a very promising and acceptable result. However, in the final analysis it is the electron density map and subsequent refinement which provide the final proof of correctness of the direct phase set. The R_E value for this best phase set prior to examination of the electron density had a value of 22.4% which is very satisfactory at that stage.

Table 8.11a Input data for crystal BW202W92(R) to SHELXS-97

Input file BW202W92(R).ins

```
TITL          BW202W92(R)
CELL          1.54180 8.385 16.984 12.481 90.000 104.15 90.000
ZERR          4.00 0.002 0.003 0.003 0.000 0.02 0.000
LATT          −1
SYMM          − X, 1/2 + Y, - Z
SFAC          C H O N CL S F
UNIT          48 48 12 16 12 4 4
HKLF          3
END
```

Explanation of input file BW202W92(R).ins

TITL	Title
CELL	Wavelength and unit-cell parameters
ZERR	Number of chemical formula units per unit cell, with esds
LATT	Non-centrosymmetric, -1 (centrosymmetric, 1)
SYMM	Equivalent position(s) related to x,y,z (space group is $P2_1$)
SFAC	Atom types in the structure
UNIT	Number of each atom type in the unit cell (in the order of the previous line)
HKLF	Data input order: h,k,l, F_o, and $\sigma(F_o)$

Note: This HKLF input is not the recommended form. The more usual input data would be HKLF 4 corresponding to h,k,l, $F_o{}^2$, and $\sigma(F_o{}^2)$. These data were collected on a CAD4 diffractometer, which produces output data as F_o and $\sigma(F_o)$; thus, the input is restricted to HKLF 3. However, refinement is still carried out on $|F|^2$

Input file BW202W92(R).hkl

```
−8    0    0    43.79    1.30    1
−8    1    0     8.19    2.75    1
−8    3    1    11.18    2.59    1
−8    1    1    22.52    1.67    1
−8    0    2    48.56    1.28    1
−8    2    2    29.63    1.57    1
−8    3    2    10.44    2.63    1
−8    4    2    13.78    2.03    1
−8    4    3     7.48    4.08    1
−8    3    3    20.62    1.65    1
−8    2    3    13.71    2.33    1
```

. . .and so on, to the end of the *hkl* data set. At the end of each line is the batch number, which is usually 1 for contemporary data measurement schemes. A total of 3192 *hkl* values formed this data set, corresponding to a resolution of 0.800 Å

It is of interest to compare the resulting E maps from SHELXS-86 (Fig. 8.11b) and SHELXS-97 (Fig. 8.11c). The SHELXS-97 program recognizes that the structure contains the heavy atoms Cl (chlorine) and S (sulfur) and has attempted to assign the highest peaks in the electron density calculated from the best phase set as either Cl or S (Fig. 8.11c). From Fig. 8.11c we can confirm that the program has in fact assigned the correct peaks as heavy atoms but has not been able to correctly identify all of the atom types: only one S atom has been correctly assigned as such. The earlier program SHELXS-86, the results of which are shown in Fig. 8.11b, did not attempt to make these heavy-atom assignments at all, and in addition did not apply the necessary symmetry relationship needed to bring the two halves of one of the molecules together. In both respects SHELX-97 is the superior program. Figure 8.11c shows the peaks generated by SHELXS-97 and displayed by the WinGX facility model/SXGRAPH, which has the following features: three-dimensional rotation, translation, magnification, and demagnification of the set of peaks displayed; calculation of bond lengths, bond angles and torsion angles for selected peaks; deletion of unwanted peaks; assignment of individual

Table 8.11b Selected output from the SHELX-97 direct methods calculations with WinGX

Summary of parameters for BW202W92(R)

```
ESEL |E|min 1.200 |E|max 5.000 DelU 0.005 renorm 0.700 axis 0
OMIT s 4.00 2theta(lim) 180.0
INIT nn 15 nf 16 s+ 0.800 s- 0.200 wr 0.200
PHAN steps 10 cool 0.900 Boltz 0.300 ns 282 mtpr 40 mnqr 10
TREF np 256. nE 420 kapscal 0.800 ntan 3 wn -0.750
FMAP code 8
PLAN npeaks -65 del1 0.500 del2 1.500
MORE verbosity 1
TIME t 9999999.
```

This first part of the output consists of various parameters which in this case have been set by default. For difficult structures some manual intervention may be necessary

```
282 Reflections                                   4356 unique TPR for phase annealing
420 Phases refined, using                         12351 unique TPR used
443 Reflections                                   13782 unique TPR for R(alpha)
7416 Unique negative quartets found               3768 used for phase refinement
Highest memory used to derive phase relations     5424 / 71182
```

This second selected part of the output lists some information about the direct methods results, including how many triplet phase relationships (TPRs) have been found. Valid phases have been calculated from a random starting phase set using TPRs and NQRs

Measure of fit results

FOM range	Frequency
0.000 - 0.020	0
0.020 - 0.040	0
0.040 - 0.060	184
0.060 - 0.080	0
0.080 - 0.100	2
0.100 - 0.120	15
0.120 - 0.140	4
0.140 - 0.160	3
0.160 - 0.180	6
0.180 - 0.200	5
0.200 - 0.220	1
0.220 - 0.240 and so on	

atom numbers, types and thermal parameters; saving of the present model as the new .ins file (for refinement); and initiation of structure refinement. After completion of the refinement, programs within WinGX assess the quality of refinement, produce files (.cif)[3] for checking and publication, and prepare high quality diagrams, either using the programs ORTEP/RASTER in WinGX or from the .cif[3] file or a . pdb file, using freely downloadable programs such as MERCURY or RASMOL (see Appendix D for references to these programs).

8.3 Patterson Search Methods

In our earlier discussion in this chapter, we showed that a Patterson synthesis must contain a complete set of peaks, that is, $N^2 - N$ non-origin peaks for a crystal containing N atoms in the unit cell. Since we can always calculate the vector set for a model structure, or just a part of it, the Patterson function could be unscrambled, wholly or partially, in terms of a set of atomic coordinates. This idea has led to a technique in structure analysis called Patterson search methods.

[3] Crystallographic Information File.

8.3.1 General Comments for Small Molecules and Macromolecules

We introduce here a method for structure determination that is useful and applicable either to small molecules, for which the method would normally be used in case of failure of direct methods, or to macromolecules, for which the method is now frequently used as a first choice where possible. The method may be designated Patterson Search for small molecules and Molecular Replacement (MR) in the case of macromolecules. In order to apply the method to solve a structure, the initial requirements are:

1. For the crystal under investigation, the *target structure*, a set of $F_o(hkl)$ data to as high a resolution as possible, which would normally be atomic resolution for small molecules, but generally less for a macromolecular crystal because such crystals tend to diffract X-rays weakly, Sects. 4.1 and 10.4.7
2. The availability of the coordinates of a good quality structure, the *search structure*, which forms a relatively small fragment of the target structure for small molecules, whereas with macromolecules the search structure should ideally be similar in size and structure to the target molecule and/or with at least 40% sequence homology
3. A sound understanding of the principles and practices involved
4. State-of-the-art software and hardware

Conceptually the basic principle of Patterson Search or Molecular Replacement (MR) is quite straightforward. Remember, we know that the intensity data $F_o^2(hkl)$ contain phase information. The MR method as proposed by Rossman and Blow [27] for protein structures involves a critical and quantitative comparison of the Patterson functions of the target and search models. Similar comments apply to a Patterson search method implemented successfully some years ago by Braun et al. in the Vector Verification Method [28] for small molecules. Although quite successful and relatively easy to use, this method has now been superseded by the program PATSEE, which will be described in detail below.

In contemporary software for carrying out MR applications, the method is strengthened through the use of a variety of other lengthy crystallographic techniques, which are now within the capabilities of modern computers. For example, with proteins:

1. In the MR module of CCP4 [29] an automated procedure called MrBump is used
2. Model generation is used with the program Chainsaw
3. The program Phaser is used in CCP4

Similar routines are used with the programs:

4. MOLREP (developed from AmoRe)
5. BALBES
6. X-PLOR
7. CNS

In a similar manner, PATSEE is strengthened by the use of direct methods.

The Patterson function has peaks of high density at locations corresponding to the ends of atom–atom vector pairs, with one atom of each pair occupying the common origin. For complex structures like proteins, the interatomic vectors are densely packed in the unit cell, and most of them will not be resolved in the Patterson map; lack of atomic resolution in the X-ray data will also cause a further blurring of the vector distribution density. The two atoms forming the Patterson vector can be in either the same molecule, *intramolecular*, or between atoms in either symmetry-related molecules or nonsymmetry-related molecules, *intermolecular*.

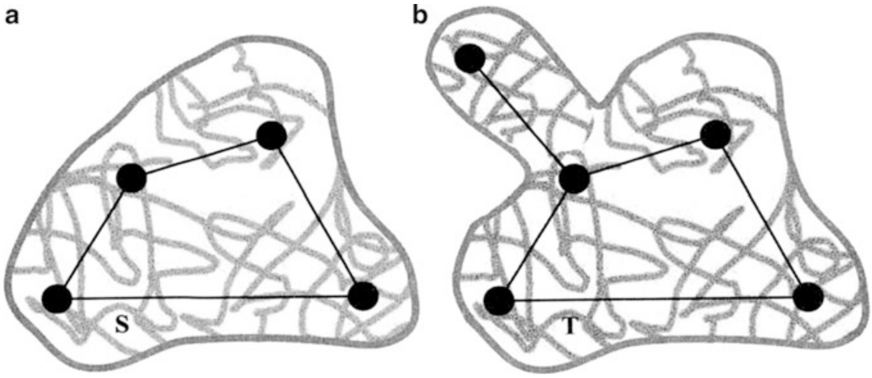


Fig. 8.12 (**a**) Known, search model *S* to be used in Patterson Search or Molecular Replacement. (**b**) The unknown target molecule *T*, that is similar to the search model but has an extra structural feature

8.3.2 Intramolecular Interatomic Vectors and Molecular Orientation

Intramolecular vectors tend to be shorter than the intermolecular vectors that span different molecules in the unit cell. This *self-vector set*, as it is also known, is consequently situated around the origin of the Patterson function, the longest vectors arising between atoms at extreme ends of the molecule. Each atom, in theory, images [30] both the structure and its inverse, that is, vector types AB and BA. Because of the lack of resolution in most protein structures, this will be in the form of a *blurred molecular envelope* of density, whereas for small molecules at atomic resolution the individual images will be much more clearly defined. There will be one such image of the structure, plus its inverse, per atom, forming a centrosymmetric distribution. Figure 8.12 shows two similar (homologous) two-dimensional "molecules," in which *S* represents a suitable search molecule, and *T* the target molecule whose structure is being determined. In the case of a small molecule we can consider the four or five discs to represent resolvable atoms. For macromolecules, the surrounding sheaths in these diagrams represent the overall molecular shape that MR technique is seeking.

Figure 8.13 shows the two actual structures in their different unit cells. In small-molecule structures, where the search and target molecules include a relatively small part, maybe 50%, of the structure in common, we would expect the two unit cells and space groups to be different. For macromolecular structures, which are more similar to each other, this need not be so, although frequently it is. As can be seen in this example, the known search structure *S* is based on a non-orthogonal unit cell, while the unit cell for the target structure *T* is orthogonal. Both cells incorporate twofold symmetry; Fig. 8.13b shows some of the intermolecular interatomic vectors, as discussed in Sect. 8.3.3. For the two molecules *S* and *T*, Fig. 8.14 shows the corresponding resolved *self-vector set* peak positions in the Patterson functions for small molecules and the corresponding simulated vector sheaths representing the envelope of the Patterson function, for macromolecules; for clarity only those peaks *not* related by the Patterson center of symmetry are shown.

Rotation Stage of Patterson Search or Molecular Replacement

In the above simulated example the vectors defining the *orientation* of the unknown structure *T* in its unit cell, Fig. 8.14b, can be seen to occur in the Patterson of the known search structure *S*, Fig. 8.14a. To demonstrate this fact the reader should make a *transparent* copy of Fig. 8.14a, which shows the Patterson peaks of the known molecule *S*, and place it over Fig. 8.13b, the Patterson function of the unknown structure *T*. Locate the two diagrams such that their *origins are in register*, and show that the maximum correspondence occurs for an *anticlockwise rotation* of your copy by 67°.

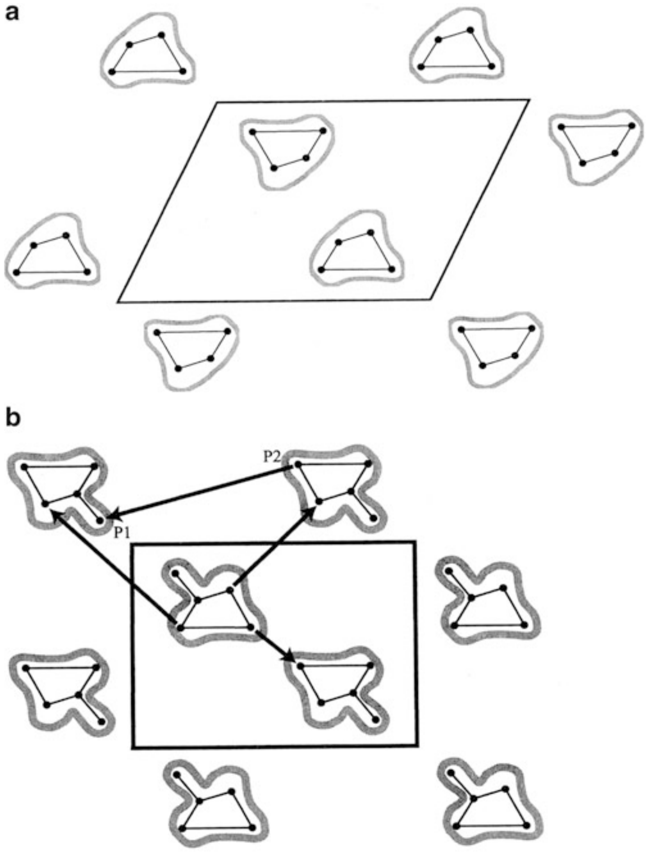


Fig. 8.13 (**a**) The search model *S*, and (**b**) the target molecule *T* packed in their respective unit cells. It is necessary that the structure of *S* is known and that a set of atomic coordinates is available. The structure (**b**) is to be determined by Patterson Search or Molecular Replacement. Some of the intermolecular Patterson vectors are shown in (**b**)

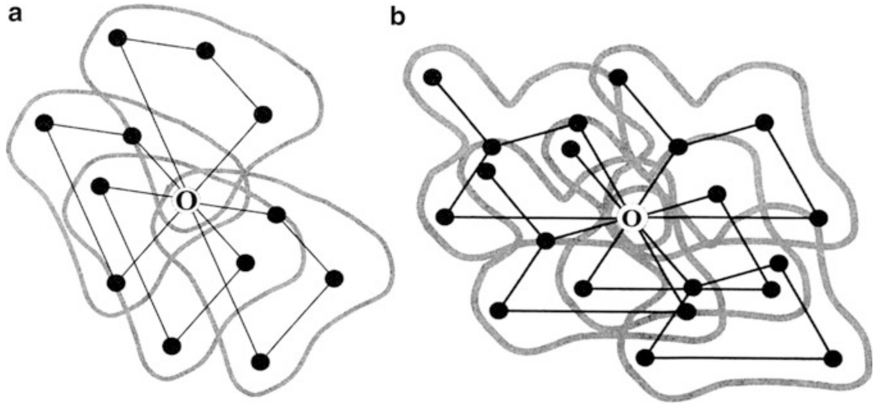


Fig. 8.14 Interatomic vectors for (**a**) the search model *S*, and (**b**) the target molecule *T*. Only those vectors independent of the Patterson center of symmetry are shown. Both diagrams contain multiple images, or part images, of structures. There is a rotational relationship between the two Patterson functions that the reader should try to determine (see text). Note that the sets of vectors shown here are arranged around the origin of the Patterson function at distances corresponding to the distances between atoms within a given molecule. The high vector density which always occurs at the Patterson origin has been omitted

In a real structure determination this orientation angle would be calculated by one of the available computer programs, as described below. Note that because the unknown structure Fig. 8.12b contains an additional moiety, vectors for this part of the structure are missing in the search Patterson, Fig. 8.14a. For small-molecule analysis the missing vectors may form a large fraction of the target molecule, whereas for macromolecules only a small fraction would normally be involved. Subsequent Fourier and least-squares refinement of the target structure, once the search model has been located, will serve both to locate any missing atoms or groups and to eliminate others which may have been included, but which do not form part of the target structure.

In practice the process of matching the orientations of the two Patterson functions, the *rotation stage*, is carried out by computation, testing over a series of orientation angles. For three-dimensional structures, it is necessary to perform these rotations about three independent axes, normally called α, β, and γ—not to be confused with the unit-cell angles. Axial systems used for the rotation procedure differ among different programs, and it is advisable to read the program manual in each case for information on this point (see Appendix D). The angular rotation ranges and intervals have to be chosen carefully in order to cover a sufficient number of possibilities, thus ensuring that the correct angular triplet is not overlooked. Although computationally wasteful, it is better to include too many trials rather than too few. In PATSEE it is not unusual to use several thousand random angle triplets in the rotation stage of the analysis.

On account of the complexity of protein structures, the following two conditions apply to the rotation function:

1. In order to limit the Patterson vectors to lengths that include self-vectors (intramolecular vectors) but exclude cross vectors (intermolecular vectors), the rotation function $R(\alpha, \beta, \gamma)$ should be calculated over a restricted spherical volume U centered at the origin, and having a radius known as the *radius of integration*.
2. The large number of values of $R(\alpha, \beta, \gamma)$, calculated over the required angular range, generally contains many peaks in addition to those that belong to the correct solution, and having comparable magnitudes. These peaks are simply signifying that there is a degree of correspondence between the two Patterson functions in this orientation, albeit a wrong one. Because of this uncertainty, some programs, such as AmoRe, retain all peaks greater than 50% of the highest peak (even more in some programs) for transference to the *translation stage* of MR. The problem is less significant for small molecules but nevertheless must not be overlooked. Figures of merit are used to discriminate between the true and false solutions.

8.3.3 Intermolecular Interatomic Vectors: Translation Stage of MR

Assuming that the correct orientation of the search fragment or molecule has been determined in the rotation stage, the correctly oriented structure must then be located spatially in its true position in the unit cell of the target crystal by means of a *translation stage*. The origin with respect to this translation process is usually governed by the space group. Translation is carried out rigorously by placing the oriented search fragment in a large number of test positions located on a fine grid. This process must cover a sufficient number of finely selected translational increments in three dimensions, designated as either Δt_1, Δt_2, Δt_3, or t_x, t_y, t_z, in order to ensure that the correct location of the molecule is scanned. In essence, the correct structure is recognized initially as corresponding to the highest degree of overlap between the calculated and observed Patterson functions, when superimposed after applying the given translation vector.

During the above process all of the lattice and symmetry operations for the target crystal are fully applied. It is therefore absolutely essential that the space group of the target crystal has been correctly

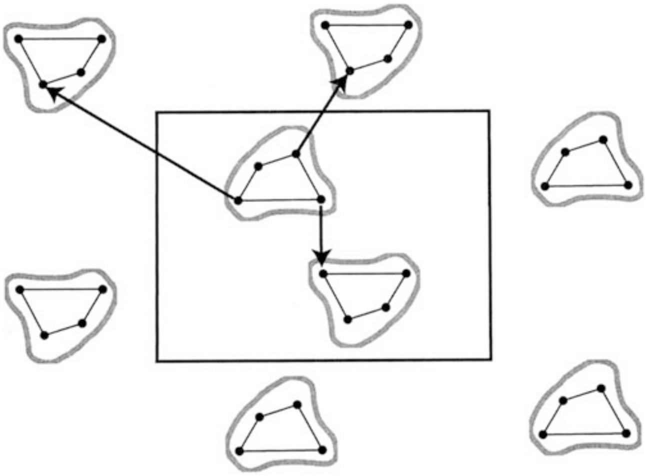


Fig. 8.15 A possible solution for Patterson Search or Molecular Replacement, where the *correct rotation* (or orientation) has been applied to the search model S but the subsequent translation is incorrect (compare with Fig. 8.14b). Three intermolecular vectors are shown for this situation which, because the translation is incorrect, will not be present in the Patterson function for the target molecule. Consequently, there will be a lack of correspondence between the S and T Patterson functions, resulting in poor figures of merit for this solution; such solutions will therefore be rejected

assigned. If this is not the case there will be no outstanding solution at the translation stage, and further refinement of the structure will not be possible. Figure 8.13b, which represents the correct solution for the target structure, includes some of the intermolecular interatomic vectors. For comparison the corresponding vectors are shown in Fig. 8.15, in which the search molecule S is correctly oriented but *incorrectly* translated into the target unit cell. Very large changes in these vectors are evident and would have a profound effect on the quality of this particular translation when tested computationally. The changes in how the molecules are packed (see below), while retaining the twofold symmetry and unit-cell translations in the structure, should be noted by comparing Figs. 8.13b and 8.15.

8.3.4 Crystal Packing and Refinement of the Structure

We consider next the packing of the rotated–translated (RT) molecules in the unit cell. An incorrect RT solution will produce incorrect crystal packing, which can be easily detected from a thorough calculation of interatomic distances. What usually happens is that some atoms in different molecules will be unacceptably close to each other; see Tables 8.21 and 8.22. Evidence of the correctness of the combined RT search results can thus be easily achieved by inspecting the crystal packing generated through each promising set of rotation and translation parameters that has been retained, usually because the figures of merit are encouraging. Crystal packing can be examined readily with the use of a suitable molecular graphics program. For small molecules this can be carried out, for example, in the program package WinGX using the routine PLATON99, while for macromolecules the program MOLPAK has been found to be useful and reliable (see Appendix D). These programs provide not only graphical representations of the packed molecules but also quantitative values for the intermolecular contacts; in particular, any violent clashes between the symmetry-related molecules that would exclude a physically viable solution are flagged.

Expansion and Refinement of the Structure

The final stage of this procedure, as with other methods of structure analysis, attempts to locate atoms that are missing from the model and to refine the positions of all atoms to produce a value for the R-factor, Sect. 7.6.1, that is compatible with the accuracy of the intensity data. As we have seen, missing atoms can be located by calculating the electron density, based on phasing from the partial model. Initially the calculated structure factors will be of low accuracy, but with a reasonably sized fragment, as will be seen in the following examples, the electron density for the correct solution can be expected to reveal at least some of the missing atoms, which are then added to the phasing process and passed on to the next cycle.

Program suites like WinGX have facilities for checking the molecular geometry at each stage, so as to reduce the possibility of adding incorrectly placed atoms to the model. The partial model may be subjected to a few cycles of least-squares refinement, as this can accelerate the procedure toward convergence. However, it should be noted that only when all atoms are present in the model will the R factor be meaningful. This refinement procedure will be designated loosely as "Fourier-least-squares," see Sect. 8.4.2.

8.3.5 Patterson Search Methods for Small Molecules

In the description of the heavy-atom method, Sect. 7.5, we showed that, for a structure containing a small number of relatively heavy atoms, the Patterson function can lead to a successful determination of the atomic coordinates of the heavy atoms, the remaining atoms being located by subsequent calculation of electron density maps prior to refinement. Direct methods can be used to solve light-atom structures, for which there are none of the predominant interatomic vectors that would be necessary for an application of the heavy-atom method. We have seen that the Patterson function necessarily contains the complete set of interatomic vectors for the given crystal structure, in the form of peaks located at the ends of such vectors, one end of each vector being located at the origin. Reference to Fig. 8.15 shows that in effect each atom forms an image of the structure with itself at the origin. When the Patterson center of symmetry is added, there result $2N$ displaced images of the structure present around the origin. A *single weight* peak, that is, one generated by a given pair of atoms 1 and 2, will have a value or peak height roughly proportional to the atomic number product Z_1Z_2.

The total of $N^2 - N$ non-origin Patterson peaks per unit cell containing N atoms increases rapidly with N: the peaks may become overcrowded and overlapping, and unresolvable for larger structures. In this section we describe applications of Patterson methods to approximately equal-atom structures. These structures are usually of relatively low molecular weight, containing 30–40 carbon-like atoms, with crystals that diffract to atomic resolution, leading to well-resolved interatomic vectors in the Patterson function. Macromolecules, which generally diffract far short of atomic resolution, will be discussed in Chap. 10.

The Patterson function for small molecules necessarily contains the complete set of interatomic vectors for the crystal structure. Thinking of the logistics in reverse: for a *known* light-atom structure we could use the atomic coordinates to generate the set of interatomic vector coordinates, and these would match the Patterson function within limits of error that are dictated by the quality of the F_o data, assuming that peak overlap is not a major problem. Since each atom in the structure forms in the Patterson function an image of the structure in itself, Sect. 7.4.2, there are N such images scrambled, or convoluted together, in $P(uvw)$. The complexity of the Patterson function generally makes deconvoluting it in terms of the individual atomic coordinates almost impossible unless some additional information can be used initially. This information is usually in the form of a reasonably precise model of the geometry of a fragment of the molecule whose three-dimensional structure is known.

The method thus depends on a good knowledge of the chemical identity of the molecule in question. Knowing the geometry of such a fragment of the structure, interatomic vectors for the fragment may be calculated and searched for in the Patterson function. Coordinates for the molecular fragment to be used may be derived from the library of known crystal structures in the Cambridge Crystallographic Data Base, or generated from graphics programs, such as the excellent ChemSketch [31], which incorporate standard bond lengths and angles; see Sect. 8.7. For example, a suitable molecular fragment may simply be a benzene ring, which comprises a flat regular hexagon with sides of 1.40 Å and internal angles of 120°. The coordinates for the search model in this case could be obtained by drawing (see Problem 8.9). If the structure contains a cyclohexane ring, this structure is truly three-dimensional, and one of the other two methods mentioned above must be used to derive its coordinates; see also Sect. 13.6.6. The examples discussed below both involve somewhat more complex search fragments.

8.3.6 The Program PATSEE

In the discussion so far, we have established that a Patterson Search in vector space consists of the following stages:

1. Acquisition of a set of accurate atom coordinates for a suitable search model
2. Calculation and storage of the Patterson function (self-vector set) for the model
3. Calculation and storage of the Patterson function for the target crystal, based on the $F_o(hkl)$ data
4. Rotation search, which provides several possible orientations for the search model to occupy in the target unit cell, listed in order of likely feasibility, or precedence
5. Translation search which attempts to place each of the rotation solutions in turn into the correction position in the target unit cell
6. Refinement and expansion of the rotated and translated models, again in order of precedence, to finally converge on the correct solution for the target structure

The program PATSEE follows stages (1)–(5) in the above list, using some important and powerful new algorithms in its implementation. Stage (6) reverts to the standard procedures for structure refinement described in the next section. The PATSEE program is highly recommended, is easy to use, and readily available to crystallographers, being incorporated into the structure analysis suite WinGX. As with other methods that utilize Patterson Search, the program determines the orientation of the search fragment from the rotation function. However, instead of then using the conventional translational function, direct methods are applied in order to determine the position of the oriented fragment in the unit cell. This stage involves the use of several triple-phase invariants, Sect. 8.2.12, selected by the program as being sensitive to the location of the fragment within the unit cell. The weighted sum of the cosines of these phases is maximized with respect to the position of the search fragment in order to determine its most probable location.

As a very large number of possible solutions for the orientation and position of the search fragment in the unit cell are explored, it is necessary to try to pinpoint the correct solution by using a figure of merit (FOM) based upon:

1. The agreement with the Patterson function
2. The triple-phase consistency
3. An R-index between the observed $|E|$, Sect. 8.2.1, and the $|E_c|$ values, Sect. 13.4.10, for the partial structure

The Patterson function is calculated using $|E|F_o$ as coefficients instead of F_o^2 in order to sharpen or improve the resolution of peaks. The complete Patterson function is then stored in the computer, each grid value being represented by a digit between 0 and 7 so that it can be stored in three bits of

Table 8.12 Rotation search in PATSEE

Compilation of the intramolecular self-vector set from the model coordinates with distances between 2 Å, which are too short to provide orientation data, and 6 Å, which is the point at which errors in distance affect the accuracy with which vectors superimpose
Generation of random orientations, typically between 10000 and 60000 angle triplets (α, β, γ) at about 7° intervals
For each orientation, calculation of the correlation between the rotated intramolecular vector set and the Patterson function. This is computed as a sum function to give a figure of merit (FOM) $\mathrm{RFOM} = (1/n)\sum_{i=1}^{n} P_i/w_i \quad (8.63)$ where $n \approx 0.3n_{\mathrm{total}}$, w_i is the calculated vector weight, P_i is the nearest Patterson grid value, and n_{total} is the number of worst-fitting vectors, that is, those with the lowest P_i/w_i value
An overlap or packing test ensures that when symmetry and lattice translations are applied there are no serious interatomic clashes
An equivalence test which excludes similar solutions from being retained
Sorting of the solutions in descending order of RFOM
Refinement of the best solutions, carried out by testing up to 1000 additional random rotations around each retained solution, at approximately 2° intervals

computer memory. This is one step better than the method employed in the Vector Verification, which employed only a two-bit representation. In order to make efficient use of computer memory, the Patterson values are encoded according to seven test levels, with level 2 equal to the median of the cumulative Patterson distribution, the difference between two successive test levels being about half the expected height of the highest single vector. The user can, in fact, supply different test levels from those noted above but it is probably not necessary.

Rotation Search Strategy Used in PATSEE

The rotation search procedure used in PATSEE is summarized in Table 8.12.

Translation Search Strategy Used in PATSEE

The translation search procedure used in PATSEE is summarized in Table 8.13.

8.3.7 Examples of Structure Solution Using PATSEE

The following examples, 5,7-methoxy-8-(3-methyl-1-buten-3-ol)-coumarin and atropine, indicate various other features of the PATSEE program. In particular, the use of the figures of merit RFOM, TPSRSUM, and CFOM is to pinpoint the correct Patterson Search solution.

Structure of 5,7-Methoxy-8-(3-Methyl-1-Buten-3-ol)-Coumarin

The crystal and molecular structure of the antimalarial compound 5,7-methoxy-8-(3-methyl-1-buten-3-ol)-coumarin, $C_{16}H_{18}O_5$, $M_r = 290.3$ has been reported [32]. This molecule, Fig. 8.16a, was selected for investigation of the features of the PATSEE program, because the coumarin moiety, Fig. 8.16b, is known to be fairly rigid and planar (rings A and B) and as such is an ideal search molecule. The presence of potentially more flexible side groups provides an element of challenge to the method. The atomic coordinates of the 11 atoms in the coumarin moiety are readily available from either the published structure [33] or the Cambridge Crystallographic Data Base [34]. In fact, we chose the method of molecular graphics employing the Chem-X package to generate coumarin search model 1. The program Chem-X is no longer available but ChemSketch [31] is a good alternative for this type of work.

Table 8.13 Translation search in PATSEE

Search for the most probable direct methods TPRs

Calculation of |E(*hkl*)| for a given orientation of the search model

Selection of suitable phase relationships from a *relatively small number* of large |E(*hkl*)| data. This strategy considerably speeds up the procedure and especially enhances the efficiency for larger structures. Note that in small-molecule work, the search fragment is of course very incomplete and possibly inaccurate. Nevertheless, if its scattering power $\left(\approx \sum_j Z_j^2\right)$ is large enough, the TPRs used here should hold, at least approximately, for the correct solution, and be nonrandom in character. Hence the importance of having a sufficiently large fragment of the whole structure as the search model, as discussed further in the next section. The reader should refer to Sect. 8.2 to aid the appreciation of these three steps

For each oriented search model from the rotation stage, sets of atom coordinates are generated, the oriented search model being placed at a position in the unit cell generated in a random manner. Each of these sets of coordinates is used for calculation of trial phases and assessment through their agreement with the selected TPRs from stage 1

A packing test is carried out on rotated-translated fragments from stage 4 and a model is eliminated if short intermolecular distances occur

Initial refinement of an RT fragment: this procedure involves optimizing a figure of merit TPSRSUM by fine tuning the position of the fragment:

$$\mathrm{TPSRSUM} = \left[\sum |\mathrm{E}_{\mathbf{h}}||\mathrm{E}_{\mathbf{k}}||\mathrm{E}_{-\mathbf{h}-\mathbf{k}}| \cos(\phi_{\mathbf{h}} + \phi_{\mathbf{k}} + \phi_{-\mathbf{h}-\mathbf{k}})\right] / \left[\sum |\mathrm{E}_{\mathbf{h}}||\mathrm{E}_{\mathbf{k}}||\mathrm{E}_{-\mathbf{h}-\mathbf{k}}|\right] \tag{8.64}$$

the summations being taken over all selected three-phase structure invariants. TPSRSUM is expected to be large and positive for the correct solution, up to a maximum value of 1.0. During this process the step sizes are reduced from around 0.2 to 0.05 Å

A further distance test is then carried out to check the packing of the fragment

Solutions that have survived all of these rigorous tests are further tested against the Patterson function of the target crystal. In earlier Patterson Search programs, such as Vector Verification, this stage was carried out immediately after the rotation stage [87], and consequently more time consuming and less likely to succeed.

In PATSEE the correlation between the Patterson function and the fragment-derived intermolecular vector set is examined by comparing the weight of each vector with the nearest grid value. The fit is measured by calculating a further FOM:

$$\mathrm{TFOM} = (1/n) \sum_{i=1}^{n} P_i / w_i \tag{8.65}$$

where $n \approx 0.2 n_{\text{total}}$, and the other parameters are defined as before

Final selection and ordering of the possible solutions. At this stage a small number of the most promising solutions according to TPSRSUM and TFOM will have been stored in the computer. An R_{E} index is calculated as

$$R_{\mathrm{E}} = \left[\sum (\mathrm{E_o} - |\mathrm{E_c}|)/p\right] / \sum \mathrm{E_o} \tag{8.66}$$

where p^2, equal to $\sum Z_{\text{frag}}^2 / \sum Z_{\text{molecule}}^2$, is the fractional scattering power of the search model (frag) compared to that of the whole model. Only positive terms in (8.66) are considered, as it is assumed that negative terms indicate complete agreement. The solutions are then sorted according to a combined FOM:

$$\mathrm{CFOM} = 0.1(\mathrm{RFOM} + \mathrm{TFOM})(\mathrm{TPRSUM}^{1/2})/R_{\mathrm{E}} \tag{8.67}$$

Search Model 1 from Chem-X

The material crystallizes in the monoclinic space group $P2_1/c$ with four molecules per unit cell of dimensions $a = 8.9044(9)$ Å, $b = 17.623(1)$ Å, $c = 10.175(1)$ Å, $\beta = 113.97(1)°$. The X-ray intensity data were collected on a Nonius CAD4 diffractometer, Sect. 5.6ff, using Cu $K\alpha$ radiation.

A total of 3193 reflections was collected, to a $\theta_{\max}$ of 74.22°, of which 2972 are independent with an R_{int}, Sect. 10.4.7, of 0.0175. Using the coumarin ring system as the search model provides a fragment consisting of 11 out of 21 non-hydrogen atoms. The fractional scattering power (p^2), defined in the previous section, for this model is very high (50.4%). The coordinates generated by Chem-X for the first search model are shown in Table 8.14, which is a complete listing of the PATSEE input data, whereas Table 8.15 summarizes the search results.

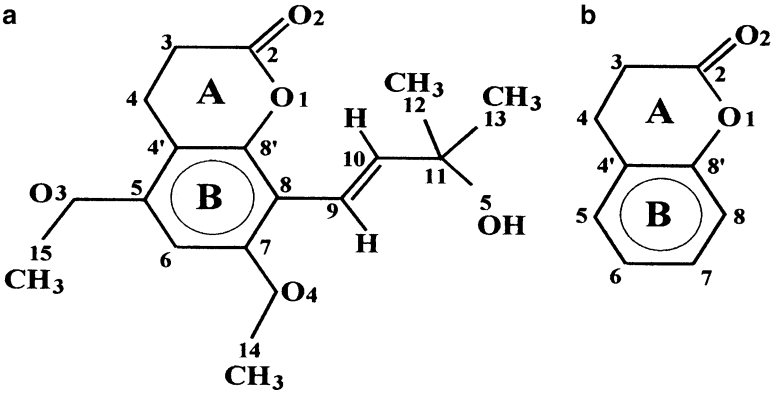


Fig. 8.16 Chemical formulae: (**a**) the Coumarin derivative; (**b**) Coumarin: rings A and B form a planar group for use in Patterson Search (Diagrams produced by Chemwindow (Softshell International Limited))

Table 8.14 Input data for coumarin model 1 PATSEE search

TITLE	Coumarin derivative Chem-X model for PATSEE
CELL	1.5418 8.9044 17.6236 10.1757 90.0 113.97 90.0
ZERR	4 0.0009 0.001 0.001 0 0.01 0
LATT	1
SYMM	*X*, 0.5 − *Y*, 0.5 + *Z*
SFAC	C H O
UNIT	64 72 20
ROTS	10000 20
TRAN	
FRAG	1 100.00 100.00 100.00 90.0 90.0 90.0
O1	3 −0.01015 0.00495 0.000
C2	1 −0.02219 −0.00188 0.000
O2	3 −0.03271 0.00429 0.000
C3	1 −0.02228 −0.01572 0.000
C4	1 −0.01034 −0.02273 0.000
C4′	1 0.00169 −0.01589 0.000
C5	1 0.01363 −0.02289 0.000
C6	1 0.02566 −0.01605 0.000
C7	1 0.02576 −0.00221 0.000
C8	1 0.01382 0.00479 0.000
C8′	1 0.00179 −0.00205 0.000

Notes: The above data have the following interpretation: CELL: wavelength and target unit-cell parameters; ZERR: number of molecules in target cell and errors in experimental unit-cell parameters a, b, c, α, β, γ; LATT 1: primitive unit cell and centrosymmetric space group; SYMM: other space-group symmetry operations; SFAC: atom types in molecule; UNIT: number of each atom type in unit cell; ROTS: 10000 20 (use 10,000 random test orientations and retain the best 20); TRAN: initiate translation search using nt random positions calculated by the program (the user also has the option of supplying a value for this number here); FRAG: 1 = fragment 1 (a second fragment can be supplied in addition as in the example below). The next six values are the unit-cell parameters for the fragment. This is not a real crystal cell as this fragment was model built in Chem-X. The final 11 lines of data here represent the model atom names, types, and coordinates in the FRAG unit cell

Table 8.15 Summary of results for coumarin model 1

		Rotation stage			Translation stage
Solution	RFOM	α (°)	β (°)	γ (°)	CFOM
1	0.514	5.103	5.844	2.295	0.559
2	0.514	5.109	5.887	0.527	0.678
3	0.514	1.261	3.580	0.844	0.540
4	0.514	1.317	2.667	0.593	0.665
5	0.514	3.524	4.462	1.948	0.542
6	0.514	3.185	1.821	0.953	0.854
7	0.514	1.325	3.594	2.521	0.698
8	**0.514**	**6.163**	**5.078**	**0.884**	**1.053**←
9	0.514	5.153	5.841	0.462	0.909
10	0.514	1.832	0.475	0.891	0.881
11	0.514	2.868	4.993	1.146	0.554
12	0.514	5.136	0.437	2.633	0.716
13	0.491	1.939	5.704	2.240	0.470
14	0.491	1.862	5.706	2.253	0.594
15	0.490	1.315	2.543	2.291	0.547
16	0.364	5.305	0.977	1.172	0.103
17	0.356	0.438	1.273	2.666	0.409
18	0.329	4.078	4.008	1.198	0.094
19	0.299	5.091	0.103	2.478	0.089
20	0.298	1.162	2.826	2.258	0.132

Discussion of Results and Expansion and Refinement of the Model

From the extract from the PATSEE output, Table 8.15, we can see:

1. The top 15 values of RFOM, after the rotation stage, are quite similar
2. The values of CFOM, after the translation stage, are more widely spread and only solution 8 has a value greater than 1.0
3. Solution 9 has CFOM = 0.909 and somewhat similar rotation solution angles

Using the coordinates of the coumarin fragment corresponding to solution 8, all of the atoms in the structure were located and refined in two iterations of Fourier-least-squares in the program WinGX. Figure 8.17 shows how the structure developed at these three stages. At stage (a), Fig. 8.17a, the *R* factor was very high, as expected, with a value of 57.3%, reducing to 42.3% after the stage of Fig. 8.17b, and dramatically to 15.5% represented by Fig. 8.17c, which included all non-hydrogen atoms. Experience tells us that a structure which refines with isotropic temperature factors to this sort of *R*-value is probably correct. Further refinement of the model led to the published structure for which $R = 3.9\%$, Fig. 8.17d. It has been shown that the use of model 1 with PATSEE has led to a successful determination of the structure with the RT solution corresponding to the highest CFOM value. The strategy of carrying 20 rotation stage solutions through to the 00translation stage was necessary in view of the lack of discrimination in the RFOM values.

Search Model 2 from Chem-X

In order to probe further the working and effectiveness of the PATSEE program, the above calculations were repeated with a smaller search model. A smaller, six-atom fragment, model comprising ring B, Fig. 8.16b, was generated again using Chem-X. This is essentially a benzene ring with six equivalent bonds and angles. The fractional scattering power p^2 is now much lower,

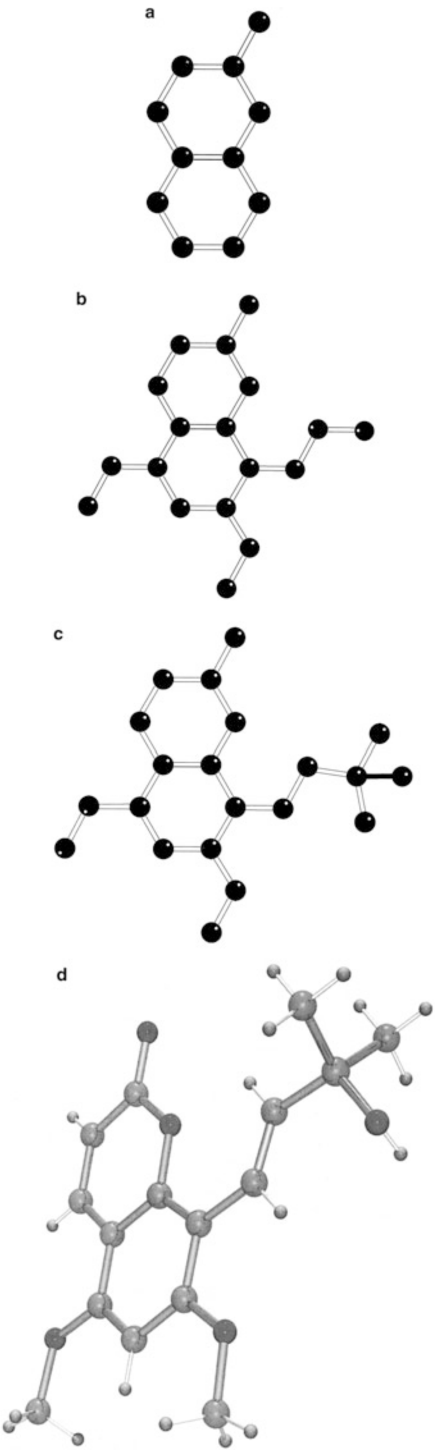


Fig. 8.17 Stages in determination of the structure of the coumarin derivative. (**a**) The search model. (**b**) The partial structure developed after one cycle of refinement. (**c**) After two cycles of refinement. (**d**) After further refinement (diagrams by POV-Ray™ VERSION 3.1 (http://www.povray.org), as implemented in WinGX and generated by Ortep-3 for Windows)

Table 8.16 Summary of RT results for coumarin model 2

		Rotation stage			Translation stage
Solution	RFOM	α (°)	β (°)	γ (°)	CFOM
1	0.604	0.515	1.328	2.756	0.960
2	0.604	5.015	5.907	2.162	0.661
3	0.604	5.088	5.846	0.637	0.881
4	0.604	3.746	1.826	1.492	0.437
5	0.604	5.088	5.284	2.035	0.264
6	0.604	1.973	5.816	0.615	0.983
7	0.604	5.956	4.351	2.546	1.335
8	0.604	1.278	2.682	2.257	1.234
9	0.604	1.895	5.873	0.623	1.071
10	0.604	0.353	1.714	1.263	0.783
11	0.604	3.158	4.603	2.316	0.787
12	0.604	3.353	4.417	2.025	0.879
13	0.604	2.724	1.756	0.429	0.830
14	0.604	4.548	3.606	0.925	0.715
15	0.604	1.468	3.602	0.781	0.691
16	**0.604**	**6.289**	**4.886**	**0.936**	**1.478** ←
17	0.604	2.741	4.932	1.130	0.697
18	0.604	4.301	3.585	2.717	1.055
19	0.604	1.843	5.736	2.330	1.174
20	0.604	1.217	3.681	2.402	0.650

24.1% instead of 50.4% for model 1, and is therefore expected to be less effective in the Patterson search. In other words even if PATSEE can locate the model correctly in the unit cell, this may not be the top solution and may be more difficult to expand to the full molecule. The program was again instructed to retain the top 20 rotation solutions, which were then passed on to the translation procedure, with the results as summarized in Table 8.16.

Expansion and Refinement of the Model

The coordinates for the six atoms corresponding to the best RT solution (solution 16) were input to the WinGX package and refined with SHELX-97 for two least-squares cycles. Inspection of a subsequent electron density map, using the program SXGRAPH in WinGX, allowed all of the missing structure to be built in and refined with two iterations of Fourier and isotropic least squares to an *R*-factor of 15.5%. The above result is quite pleasing, and may be somewhat surprising in view of the relatively few atoms used in the search.

Search Model 3 from Chem-X

The search model was further modified and reduced by removing atoms C4′ and C8′, Fig. 8.16b, and the PATSEE procedure was repeated. A correctly rotated and translated search model was found in the PATSEE RT listing. It was expanded and refined using WinGX to the same isotropic *R*-factor as in the previous cases, but with a little more difficulty. The reason this model did not produce the top RT solution is undoubtedly due to its size. Compared to models 1 and 2, the fractional scattering power for model 3 is only 16.1%. It is therefore a tribute to the method that the correct crystal structure could be generated with PATSEE without too much difficulty; remember though, that it is often much easier to solve a problem when the answer is known.

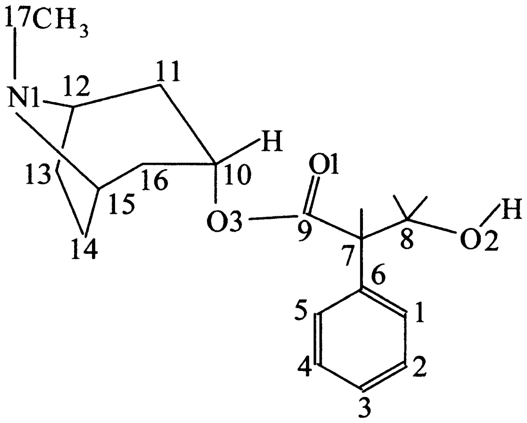


Fig. 8.18 Chemical formula of atropine (diagram produced by Chemwindow (Softshell International Limited))

The above experiments with a second and third model emphasize the power of the method and it should be remembered that some structures do still fail to be determined by direct methods. We know that a 21-atom structure can be solved with a 4-atom search model. Similar models may be useful for solving small polypeptide structures, for example, using the well-defined planar peptide group in the initial search. A dipeptide search model with 1° of rotational freedom may enable a tetrapeptide structure to be solved. This procedure will become clearer after studying the next section.

Structure of Atropine: *α-[Hydroxymethyl]Benzeneacetic Acid 8-Methyl-8-Azabicyclo[3.2.1]oct-3-yl Ester*

The chemical formula of the atropine molecule is shown in Fig. 8.18. Until recently [35] there were no reports of the structure of this classic molecule in the literature. Atropine is a competitive antagonist at central and peripheral synapses and has been used, somewhat unadvisedly, as a beautifying agent; hence its alternative, better known name *Belladonna*.

Atropine, $C_{17}H_{23}NO_3$, $M_r = 289.4$, is α-[hydroxymethyl]benzeneacetic acid 8-methyl-8-azabicyclo [3.2.1]oct-3-yl ester, and is known also as tropine tropate.

Practical Details

Unit-cell determination and refinement, data collection, and data reduction were carried out on a Nonius Kappa CCD diffractometer [36] using Mo $K\alpha$ radiation at −173 °C (liquid nitrogen temperature) with the aid of an Oxford Cryostreams cooler. At this temperature, the crystal diffracted strongly over the 90 min period of data collection. Data frames were processed using the Nonius software. The material crystallizes in the orthorhombic system, with the unusual space group, *Fdd*2, Fig. 8.19, with 16 molecules per unit cell of dimensions $a = 24.291(5)$ Å, $b = 39.538(8)$ Å, $c = 6.472(1)$ Å. A total of 2701 independent reflections were collected, to $\theta_{\max}$ 25.02°.

Patterson Search Models

The molecule contains a benzene ring, as in the previous example, which is an obvious fragment to use for Patterson Search. This can be easily extended by adding atom C7 to form a slightly larger fragment comprising 7 of the 21 non-hydrogen atoms, with a fractional scattering power $p^2 = 0.30$. However with this model, trials with PATSEE failed to identify an RT solution that could be expanded and refined into the complete molecule.

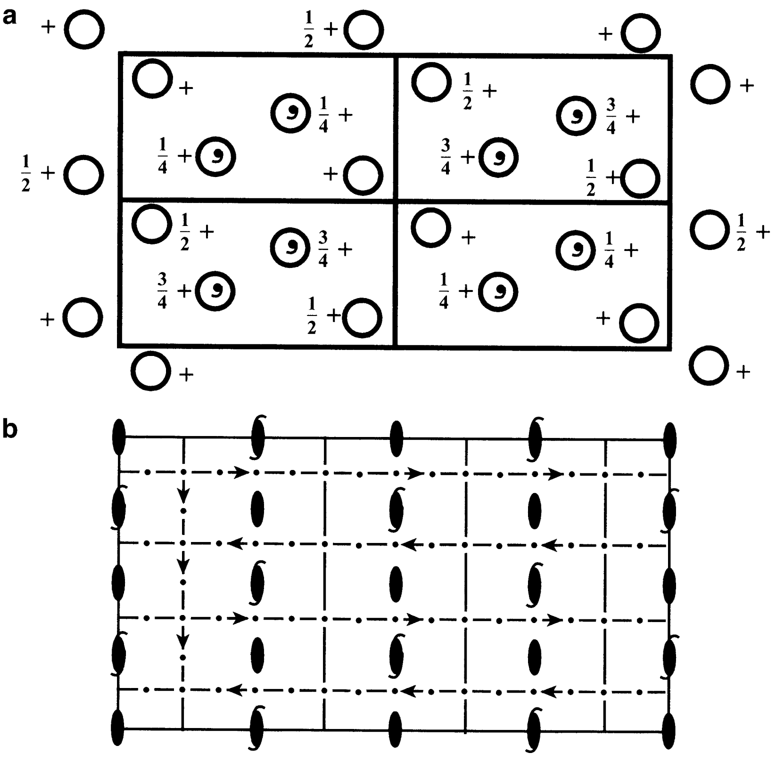


Fig. 8.19 Space group diagrams for *Fdd*2. (**a**) General equivalent positions. (**b**) Symmetry elements

It can be demonstrated, by applying the usual Fourier-least-squares procedure to the *known* atom coordinates for this fragment, that phasing on the 7 atoms of this model leads to a complete and refinable structure. We conclude therefore that if the correct solution is in fact produced by PATSEE for this small fragment, it does not appear as one of the best solutions, as judged by the figures of merit.

In order to overcome this problem and further test the features of PATSEE, a model was built with Chem-X, which included fragment 1 (atoms 1–6) and five other atoms, C7, C9, O1, O3, and C8, as fragment 2, Fig. 8.20. Fragment 2 is itself structurally rigid, but for the combined search model of fragments 1 and 2 there is *rotational freedom* about the bond C6–C7. The PATSEE program allows the linkage torsion angle, Sect. 8.5.2, between two such fragments to be systematically varied during Patterson Search to produce a series of test models. Each new set of coordinates produced by a change of this torsion angle is then treated as an independent search model. The computer time required to complete the rotation search is thus multiplied by the number of individual models explored. The best models are again passed to the translation stage.

Patterson Search for Atropine

The input data are shown in Table 8.17. The reader should identify the differences between the data in this table and those in Table 8.14 used for coumarin. Obvious changes arise from the differences in unit-cell dimensions and space group and molecular formula. The reader should study Fig. 8.19 so as to identify the entries that appear in Table 8.17; LATT −4 is the code for a non-centrosymmetric space group with an *F* unit cell. The rotation (ROTS) and translation (TRAN) instructions have been retained from the coumarin example without change; see Table 8.14 and footnote, especially for information on ROTS and TRAN. The next major change occurs in the FRAG listing of atom positions: the instruction TWIS 0 2 360 causes the program to vary the

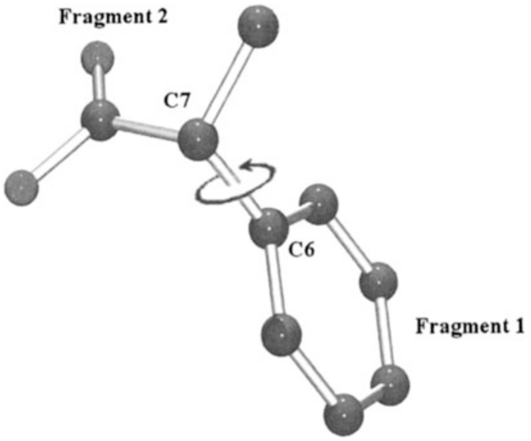


Fig. 8.20 Search model for atropine; torsional flexibility between the two rigid groups occurs about the C6–C7 bond indicated (diagram produced by POV-Ray™ VERSION 3.1 [85], as implemented in WinORTEP [86])

Table 8.17 Input data for PATSEE search for atropine fragment

TITL	Atropine PATSEE using dual model
CELL	0.71073 24.2913 39.5380 6.4727 90.000 90.000 90.000
ZERR	16.00 0.0049 0.0079 0.0013 0.000 0.000 0.000
LATT	−4
SYMM	$-X, -Y, Z$
SYMM	$0.25 + X, 0.25 - Y, 0.25 + Z$
SYMM	$0.25 - X, 0.25 + Y, 0.25 + Z$
SFAC	C H N O
UNIT	272 368 16 48
ROTS	10000 20
TRAN	
FRAG	1 100.00 100.00 100.00 90.0 90.0 90.0 (*Fragment via Chem-X*)
C1 1	0.12068 0.04724 0.06165
C2 1	0.12991 0.05601 0.06751
C3 1	0.13407 0.05393 0.08046
C4 1	0.12927 0.04327 0.08761
C5 1	0.12003 0.03461 0.08187
C6 1	0.11568 0.03661 0.06881
TWIS	0 2 360
C7 1	0.10546 0.02722 0.06245
C8 1	0.07132 0.04096 0.06734
C9 1	0.09194 0.02896 0.06908
O3 3	0.08480 0.03817 0.06248
O1 3	0.08835 0.02332 0.07902

Table 8.18 Atropine fragment rotation search

Solution	RFOM	α (°)	β (°)	γ (°)	Torsion angle (°)
1	1.742	2.942	2.912	0.174	182
2	1.732	5.892	3.527	1.736	2
3	1.701	6.036	3.359	2.980	180
4	1.676	6.052	3.352	2.984	182
5	1.563	2.932	2.916	0.176	184
6	1.544	5.827	3.474	1.766	358
7	1.535	5.933	3.540	1.737	360
8	1.535	5.933	3.540	1.737	0
9	1.522	6.101	3.295	2.999	184
10	1.438	5.938	3.567	1.705	4
11	1.403	2.883	2.927	0.158	176
12	1.353	2.690	5.981	1.726	35
13	1.348	2.912	2.876	0.179	180
14	1.334	6.105	3.408	2.969	186
15	1.332	2.702	5.949	1.783	354
16	1.312	5.862	3.549	1.745	6
17	1.285	2.946	2.883	0.176	178
18	1.272	6.114	3.360	3.008	188
19	1.246	2.594	6.027	1.757	352
20	1.222	2.588	5.984	1.748	356

Note: Each entry in this table is the result of 10,000 test rotations

torsion angle systematically about C6–C7 from 0 to 360° in steps of 2°, thus involving 181 different search models in the rotation procedure. The final application of the torsional change at 360° gives the same result as for the first at 0 and acts merely as a check.

Rotation Stage

The results of the rotation analysis are summarized in Table 8.18. In this table and in Table 8.19, the torsion angle refers to the relative orientation, about the C6–C7 bond, of the two rigid fragments of the search model (Fig. 8.20).

Translation Stage

The possible rotation function solutions are then each transferred to the translation algorithm in turn. The best 20 solutions, Table 8.19, were subjected to the translation algorithms as before. Values of CFOM after translation and optimization range from 0.929 to 2.640, as shown in the table.

In this analysis the order of precedence has changed when using CFOM as the final discriminator. As would be expected the best solutions (arrowed) all have similar rotational and torsional parameters. Although not shown here, the corresponding translation components derived in the second stage of PATSEE are also quite similar, as would be expected.

Expansion and Refinement of the Model

The overall best solution was taken to be number 2 in Table 8.19, and the model coordinates corresponding to this solution were subjected to Fourier-least-squares expansion and refinement as before. After two iterations of Fourier-least squares all non-carbon atoms were located and isotropically refined to an R factor of 0.105. In the published structure *anisotropic* refinement led to final R factor of 0.0453. The completed structure of atropine is shown in Fig. 8.21.

Table 8.19 Atropine fragment translation search results

Solution	RFOM	Torsion angle (°)	CFOM
1	1.742	182	1.408
2	**1.732**	**2**	**2.640** ←
3	1.701	180	1.373
4	1.676	182	1.250
5	1.563	184	1.424
6	1.544	358	2.284 ←
7	1.535	360	1.232
8	1.535	0	1.232
9	1.522	184	1.029
10	1.438	4	2.342 ←
11	1.403	176	1.081
12	1.353	356	1.342
13	1.348	180	1.039
14	1.334	186	1.114
15	1.332	354	1.307
16	1.312	6	1.793
17	1.285	178	1.077
18	1.272	188	0.929
19	1.246	352	1.763
20	1.222	356	1.694

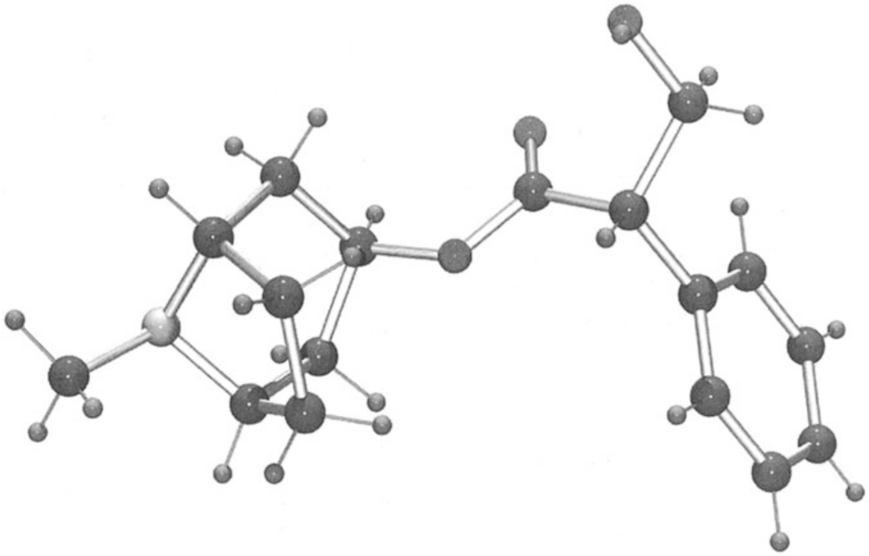


Fig. 8.21 The completed structure of atropine (diagram produced by POV-Ray™ VERSION 3.1 [85], as implemented in WinORTEP [86])

Conclusions

These examples illustrate many of the features of the powerful Patterson search technique. However, for small-molecule analysis, because direct methods are so much more easy to apply, if not to understand, Patterson search will remain as a reserve technique, albeit a very useful and powerful one, in the crystallographer's armory.

We have discussed two small-molecule analyses where Patterson search with PATSEE has led successfully to refineable models of the crystal structure. Provided that a suitable search model is available there is no reason to doubt that most small-molecule structures could be solved in a similar way. The most useful discriminator provided by PATSEE is the combined figure of merit (CFOM). For all but the smallest search models employed in the examples discussed, the RT solution with the largest value of CFOM has proved to be correct.

A search model with a fractional scattering power as low as 16.1% and including only 6 out of a total of 21 non-hydrogen atoms can provide the correct answer, but it requires more sifting of coordinate sets, using Fourier and least squares, to be selected. We shall see in Sect. 10.6.4 that the related method of Molecular Replacement (MR) is routinely used in macromolecular structure analysis, and for this reason the examples discussed in the present chapter will be invaluable for a sound understanding of the principles of the method when we come to discuss the large molecule studies.

8.3.8 Shake and Bake

We mention here, very briefly, another procedure that is a further stage in solving really complicated structures. Like PATSEE the program is readily available [37] and is quite fun to use, as the name *Shake and Bake* suggests.

In outline, Shake and Bake is a direct structure solving procedure that carries out phase refinement in reciprocal space and electron density space alternately, in order to achieve a true minimum of $R(\Phi)$, a quantity known as the minimal function [38].

$$\begin{aligned} R(\Phi) = \Bigg\{ & \sum_{\mathbf{H,K}} A_{\mathbf{HK}} \left[\cos T_{\mathbf{HK}} - \frac{J_1(A_{\mathbf{HK}})}{J_0(A_{\mathbf{HK}})}\right]^2 + \sum_{\mathbf{L,M,N}} |B_{\mathbf{LMN}}| \left[\cos Q_{\mathbf{LMN}} - \frac{J_1(B_{\mathbf{LMN}})}{J_0(B_{\mathbf{LMN}})}\right]^2 \\ & + \sum_{\mathbf{L,M,N}} |B_{\mathbf{LMN}}| \left[\cos Q_{\mathbf{LMN}} - \frac{J_1(B_{\mathbf{LMN}})}{J_0(B_{\mathbf{LMN}})}\right]^2 \left[\sum_{\mathbf{H,K}} A_{\mathbf{HK}} + \sum_{\mathbf{L,M,N}} |B_{\mathbf{LMN}}|\right]^{-1} \Bigg\} \end{aligned} \tag{8.68}$$

where $A_{\mathbf{HK}}$ and $B_{\mathbf{LMN}}$ are given by (8.69).

$$\begin{aligned} A_{\mathbf{HK}} &= (2/N^{1/2})|\mathrm{E}_{\mathbf{H}}\mathrm{E}_{\mathbf{K}}\mathrm{E}_{\mathbf{H+K}}|B_{\mathbf{LMN}} \\ &= (2/N)|\mathrm{E}_{\mathbf{L}}\mathrm{E}_{\mathbf{M}}\mathrm{E}_{\mathbf{N}}\mathrm{E}_{\mathbf{L+M+N}}|(|\mathrm{E}_{\mathbf{L+M}}|^2 + |\mathrm{E}_{\mathbf{M+N}}|^2 + |\mathrm{E}_{\mathbf{N+L}}|^2 - 2) \end{aligned} \tag{8.69}$$

The structure triple $T_{\mathbf{HK}}$ is defined by

$$T_{\mathbf{HK}} = \phi_{\mathbf{H}} + \phi_{\mathbf{K}} + \phi_{\mathbf{-H-K}} \tag{8.70}$$

and an estimate of cos $T_{\mathbf{HK}}$ follows the equation

$$\cos\, T_{\mathbf{HK}} = J_1(A_{\mathbf{HK}})/J_0(A_{\mathbf{HK}}) \tag{8.71}$$

where I_0 and I_1 are the modified Bessel functions.

The structure quartet $Q_{\mathbf{LMN}}$ is given by

$$Q_{\mathbf{LMN}} = \phi_{\mathbf{L}} + \phi_{\mathbf{M}} + \phi_{\mathbf{N}} + \phi_{\mathbf{-L-M-N}} \tag{8.72}$$

and the phases themselves are functions of the atomic positions:

$$E(\mathbf{h}) = |E(\mathbf{h})| \exp \mathrm{i}2\pi\phi(\mathbf{h}) = \frac{1}{\sqrt{N}} \sum_j \exp \mathrm{i}2\pi\mathbf{h} \cdot \mathbf{r}_j \tag{8.73}$$

The Shake and Bake procedure is initiated by generating triplet and negative quartet structure invariants that are based on a trial structure or partial structure model comprised of N atoms in the unit cell, chosen such that no two atoms are closer than ca. 1.2 Å and no atom lies within bonding distance of four other atoms. Normalized structure factors $\mathrm{E}_{\mathbf{h}}$ are determined and with the phases a first electron density map calculated. The phases are then subjected to a cyclical shake-and-bake phasing procedure consisting of phase refinement, to minimize $R(\varphi)$, and constrained electron density maps to obtain atom positions. The time per cycle depends on the number of atoms and is about 10 s for a 30-atom structure and about 90 s for a structure with 400 atoms in the unit cell. The technique has solved a number of structures, including one with over 600 atoms in the asymmetric unit. For further details the reader is referred to the Appendix D and the literature [39–41]. The procedure has also been used successfully on the 1001 non-hydrogen atom structure of lysozyme [38] and on a range of other protein structures [42].

8.4 Least-Squares Refinement

If we have two pairs of values of X and Y for measurements which are related by the equation

$$Y = mX + b \tag{8.74}$$

we can obtain a unique answer for the constants m and b. Sometimes, as in the Wilson plot, we have several pairs of values, which contain random errors, and we need to obtain those values of m and b that best fit the complete set of observations. In practical problems, we have often a situation in which the errors in the X values are negligible compared with those in Y.

Let the best estimates of m and b under these conditions be m_0 and b_0. Then, the error of fit in the ith observation is

$$e_i = m_0 X_i + b_0 - Y_i \tag{8.75}$$

The principle of least squares states that the best-fit parameters are those that minimize the sum of the squares of the errors. Thus,

$$\sum_i e_i^2 = \sum_i (m_0 X_i + b_0 - Y_i)^2 \quad (i = 2, \ldots, N) \tag{8.76}$$

has to be minimized over the number N observations. This condition corresponds to differentiating partially with respect to m_0 and b_0, in turn, and equating the derivatives to zero. Hence,

$$m_0 \sum_i X_i^2 + b_0 \sum_i X_i = \sum_i X_i Y_i \tag{8.77}$$

$$m_0 \sum_i X_i + b_0 N = \sum_i Y_i \tag{8.78}$$

which constitute a pair of simultaneous equations (*normal equations*) easily solved for m_0 and b_0.

In a crystal structure analysis, we are always manipulating more observations than there are unknown quantities; the system is said to be overdetermined. We shall consider some crystallographic applications of the method of least squares.

8.4.1 Unit-Cell Dimensions

In Chap. 5, we considered methods for obtaining unit-cell dimensions with moderate accuracy from photographic and diffractometer measurements. Generally, we need to enhance the precision of these measurements, which may be achieved by a least-squares analysis. Consider, for example, a monoclinic crystal for which the θ *values* of a number of reflections of known indices, preferably high-order, have been measured to the nearest 0.01°. In the monoclinic system, sin θ is given, Table 2.4 with $k = \lambda$, and (3.43), by

$$4\sin^2\theta = h^2a^{*2} + k^2b^{*2} + l^2c^{*2} + 2hla^*c^* \cos\beta^* \tag{8.79}$$

In order to obtain the best values of a^*, b^*, c^*, and cos β^*, we write, following (8.76),

$$\sum_i \left(h_i^2a^{*2} + k_i^2b^{*2} + l_i^2c^{*2} + 2h_il_ia^*c^* \cos\beta^* - 4\sin^2\theta_i\right)^2 \tag{8.80}$$

and then minimize this expression, with respect to a^*, b^*, c^*, and cos β^*, over the number of observations i. The procedure is a little more involved numerically; we obtain four simultaneous equations to be solved for the four variables, but the principles are the same as those involved with the straight line.

8.4.2 Least-Squares Parameters

Correct trial structures are refined by the least-squares method. In essence, this process involves adjusting a scale factor and the positional and temperature parameters of the atoms in the unit cell so as to obtain the best agreement between the experimental F_o values and the $|F_c|$ quantities derived from the structure model. In its most usual application, the technique minimizes the function.

$$R' = \sum_{\mathbf{h}} w(F_o - G(|F_c|)^2 \tag{8.81}$$

where the sum is taken over the set of crystallographically independent terms **h**, w is a weight for each term, and G is the reciprocal of the scale factor K for F_o. Let p_j ($j = 1, 2,\ldots, n$) be the variables in $|F_c|$ whose values are to be refined. Then

$$\frac{\partial R'}{\partial p_j} = 0 \tag{8.82}$$

or

$$\sum_{\mathbf{h}} w\Delta \frac{\partial|\mathrm{F_c}|}{\partial p_j} = 0 \tag{8.83}$$

where Δ is $F_\mathrm{o} - |\mathrm{F_c}|$. For a trial set of parameters not too different from the correct values, Δ is expanded as a Taylor series to the first order:

$$\Delta(\mathbf{p}, \boldsymbol{\xi}) = \Delta(\mathbf{p}) - \sum_{i=1}^{n} \xi_i \frac{\partial|\mathrm{F_c}|}{\partial p_j} \tag{8.84}$$

where the shift ξ_i is the correction to be applied to parameter p_i; $\mathbf{p}$ and $\boldsymbol{\xi}$ represent the complete sets of variables and corrections. Substituting (8.84) in (8.83) leads to the normal equations

$$\sum_{i=1}^{n} \left[\sum_{\mathbf{h}} w \frac{\partial|\mathrm{F_c}|}{\partial p_i} \frac{\partial|\mathrm{F_c}|}{\partial p_j} \right] \xi_i = \sum_{\mathbf{h}} w\Delta \frac{\partial|\mathrm{F_c}|}{\partial p_j} \tag{8.85}$$

The n normal equations may be expressed neatly in matrix form:

$$\mathbf{A}\,\boldsymbol{\xi} = \mathbf{b} \quad \text{or} \quad \sum_{i} a_{ij}\xi_i = b_j \tag{8.86}$$

where

$$a_{ij} = \sum_{\mathbf{h}} w \frac{\partial|\mathrm{F_c}|}{\partial p_i} \frac{\partial|\mathrm{F_c}|}{\partial p_j} \tag{8.87}$$

and

$$b_j = \sum_{\mathbf{h}} w\Delta \frac{\partial|\mathrm{F_c}|}{\partial p_j} \tag{8.88}$$

The solution of the normal equations is a well- documented mathematical procedure that we shall not dwell upon. Instead, we draw attention to certain features of least-squares refinement. It is important to remember that least squares provides the best fit for the parameters that have been put into the model. Hence, it is essential to examine a final difference-Fourier map at the completion of a least-squares refinement, after several cycles of calculations have led to negligible differences ξ_i. The techniques of least squares have been reported fully at Crystallographic Computing Conferences (see Bibliography).

Temperature Factors

The Wilson technique, Sect. 4.2.1, leads to an overall isotropic temperature factor for the structure, and we have discussed the theory of one-dimensional isotropic thermal vibration in Sect. 4.1.8. As indicated there, a better procedure allots a B parameter to each atom and refines the values as least-squares parameters. We write the temperature correction factor T_i as

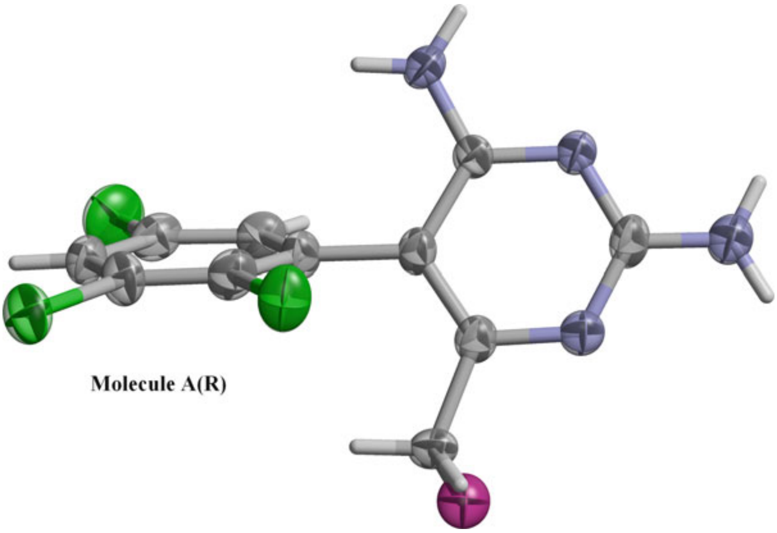


Fig. 8.22 Anisotropic thermal ellipsoids for molecule A of BW202W92(R) [26]. The ellipsoids are plotted at the 50% probability level, which means that at the maximum radius drawn the exponential expression of (8.91) is 0.5. In terms of an isotropic thermal surface, the radius would be almost 0.6 Å corresponding to $\sin^2\theta/\lambda^2$, or $(\frac{1}{2}d)^2$, of 0.694 Å^{-2} (drawn in WinGX with ORTEP)

$$T_i = \exp[-(B_i\lambda^{-2}\sin^2\theta)] \tag{8.89}$$

and the equation

$$B_i = 8\pi^2\overline{U_i^2} \tag{8.90}$$

relates the isotropic temperature factor B_i to the mean square amplitude $\overline{U_i^2}$ of the ith atom; the three-dimensional surface of isotropic vibration is a sphere. A more sophisticated treatment, as we indicated already in Sect. 4.1.8, describes the vibrations of each atom by a symmetrical tensor **U** having six independent components in the general case:

$$\begin{aligned}T_i = \exp[-2\pi^2(&U_{11}h^2a^{*2} + U_{22}k^2b^{*2} + U_{33}l^2c^{*2} + 2U_{23}klb^*c^* + 2U_{31}lhc^*a^*\\&+ 2U_{12}hka^*b^*)]\end{aligned} \tag{8.91}$$

and the (anisotropic) U_{ij} parameters are refined as part of the model. The surface of vibration is now a biaxial (thermal) ellipsoid, and the mean-square amplitude of vibration in the direction of a unit vector $\mathbf{L}(L_1, L_2, L_3)$ is given by

$$\overline{U_i^2} = \sum_{i=1}^{3}\sum_{j=1}^{3} U_{ij}L_iL_j \tag{8.92}$$

Biaxial ellipsoids for molecule A of the R-enantiomer BW202W92(R) are illustrated in Fig. 8.22; see also Sects. 8.2.12 and 8.2.16. Since **L** is defined with respect to the reciprocal lattice, the component of **U** with $\mathbf{L}(1,0,0)$, parallel to a^*, is

$$\overline{U_i^2} = U_{11} \tag{8.93}$$

In an orthorhombic crystal, for example, a direction 30° from a^* in the a^*b^* plane has $\mathbf{L} = (\frac{\sqrt{3}}{2}, \frac{1}{2}, 0)$, and the component of $\mathbf{U}$ in that direction is

$$\overline{U_i^2} = U_{11}(\tfrac{\sqrt{3}}{2})^2 + U_{22}(\tfrac{1}{2})^2 + 2U_{12}(\tfrac{\sqrt{3}}{2})(\tfrac{1}{2}) \tag{8.94}$$

The following relationships among the values of B, $\overline{U_i^2}$, and the root mean square (rms) amplitude are often useful:

B/Å^2	U^2/Å^2	rms Amplitude (Å)
0.10	0.0013	0.036
0.50	0.0063	0.080
1.0	0.013	0.11
5.0	0.063	0.25
10	0.13	0.36

The smallest rms amplitudes encountered are ca. 0.05 Å. Values of B between 3 and 10 Å^2 are found in organic structures at ambient temperatures. The larger rms amplitudes require caution in interpreting them in terms of bond lengths and their precision. For example, static or dynamic disorder, Sect. 8.9, not included in the model, may be manifested as abnormally large temperature factors.

Scale Factor

In least-squares refinement the F_o data must not be adjusted, and so the parameter G in (8.81) is introduced. The inverse of the refined value of G may be applied to F_o at the end of a refinement cycle. Several cycles of refinement may be needed before the parameters reach a sensibly constant value. Generally full-matrix least-squares refinement is to be preferred. However, where the number of parameters is very large or where computer availability is limited, an approximation may be used. One such method is the block-diagonal refinement, in which certain off-diagonal a_{ij} terms in (8.86) are neglected. Generally, more cycles are necessary in this procedure.

Weights

In the initial stages of refinement, weights may be set at unity or chosen so as to accelerate the process, such as down-weighting reflections of small F_o or of high order, or both. In the final stages, weights should be related to the precision of F_o, which can be achieved in two ways:

1.

$$w(hkl) = 1/[\sigma^2(F_o(hkl)] \tag{8.95}$$

where the estimated standard deviation, $\sigma[F_o(hkl)]$, is obtained from counting statistics in diffractometer data by the relationship

$$\sigma = \sqrt{N} \tag{8.96}$$

N being related to the total counts, peak, and background, for the given reflection. Sometimes a quantity $pF_o{}^2$ is added to the right-hand side of (8.95), where p is adjusted so that $w\Delta^2$ is constant over ranges of F_o, where Δ is $F_o - |\mathrm{F_c}|$.

2.
$$w(hkl) = \Phi(F_o) \tag{8.97}$$

where the function Φ is chosen so that $w\Delta^2$ is again sensibly constant over ranges of F_o. Another weighting scheme is given by

3.
$$w = (A + F_o + BF_o{}^2 + CF_o{}^3)^{-1} \tag{8.98}$$

where the constants A, B, and C may be obtained by a least-squares fit of mean values of Δ, in ranges of F_o, to the inverse of the right-hand side of (8.98).

Precision

The choice of absolute weights (8.95) should yield parameters of lowest variances:

$$\sigma^2(p_j) = (a_{jj})^{-1} \tag{8.99}$$

where $(a_{jj})^{-1}$ is an element of the matrix inverse to that of the a_{ij} elements (8.87). With weights related to F_o,

$$\sigma^2(p_j) = (a_{jj})^{-1} \frac{\sum_{\mathbf{h}} w\Delta^2}{m - n} \tag{8.100}$$

where m is the number of reflections and n the number of parameters to be refined. Generally, $m/n \geq ca$ 10 leads to a high quality analysis. In a block-diagonal approximation, the standard deviations are usually underestimated by 15–20%.

Atoms in Special Positions

If any symmetry operation of the space group of a structure leaves an atom invariant, the atom is on a special position. Consider one molecule of formula AB_2 in the unit cell of space group $P2$. The A atoms occupy, say, the special positions 0, y, 0, and the B atoms occupy the general positions x, y, z and $\bar{x}, y, \bar{z}$. Several important points arise:

1. The x and z coordinates of atoms A remain invariant at zero during refinement.
2. In this space group, the origin is along the twofold axis y and must be specified by fixing the y coordinate of an atom; the heavier the atom the better. It could be $y_B = 0$, in which case this parameter also remains invariant.
3. With respect to the symmetry operations of the space group, atoms A must be given an atom multiplicity factor of $\frac{1}{2}$ so that a total of one A atom per unit cell obtains.
4. In an isotropic refinement, one of the three twofold axes of the biaxial thermal ellipsoid must lie along the twofold axis y of the space group. In this case, the U_{12} and U_{23} elements of $\mathbf{U}$ remain invariant at zero value.

8.4.3 Theory of Least-Squares Refinement and Strategies to Use

The calculations of least squares, although lengthy, have been implemented in now well-tested program systems, such as WinGX and SHELX, that are available to workers in crystal structure determination. In a number of instances, the application of least-squares calculations is straightforward, and results of sufficient precision are obtained. In other cases, and perhaps more generally, a consideration of the strategy to be employed, within the constraints of the program system, is necessary and desirable. In the structure determination process, a model is proposed on the basis of the diffraction data, and then subjected to adjustments (refinement) of the parameters of that model so as to reach a solution with a defined precision.

Model

For a least-squares refinement to be successful, the model must be sufficiently close to the truth and contain all the components of the structure. The multidimensional surface that corresponds to an n-dimensional refinement is complex and contains many false minima, into any of which an insufficiently correct model may converge. Hence, the correctness of a structure should be based on several criteria, Sect. 8.7.

Data Errors

Experimental data are subject to systematic and random errors. In X-ray diffraction, a systematic error may be, for example, the lack of an absorption correction: the values of F_o will often be less than those of $|\mathrm{F_c}|$ for the partially refined structure, noticeably for the low-order reflections of high intensity. The minimization of the difference between F_o and $|\mathrm{F_c}|$ will lead, in this case, to unwarranted discrepancies in the temperature factors of the atoms and the scale factor of the data.

Increasing the temperature factors and decreasing G, the scale factor applied inversely to $|\mathrm{F_c}|$ during refinement, will tend to decrease the $|\mathrm{F_c}|$ values so as to compensate for the effect of absorption. Random errors are unavoidable in any physical experiment, but they can be minimized by careful attention to the experimental procedure. For example, reflections of low intensity and consequent high probable error can be improved in reliability by increasing the measurement time. This procedure may be costly in time, but is very straightforward with a diffractometer.

The data set must be sufficient in size for the work in hand. The success of least-squares refinement is partially dependent on the fact that the calculations are appreciably overdetermined. This excess of data over variables is needed in order to average out discrepancies in individual measurements. As an approximate guide, the ratio of data to variables should be at least 8. The ratio can be improved by decreasing the number of variables. There are instances, such as with phenyl rings, where refinement of the parameters of the hydrogen atoms might be regarded as exaggerated. The hydrogen atoms can, and should, be allowed to contribute to the calculated structure factors, but unless there is reason to suppose otherwise they would be expected to display sp^2 geometry, with C–H $\approx$ 1.00 Å. Their isotropic temperature factors may be refined, or assumed to be, say, 1.3 times that of the carbon atoms to which they are attached. In a structure like euphenyl iodoacetate $\mathrm{C_{32}H_{53}O_2I}$, Sect. 1.1, the number of variables may be very significantly reduced by this type of approach. The strategy should be determined by the nature of the problem in hand, and computer programs that handle least-squares refinement have been designed to accommodate a range of constraints.

If the data-to-variables ratio is too low, the final structure may have an inherent lack of resolution that may not be immediately apparent from the numerical least-squares results, although all the expected atomic positions will be represented, provided that the model itself was complete. Least squares, unlike the Fourier process, cannot find anything that is not present in the model. It will obtain a best fit, under the given strategy, for the model supplied. A lack of resolution may arise because of an unsatisfactory termination of the data set. If the cut-off criterion for acceptable data leads to too few reflections, those that have been excised could be subjected to re-measurement for longer times so that they can be accepted under the same criterion. If this is not done, then imperfections will exist in the refinement, and may be manifested in large estimated standard deviations for some parameters or in unsatisfactory temperature factors. Incorrect values for B factors may lead to highly improbable root mean square atomic displacements or to U_{ij} values that are non-positive definite, that is, they do not define an ellipsoid.

There is some concern with published papers on structure determination that report 50% or more of the unique experimental F_o data rejected by some criterion, such as $I \leq 3\sigma(I)$, apparently to the satisfaction of both the worker and the journal editor. The cosmetic effect of this practice serves merely to reduce the R factor rather than to improve the estimated standard deviations of the structural parameters. Weak intensity data do, in fact, contain structural information. Protein crystallographers, who of necessity work with poorly diffracting crystal specimens, are not in a position to discard data in this fashion.

Least-Squares Refinement Procedure

The least-squares refinement of a model leads to those parameters that minimize $(F_o - |\mathrm{F_c}|)^2$, as in (8.76), over the whole data set. The first application of the calculations will generally not lead to the best-adjusted parameters, and it will be necessary to cycle through the calculations until the results converge; that is, until the calculated shifts are less than the estimated standard deviations of the corresponding parameters. The refinement is continued until the shifts in the parameters are some small fraction (0.1 is often quoted) of the estimated standard deviations.

A good procedure will begin by ensuring the maximum number of good observations, and by minimizing the number of variables consistent with the requirements of the problem. The starting model must be sufficiently good so that the minimization does not fall into a local, false position. This is unlikely if Fourier synthesis or difference-Fourier synthesis has been used in establishing the model and approximate scale and temperature factors have been obtained by statistical methods. Although Fourier methods of refinement are less convenient than those of least squares, they do have the power of revealing necessary information that may not be contained in an initial model. From the above, it should be clear that such information is vital to a good least-squares refinement.

Another constraint that may be considered, as well as the fixed C–H geometry, is the rigid-body specification of a group of atoms. For example, unless there is any reason to suppose otherwise, a phenyl ring can be considered to obey the geometry of Tables 8.21 and 8.22, or other equivalent compilation, and thus be refined as a single, rigid entity.

The least-squares equations outlined in Sect. 8.4.2 include a weight for each observation. They may be unity or calculated from (8.97) for the early stages of refinement. Subsequently, weighting schemes based on F_o or $\sin\theta$ or on both of these parameters may be used. The validity of any weighting scheme should be checked for the given problem, as already indicated. Attention to detail in least-squares refinement is generally rewarding, and it is worth remembering that it consumes about three-quarters of the total calculation time of a structure analysis.

8.4.4 Least-Squares Refinement Against F_o^2

Some crystallographic program packages, notably SHELX-97, employ full-matrix least-squares refinement routines based on $F_o{}^2$ and $|\mathrm{F_c}|^2$, as opposed to F_o and $|\mathrm{F_c}|$ as shown in (8.81). Apart from this major difference which will influence the form of the normal equations, the method of solution to produce parameter shifts and estimated standard deviations, σ, will be entirely parallel to the procedure described above for refinement against F_o and $|\mathrm{F_c}|$. Studies have shown that refinement against $|\mathrm{F}|^2$ allows more weighted experimental data to be incorporated into the analysis which enables it to proceed more smoothly. A further rationale in favor of this choice is that it is $F_o{}^2$ and $\sigma(F_o{}^2)$ that are, after corrections, obtained from the diffraction experiments so why refine against F_o and $|\mathrm{F_c}|$? Refinement against F_o also involves mathematical problems with very weak reflections or reflections with apparent negative measured intensity values. There are also difficulties in converting and estimating $\sigma(F)$ from the measured $\sigma(F^2)$ values for weak or zero measured intensities. Refinement against $|F|^2$ avoids these difficulties, and also reduces the probability of the refinement iterations settling into a local minimum. It also simplifies the treatment of twinned crystals and the determination of the absolute configuration in the case of non-centrosymmetric structures, Sect. 7.5.1. For these reasons, it is probably currently the most frequently used technique, although it does rely heavily on the assignment of reasonable weights to individual reflections. Further details are to be found in the SHELX references [21, 23].

8.4.5 Constraints and Restraints

The program SHELX-97 allows the user to apply a wide range of controls on the structural parameters being refined in any particular case. It is necessary in this respect to distinguish between a constraint and a restraint. A *constraint* is an exact mathematical condition that enables one or more least-squares variables to be expressed exactly in terms of other variables or constants, and as such is eliminated from the actual refinement process while still contributing as part of the structure to $|F_c|$. An example is the fixing of an atom position exactly on a center of symmetry. A *restraint* involves additional information that is not exact but is subject to a probability distribution; for example chemically but not crystallographically equivalent bonds could be restrained to be approximately equal within a specified tolerance. A restraint is treated as an extra experimental observation, with an appropriate esd that determines its weight relative to the X-ray data.

Constraints are flagged in SHELX-97 using appropriate instructions prefixed by AFIX. A wide variety of options is available as described in the manual. An HFIX instruction generates AFIX instructions and dummy hydrogen atoms bonded to specified atoms. All types of hydrogen atom groupings are available. For example, AFIX 6 fits a regular hexagon to six specified atoms, of default bond length 1.39 Å for aromatic C–C bonds, and AFIX 13 specifies an ideal $–CH_3$ group with tetrahedral angles. A DFIX instruction allows the distance between two named atoms to be restrained to a target value d within a specified standard deviation.

These facilities have a wide variety of uses. For example, to economize on parameters if the data/parameter ratio is low, or to enable hydrogen atoms to be included in the analysis when it has not been possible to locate them from an appropriate electron density map.

8.5 Molecular Geometry

When the structure analysis is complete, the results must be expressed in terms of molecular geometry and crystal packing. Thus, we need to compute bond lengths, bond angles, intermolecular contact distances, torsion angles, where necessary, all with measures of their precision.

8.5.1 Bond Lengths and Angles

Consider three atoms with fractional coordinates x_1, y_1, z_1; x_2, y_2, z_2; and x_3, y_3, z_3 in a unit cell of sides a, b, and c, Fig. 8.23. The vector $\mathbf{r}_j$ from the origin O to any atom j is given by

$$\mathbf{r}_j = x_j\mathbf{a} + y_j\mathbf{b} + z_j\mathbf{c} \tag{8.101}$$

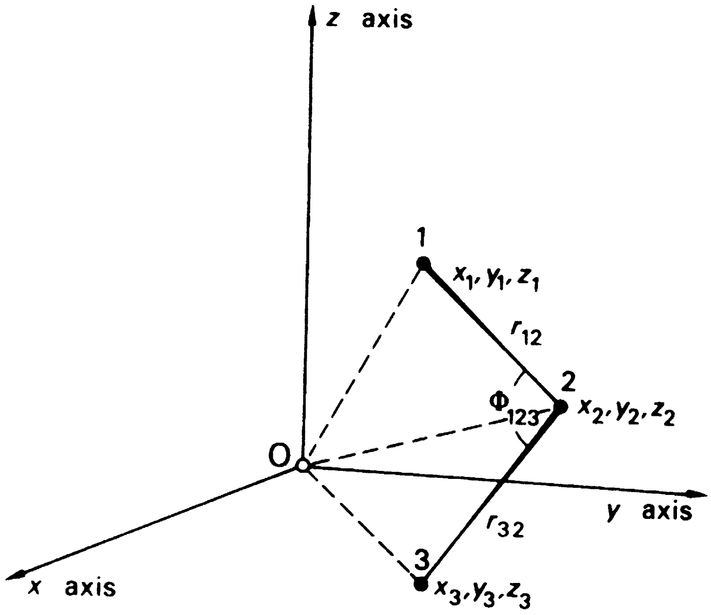


Fig. 8.23 Geometry of the calculation of interatomic distances and angles; points 1, 2, and 3 represent atomic positions

The vector $\mathbf{r}_{12}$ between atoms 1 and 2 is given by

$$\mathbf{r}_{12} = \mathbf{r}_2 - \mathbf{r}_1 \tag{8.102}$$

or, using (8.101),

$$\mathbf{r}_{12} = (x_2 - x_1)\,\mathbf{a} + (y_2 - y_1)\,\mathbf{b} + (z_2 - z_1)\,\mathbf{c} \tag{8.103}$$

Forming the dot product of each side with itself, remembering that

$$p \cdot q = pq \cos \widehat{\mathbf{pq}} \tag{8.104}$$

we obtain

$$r_{12}^2 = (x_2 - x_1)^2 a^2 + (y_2 - y_1)^2 b^2 + (z_2 - z_1)^2 c^2 + 2(y_2 - y_1)(z_2 - z_1)\,bc\cos\alpha + 2(z_2 - z_1)(x_2 - x_1)\,ca\cos\beta + 2(x_2 - x_1)(y_2 - y_1)\,ab\cos\gamma \tag{8.105}$$

This equation may be simplified for crystal systems other than triclinic. Thus, if the atoms exist in a tetragonal unit cell, for example,

$$r_{12}^2 = [(x_2 - x_1)^2 + (y_2 - y_1)^2]a^2 + (z_2 - z_1)^2 c^2 \tag{8.106}$$

In a similar manner, we can evaluate r_{32}, Fig. 8.23.

In the case of a bond angle Φ_{123} formed from atoms 1, 2, and 3, and using (8.105) for the tetragonal system,

$$\cos\Phi_{123} = \frac{[(x_2 - x_1)(x_2 - x_3) + (y_2 - y_1)(y_2 - y_3)]\,a^2 + (z_2 - z_1)(z_2 - z_3)\,c^2}{r_{12}r_{32}} \tag{8.107}$$

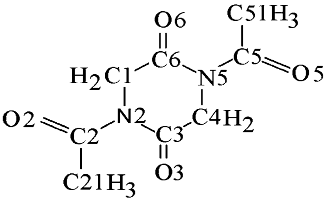
where r_{12} and r_{32} are evaluated following (8.106). Similar equations enable any distance or angle to be calculated, in any crystal system, in terms of the atomic coordinates and unit-cell dimensions.

When the asymmetric unit of a crystal contains more than one copy of a given molecule, or when similar molecules occur in different crystals, the question arises as to whether or not the several sets of molecular dimensions are significantly different. The statistical test applicable in this situation is the χ^2 test. It involves calculation of $\sum_{i=1}^{n}[\Delta_i/\sigma(\Delta_i)]^2$, which is distributed as χ^2 with n degrees of freedom; Δ_i is the difference between one measured property, such as a bond length, in a pair of molecules, and $\sigma(\Delta_i)$is the standard deviation in Δ_i, estimated as $[\sigma^2(d_{i1}) + \sigma^2(d_{i2})]^{1/2}$, assuming no correlation between d_{i1} and d_{i2}, and n is the number of pairs of measurements.

Significance Tests and χ^2 Distributions

The significance of a result can be tested by making the null hypothesis that all of the differences can be accounted for by random errors in the experimental procedures, and then obtaining from statistical tables the significance level of the test, that is, the probability P of incorrectly rejecting a good hypothesis. Normally, the test is not regarded as significant unless $P \leq 0.05$. Bond lengths and angles and their estimated standard deviations are calculated and listed by SHELX-97 through the BOND instruction in the .ins file.

As an example we will consider two aspects of the bond lengths in the X-ray structure [43] of the cyclic diamino acid peptide *N*,*N*′-diacetyl-cyclo(Gly-Gly), shown in the diagram below; the eight ring bonds, including the C=O bonds and the six side-chain bonds, will be considered in the table below. The crystal structure has two independent copies of the molecule labeled A and B.



N,*N*′-Diacetyl-Cyclo(Gly-Gly) showing the atom numbering scheme (Gly = glycine)

Using Table (b) below, we can show that:

Table a Bond lengths for *N*,*N*′-diacetyl-cyclo(Gly-Gly), molecules A and B

	Bond lengths (Å)		χ^2 Components				
Molecule	A	B	Δ_i	$\sigma(\Delta_i)$	$[\Delta_i/\sigma(\Delta_i)]$	$[\Delta_i/\sigma(\Delta_i)]^2$	
C(1)–N(2)	1.474(2)	1.472(2)	0.002	0.0028	0.714	0.51	Cα–N[a]
N(2)–C(3)	1.381(2)	1.383(2)	0.002	0.0028	0.714	0.51	N–C(O)
C(3)–C(4)	1.512(3)	1.508(3)	0.004	0.0042	0.953	0.91	C(O)–Cα
C(4)–N(5)	1.469(2)	1.474(2)	0.005	0.0028	1.786	3.19	Cα–N
N(5)–C(6)	1.386(2)	1.389(2)	0.003	0.0028	1.071	1.15	N–C(O)
C(1)–C(6)	1.507(3)	1.505(3)	0.002	0.0042	0.476	0.23	Cα–C(O)
C(6)–O(6)	1.217(2)	1.215(2)	0.002	0.0028	0.714	0.51	C(O)=O
C(3)–O(3)	1.214(2)	1.209(2)	0.005	0.0028	1.786	3.19	C(O)=O
$\sum_{i=1}^{n}[\Delta_i/\sigma(\Delta_i)]^2 = 10.20$ for $n = 8$							
C(2)–N(2)	1.416(2)	1.413(2)	0.003	0.0028	1.071	1.15	Side-chain bonds
C(2)–O(2)	1.217(2)	1.218(2)	0.001	0.0028	0.357	0.13	
C(2)–C(21)	1.492(3)	1.493(3)	0.001	0.0042	0.238	0.06	
C(5)–N(5)	1.415(2)	1.410(2)	0.005	0.0028	1.786	3.19	
C(5)–O(5)	1.213(2)	1.212(2)	0.001	0.0028	0.357	0.13	
C(5)–C(51)	1.486(3)	1.474(3)	0.012	0.0042	2.857	8.16	
$\sum_{i=1}^{n}[\Delta_i/\sigma(\Delta_i)]^2 = 12.82$ for $n = 6$							

[a] Standard peptide bond notation is used here (see Chap. 10)

Table b Chi-square distribution [45] χ^2 vs. Probability, P. This table gives a number of P values matching to χ^2 for the first 10° of freedom, df. A P-value of 0.05 or less is usually regarded as statistically significant, that is, the observed deviation from the null hypothesis is significant.

Degrees of freedom (df)	Probability (P)										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

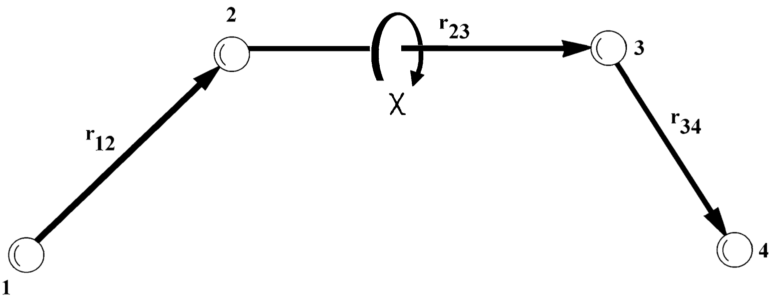


Fig. 8.24 The torsion angle $\chi(1, 2, 3, 4)$ has a positive sign for a rotation of 1, 2, about 2, 3 as shown, that is, it is positive if, looking along $\mathbf{r}_{23}$, a clockwise rotation is required to bring atom 1 into atom 4

1. For the eight ring bond lengths the probability P_{ring}, the difference between molecules A and B, is between 0.2 and 0.3, that is greater than 0.05 and is not significant
2. For the six side-chain bond lengths the probability P_{chain}, the differences between molecules A and B, lies between 0.01 and 0.05, less than 0.05, so that they are significant. Exact values of P are [44]: $P_{\text{ring}} = 0.251$ and $P_{\text{chain}} = 0.0463$, thus confirming the above results

It is not too difficult to see that the main difference in side-chain bond lengths between molecules A and B occurs for C(5)–C(51); in practice it may be necessary to seek further reasons for this type of result.

8.5.2 Torsion Angles

Torsion angles are useful conformational parameters with which to compare different, related molecules or, indeed, different conformations of one and the same molecule. In a freely rotating moiety, a torsion angle may be a function of the environment of the molecule. Consider an arrangement of four atoms 1, 2, 3, 4, Fig. 8.24. The torsion angle $\chi(1, 2, 3, 4)$ is defined by the angle between the planes 1, 2, 3 and 2, 3, 4, and lies in the range $-180° < \chi \leq 180°$; the sign is an important property of the parameter.

In the planar, eclipsed conformation shown in Fig. 8.24, χ is zero. The torsion angle is the amount of rotation of 1, 2 about 2, 3 and, looking along the direction 2 → 3; a positive value of χ corresponds to the clockwise rotation that brings atom 1 into atom 4. Let

$$\mathbf{p}_1 = \mathbf{r}_{23} \times -\mathbf{r}_{12} \tag{8.108}$$

and

$$\mathbf{p}_2 = \mathbf{r}_{23} \times \mathbf{r}_{34} \tag{8.109}$$

Then

$$\chi(1,2,3,4) = \cos^{-1}\left(\frac{\mathbf{p}_1 \cdot \mathbf{p}_2}{p_1 p_2}\right) \tag{8.110}$$

If the torsion angle is calculated by the expression atan2 $\{|\mathbf{r}_{23}|\mathbf{r}_{12}.(\mathbf{r}_{23}\times\mathbf{r}_{34}), (\mathbf{r}_{12}\times\mathbf{r}_{23}).(\mathbf{r}_{23}\times\mathbf{r}_{34})\}$ the angle is obtained with the correct sign. [atan2 is a Fortran function for $\tan^{-1}$, with two arguments]. Torsion angle calculation is provided by crystallographic software, for example, the program MOLGOM in the program suite, Sect. 13.6.5, or by using the CONF instruction in SHELX-97. WinGX has a facility for producing standard tables for publication including molecular geometry: the “Publish” button, Fig. 8.10a, invokes this facility.

8.5.3 Conformational Analysis

Confusion has arisen in the literature over the use of torsion angles in conformational analysis. It is often convenient to quote values of torsion angles as lying within certain ranges. For example, $\chi \approx 0°$ may be called *cis*, $\chi = 180°$ is *trans*, and $\chi \approx \pm 60°$ is ±*gauche*. However, because of changing conventions, it is best to quote the actual value of χ, and to state how it is defined, Fig. 8.24. This procedure will minimize ambiguities in the future.

Ring Conformations

Two types of symmetry (or pseudo-symmetry) must be considered in order to define ring conformations [46], namely mirror planes perpendicular to the dominant ring plane and twofold axes lying in the ring plane. If there is an odd number (usually 5 or 7) of atoms in the ring, all symmetry elements pass through one of the ring atoms and bisect the opposite bond, Fig. 8.25. In rings containing an even number of atoms (usually 6), symmetry elements may pass through two ring atoms located directly across the ring, or else bisect two opposite ring bonds.

Ten symmetry elements are possible in five-membered rings. The *planar* five-membered ring possesses all ten, five mirror planes and five twofold axes. The ideal *envelope* conformation has only a single *m* plane, and it passes through the out-of-plane atom. The ideal *half-chair* has one twofold axis bisecting the bond between the two out-of-plane atoms. Six-membered rings possess 12 locations for symmetry elements. In determining the ring conformation, we can ignore the two-, three-, and sixfold collinear rotation axes perpendicular to the ring plane. Figure 8.26 illustrates the symmetry elements that define the ideal forms of commonly observed conformations. The planar ring, such as in benzene, has one *m* plane and one twofold axis at each of six locations ($\frac{6}{m}mm$). The *chair* form of cyclohexane has three *m* planes and three twofold axes ($\bar{3}m$). The *boat* and *twist-boat* have point group symmetry *mm*2 and 222, respectively, while the *sofa* has symmetry *m* and the *half-chair* symmetry 2.

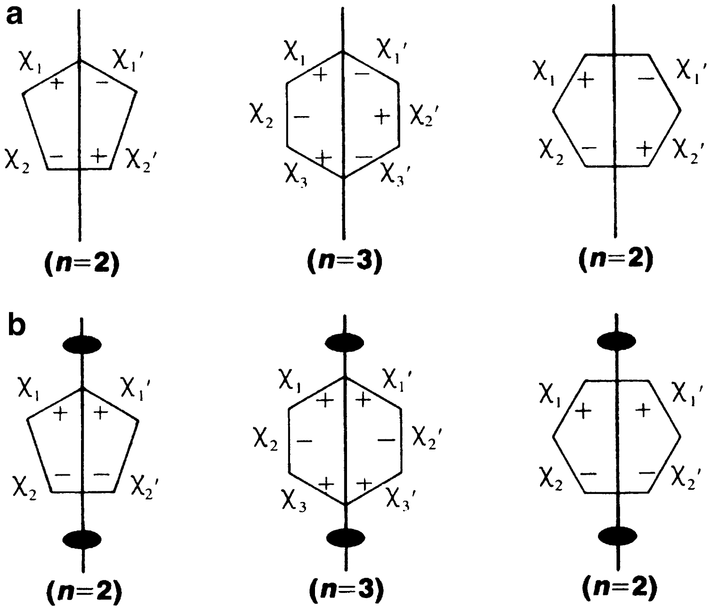


Fig. 8.25 Conformations in five- and six-membered rings. (**a**) Torsion angles related by mirror planes (—) have opposite signs. (**b**) Torsion angles related by twofold rotation axes have the same sign

Asymmetry Parameters

Once the atom coordinates are available, torsion angles may be calculated. Because of errors in the data and for stereochemical reasons, a particular cyclic structure will often depart from its ideal symmetry. The degree of this departure, its asymmetry, may be calculated in terms of asymmetry parameters. For this purpose, related or nearly related torsion angles are compared in a way that will result in a value of zero for a parameter if the corresponding symmetry is realized in the molecule. Mirror-related torsion angles have equal magnitude but opposite sign, and such torsion angles are compared by addition. The torsion angles related by twofold symmetry are equal in both magnitude and sign, and are compared by subtraction. The rms value of each discrepancy yields a measure of the deviation from ideal symmetry at the location in question. We calculate

$$\Delta C_s = \left(\frac{\sum_{i=1}^{n} (\chi_i + \chi_i')^2}{n} \right)^{1/2} \tag{8.111}$$

in respect of m symmetry, and

$$\Delta C_2 = \left(\frac{\sum_{i=1}^{n} (\chi_i - \chi_i')^2}{n} \right)^{1/2} \tag{8.112}$$

in respect of twofold symmetry; n is the number of individual comparisons, and χ_i and χ_i' are the related torsion angles in question.

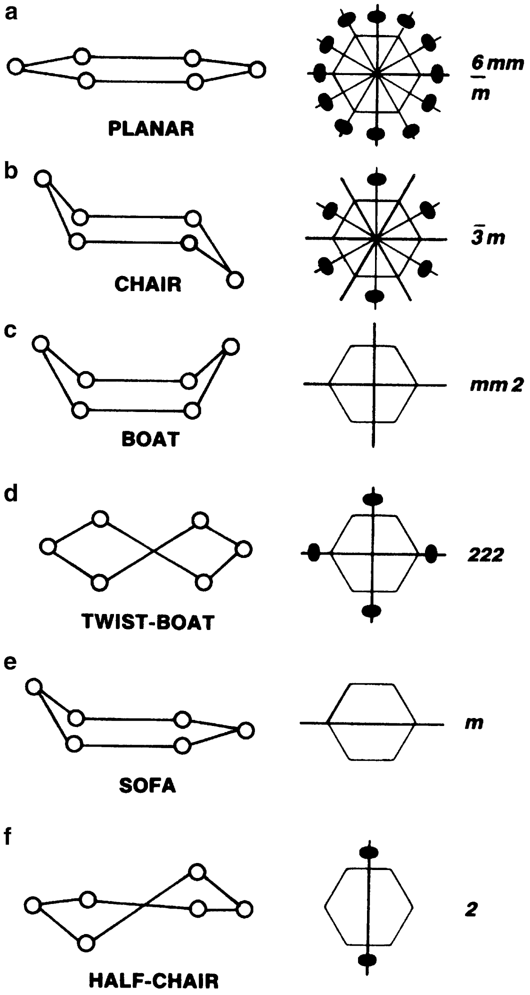


Fig. 8.26 Commonly observed conformations of six-membered rings. The mirror and twofold rotational symmetries are indicated on the *right*

8.5.4 Mean Planes

In discussing the geometry of a molecule, it may be desirable to test the planarity of a group of atoms. For a number n ($n > 3$) of atoms, the best plane may be obtained by the method of least squares. Let the plane be given by

$$PX + QY + RZ + S = 0 \tag{8.113}$$

The constants P, Q, R, and S are obtained through an extension of the least-squares procedure given in Sect. 8.4; it is desirable to work with Cartesian coordinates (see Appendix C).

8.6 Precision

Closely related to the calculations of bond lengths and angles is the expression of the precision of these quantities. The least-squares refinement procedure establishes values for an estimated standard deviation in each of the variables used in these calculations. Thus, a fractional coordinate of 0.3712 might have an esd of 0.0003, written as 0.3712(3).

We need to know further how errors are propagated in a quantity which is a function of several variables, each of which contains some uncertainty arising from random errors. The answer is provided by the statistical principle of superposition of errors. Let q be a function of several variables p_i ($i = 1, 2, 3, \ldots, N$), with known standard deviations $\sigma(p_i)$. Then the esd in q is given through

$$\sigma^2(q) = \sum_{i=1}^{N} \left[\frac{\partial q}{\partial p_i} \sigma(p_i) \right]^2 \tag{8.114}$$

A simple example may be given, through (8.103) and (8.114), for a bond between two atoms lying along the c edge of a tetragonal unit cell. Let c be 10.06(1) Å, z_1 be 0.3712(3), and z_2 be 0.5418(2). From (8.103),

$$r_{12} = (z_2 - z_1)c = (0.5418 - 0.3712)10.06 = 1.716\,\text{A} \tag{8.115}$$

and from (8.114),

$$\sigma^2(r_{12}) = (0.5418 - 0.3712)^2(0.01)^2 + (10.06)^2(0.0002)^2 + (10.06)^2(0.0003)^2 \tag{8.116}$$

Thus, $\sigma(r_{12})$ is 0.004 Å and we write $r_{12} = 1.716(4)$ Å. Similar calculations may be used for all distance and angle calculations in all crystal systems, but the general equations are quite involved numerically and best handled by computer methods.

The estimated standard deviation of a torsion angle can be calculated along the lines given in this section; the significances of differences between torsion angles may be as important as the differences themselves. Again, in a discussion of best-plane results it is essential to evaluate the perpendicular distances (deviations) of atoms from the plane, and their esds. Let the jth atom have the coordinates X_j, Y_j, Z_j. Then it is a simple exercise in coordinate geometry to show that the deviation Δ_j of this atom from the best plane (8.113) is given by

$$\Delta_j = (PX_j + QY_j + RZ_j + S)/K \tag{8.117}$$

where K is given by

$$K = (P^2 + Q^2 + R^2)^{1/2} \tag{8.118}$$

To obtain the esd in Δ_j the *esd*s in X_j, Y_j, and Z_j are first obtained, using (8.114) and results from Appendix C. Then

$$\sigma(\Delta_j) = \{[P\sigma(X_j)]^2 + [Q\sigma(Y_j)]^2 + [R\sigma(Z_j)]^2\}^{1/2}K \tag{8.119}$$

8.7 Correctness of a Structure Analysis

At this stage we may summarize four criteria of correctness of a good structure analysis. If we can satisfy these conditions in one and the same structure model, we shall have a high degree of confidence in the analysis.

1. There should be good agreement between F_o and $|F_c|$, expressed through the R factor. Ultimately, R depends upon the quality of the experimental data. At best, it will probably be about 1% greater than the average standard deviation in F_o. Assuming the desirable situation that two or more asymmetric units of data have been collected, a value R close to that for R_{int}, Sect. 10.4.7, for the data is acceptable.
2. The electron density map should show neither positive nor negative density regions that are unaccountable, other than Fourier series termination errors.
3. The difference-Fourier map should be relatively flat. This map eliminates series termination errors as they are present in both ρ_o and ρ_c. Random errors produce small fluctuations on a difference map, but they should be less than 2.5–3 times the standard deviation of the electron density $\sigma(\rho_o)$:

$$\sigma(\rho_o) = \frac{1}{V_c}\left[\sum_{hkl}(\Delta F)^2\right]^{1/2} \tag{8.120}$$

   where $\Delta F = F_o - |F_c|$ and the sum extends over all symmetry-independent reflections.
4. The molecular geometry should be chemically sensible, within the limits of current structural knowledge. Abnormal bond lengths and angles may be correct but they must be supported by strong evidence of their validity in order to gain acceptance. Normally a deviation of less than three times the corresponding standard deviation is not considered to be statistically significant. As a guide to the interpretation of acceptable stereochemistry, we include selections of ionic radii, bond lengths, and bond angles in Tables 8.20, 8.21, and 8.22, respectively; see also Table 13.4.

Table 8.20 Selected ionic radii (Å) referred to coordination number 6[a]

Ag^+	1.27	Hg^{2+}	1.02	Th^{4+}	0.94
Al^{3+}	0.54	K^+	1.44	Ti^{2+}	0.86
Ba^{2+}	1.49	La^{3+}	1.03	Ti^{4+}	0.61
Be^{2+}	0.48	Li^+	0.86	Tl^+	1.54
Ca^{2+}	1.18	Mg^{2+}	0.87	Tl^{3+}	1.03
Cd^{2+}	0.95	Mn^{2+}	0.83	Zn^{2+}	0.74
Ce^{3+}	1.01	Na^+	1.12	NH_4^+	1.66
Ce^{4+}	0.87	Ni^{2+}	0.69	H^-	1.39
Co^{2+}	0.75	Pb^{2+}	1.19	F^-	1.19
Co^{3+}	0.61	Pd^{2+}	0.86	Cl^-	1.70
Cr^{3+}	0.62	Pt^{2+}	0.80	Br^-	1.87
Cs^+	1.84	Pt^{4+}	0.63	I^-	2.20
Cu^+	0.77	Ra^{2+}	1.43	O^{2-}	1.40
Cu^{2+}	0.73	Rb^+	1.58	S^{2-}	1.70
Fe^{2+}	0.78	Sn^{2+}	0.93	Se^{2-}	1.81
Fe^{3+}	0.65	Sr^{2+}	1.32	Te^{2-}	1.97

[a] The changes in ionic radius from coordination number 6 to coordination numbers 8, 4, 3, and 2 are approximately +1.5 %, −1.5 %, −3.0 %, and −3.5 %, respectively

Table 8.21 Selected bond lengths (Å)[a]

Formal single bonds				Formal double bonds			
C4–H	1.09	C3–C2	1.45	C3–C3	1.34	C2–O1	1.16
C3–H	1.08	C3–N3	1.40	C3–C2	1.31	N3–O1	1.24
C2–H	1.06	C3–N2	1.40	C3–N2	1.32	N2–N2	1.25
N3–H	1.01	C3–O2	1.36	C3–O1	1.22	N2–O1	1.22
N2–H	0.99	C2–C2	1.38	C2–C2	1.28	O1–O1	1.21
O2–H	0.96	C2–N3	1.33	C2–N2	1.32		
				Formal triple bonds			
C4–C4	1.54	C2–N2	1.33	C2–C2	1.20	N1–N1	1.10
C4–C3	1.52	C2–O2	1.36	C2–N1	1.16		
C4–C2	1.46	N3–N3	1.45				
C4–N3	1.47	N3–N2	1.45				
				Aromatic bonds			
C4–N2	1.47	N3–O2	1.36	C2–C3	1.40	N2–N2	1.35
C4–O2	1.43	N2–N2	1.45	C2–N2	1.34		
C3–C3	1.46	N2–O2	1.41				

[a]The notation in the table indicates the connectivity of the atoms

Table 8.22 Selected bond angles

Atom	Geometry	Angle (°)
C4	Tetrahedral	109.47
C3	Planar	120
C2	Bent	109.47
C2	Linear	180
N4	Tetrahedral	109.47
N3	Pyramidal	109.47
N3	Planar	120
N2	Bent	109.47
N2	Linear	180
O3	Pyramidal	109.47
O2	Bent	109.47

8.7.1 Databases

Tables of standard (average) bond lengths and angles (both with esds) are useful aids to structure determination. In any research or advanced study, it is necessary to take cognizance of all work in the given field that has already been published. The number of crystal structures that has been solved and published is now vast, and data files have been constructed that can be interrogated by computer. The best known of these is the Cambridge Structural Database (CSD) [47]. It contains the results of both X-ray and neutron diffraction studies on organic and organometallic compounds. At the time of writing, over 544,000 crystal structures are filed in this database, summarized in Table 8.23, and the database is available in about 25 countries (see also Appendix D). The Cambridge Structural Database System (CSDS) is a single product that comprises the following components: *Cambridge Structural Database* (CSD); *CSDS Software*: *ConQuest, Mercury, VISTA,* and *PreQuest*. Knowledge bases derived from the CSD: *Mogul* and *IsoStar*. Life Sciences Products: *SuperStar*, *Hermes*, *GOLD*, *GoldMine*, *Relibase*+. Powder Diffraction Products: *DASH*.

One criticism of the CSD has been that it does not store the experimental F_o data on which each structure is based. The Protein Data Bank (PDB) does keep these data for each protein structure deposited. Not every structure in the above classes that has been ever published will be found in the CSD.

Table 8.23 Statistics of the Cambridge Structural Database (CSD), 2011

	Structures	%CSD
Total number of structures	596,810	100.0
Number of different compounds	544,565	–
Number of literature sources	1,429	–
Organic structures	254,475	42.6
Transition metal present	319,188	53.5
Li–Fr or Be–Ra present	30,134	5.0
Main group metal present	36,923	6.2
3D coordinates present	554,760	93.0
Error-free coordinates	545,085	98.3†
Neutron studies	1,534	0.3
Powder diffraction studies	2,354	0.4
Low/high temperature studies	250,328	41.9
Absolute configuration determined	11,111	1.9
Disorder present in structure	132,349	22.2
Polymorphic structures	18,386	3.1
R-factor $<$ 0.100	559,093	93.6
R-factor $<$ 0.075	506,465	84.9
R-factor $<$ 0.050	325,440	54.5
R-factor $<$ 0.030	65,897	11.0
No. of atoms with 3D coordinates	45,048,092	–

†Taken as a precentage of structures for which 3D coordinates are present in the CSD

Each entry has to pass a scrutiny that involves such checks as the consistency between the published coordinates and bond lengths. The information in the CSD falls into three categories, namely, *bibliographic*, *connective*, and *crystallographic*. The bibliographic file contains information such as the chemical name, the type of structure analysis, the chemical class, the molecular formula, and relevant literature for a given compound. The connective file contains chemical structural formulae encoded as atom and bond parameters, and the crystallographic file contains parameters relevant to the crystal structure data and its solving.

Retrieval from the database is flexible, and the software permits many different types of search, such as on chemical name, formula, or class, and compounds containing specific chemical fragments can be sought. The results of searches can be recorded by printing and plotting techniques. File formats which can be retrieved include .cif, .mol2, .pdb, SHELXL, and laztpulverix. The .cif file can be used, for example, to display the structure with MERCURY and is required to be deposited by authors to enable detailed checking of the structure determination and results.

Crystallographic programs such as SHELXL produce the .cif file at the completion of the analysis. The file can be checked for errors using programs in WinGX or other analysis systems or by using ENCIFER available through the CCDC [47]. Any alert produced by running the checking facilities should be dealt with by the authors prior to submission of their results for publication.

Similar databases have been organized: for proteins by the Resource for Studying Biological Molecules (RCSB) [48]; for inorganic structures at the Inorganic Crystal Structure Database (ICSD) [49], a database of inorganic and related structures produced cooperatively by FIZ Karlsruhe and the National Institute of Standards and Technology (NIST) [50] and searchable via two different web browser interfaces, CrystalWeb [51] and ICSD-WWW [49]; for metals, alloys, and intermetallic compounds: the CrystMet [52] database at Daresbury, UK which contains some 75,000 crystal structure data.

8.8 Limitations of X-Ray Structure Analysis

There are certain things that X-ray analysis cannot do well, and it is prudent to consider the more important of them.

Liquids and gases lack three-dimensional order, and cannot be used in diffraction experiments in the same way as are crystals. Certain information about the radial distribution of electron density can be obtained, but it lacks the distinctive detail of crystal analysis.

It is not easy to locate light atoms in the presence of heavy atoms. Difference-Fourier maps alleviate the situation to some extent, but the atomic positions are not necessarily precise. Least-squares refinement of light-atom parameters is not always successful, because the contributions to the structure factor from these atoms are relatively small.

Hydrogen atoms are particularly difficult to locate with precision because of their small scattering power and the fact that the center of the hydrogen atom does not, in general, coincide with the maximum of its electron density. Terminal hydrogen atoms have a more aspherical electron density distribution than do hydrogen-bonded hydrogen atoms, and their bond distances, from X-ray studies, often appear short when compared with spectroscopic or neutron diffraction values. For similar reasons, refinement of hydrogen-atom parameters in a structure analysis may be imprecise, and the standard deviations in their coordinate values may be as much as ten times greater than those for a carbon atom in the same structure. It is, nevertheless, very desirable to include hydrogen-atom positions in the final structure model. They lead to a best fit, and are useful when comparing the results of X-ray structure determination with those of other techniques, notably nuclear magnetic resonance.

In general, bond lengths determined by X-ray methods represent distances between the centers of gravity of the electron clouds, which may not be the same as the internuclear separations. Internuclear distances can be found from neutron diffraction data, because neutrons are scattered by the atomic nuclei. If, for a given crystal, the synthesized neutron scattering density is subtracted from that of the X-ray scattering density, a much truer picture of the electron density can be obtained. Neutron diffraction is discussed in detail in Chap. 11.

8.9 Disorder in Single Crystals

A typical small-molecule analysis may involve less than about 100 non-hydrogen atoms in the asymmetric unit. With Cu $K\alpha$ radiation and $\theta_{\max}$ tending to 70°, it would be expected to lead to the determination of bond lengths with esds of about 0.005 Å and of bond angles with esds of about 0.2°. Isotropic thermal parameters for non-hydrogen atoms usually range from 0.050 to 0.090 Å^2 and may have esds from 0.003 to 0.007 Å^2. However, it is sometimes found that the refined thermal parameters for certain atoms in a structure have atypical values. For example, U_{iso} may increase progressively and significantly toward the end of a chain-like moiety compared to the more rigid areas of the structure. The obvious and reasonable physical interpretation is simply that the atoms near the end of the chain experience greater thermal motion than do the atoms in the bulk of the molecule. For example, the hydroxyethyl side-chain atoms (excluding hydrogen) of an azasteroid derivative [53] have the following U_{iso} parameters:

>N	$–CH_3$	$–CH_2$	–OH
0.058	0.073	0.190	0.176

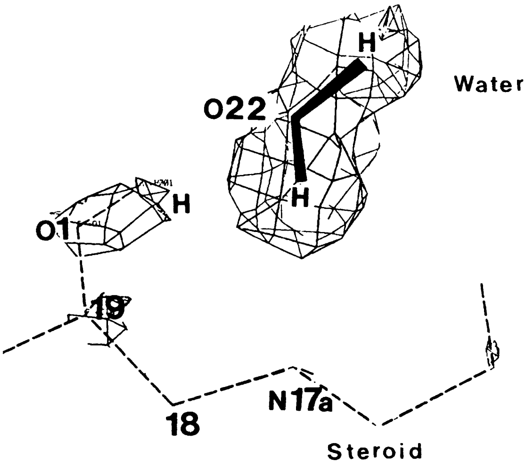


Fig. 8.27 Difference electron density map for azasteroid HS626 showing the hydrogen atom on O1 (hydroxyl) and the two water-hydrogen atoms, none of which was included in the structure factor calculation. The steroid molecule, part of which is shown by the dashed line, has been subtracted out in the difference synthesis (This figure and the next five are electron density and difference electron density maps photographed from the screen of an Evans and Sutherland Picture System 2 cathode ray tube display unit coupled to a computer that holds the electron density data. The interactive computer graphics system is programmed such that the user can simulate a three-dimensional effect by rotating the map about one or more of three mutually perpendicular axes. The contouring of the maps encloses the electron density in a cage of "chicken wire" hoops running in several directions. Unlike the sectional contour maps used elsewhere in this book, only one contour level is used, selected so as to optimize the desired features of electron density)

In the analysis of this azasteroid all atoms, including those in the side chain, were resolved and refined successfully by least squares.

Atoms in solvent of crystallization molecules may exhibit high thermal parameters, and for similar reasons. Exceptions occur from time to time, and in the above example a well-resolved solvent water-oxygen atom had a refined U_{iso} value of 0.088 Å^2 and was so well ordered that its hydrogen atoms were clearly located in a ΔF map, Fig. 8.27. In this particular case, the clarity of definition in the electron density is associated with the formation of two strong hydrogen bonds donated by each of the water-hydrogen atoms holding it firmly in position. However, in the structure of another steroid derivative [54], the carbon atoms of the side chains were so badly disordered that some atoms were not resolved in the difference electron density and appeared as diffuse patches, Figs. 8.28 and 8.29. Such disorder is probably of a statistical nature, with the atoms taking up slightly different positions from one unit cell to another. The effect can be compensated, albeit somewhat artificially, by the refinement of the isotropic thermal parameters assigned to the atoms concerned. In the example, the U_{iso} values are three to six times greater than those of the ordered atoms in the structure.

Disorder may also arise by groups of atoms either in free rotation in the solid state (dynamic disorder) or in more than one position of similar energy (static disorder). Methyl groups in large organic molecules often show this type of behavior. It may be possible to distinguish between dynamic and static disorder by a complete reexamination of the structure at a much reduced temperature.

Protein structures are of particular topical interest, and innovations in this field include the development of techniques for refining these large structures [55]. The molecules involved in protein analysis are very large, typically with more than 1000 non-hydrogen atoms in the asymmetric unit. Consequently, the crystals have large unit cells, and many possible X-ray reflections occur within a given θ-range compared to small-molecule crystals. It is customary to limit severely the maximum

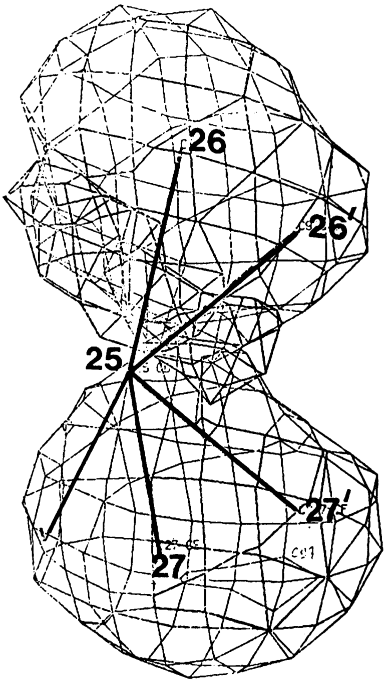
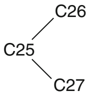


Fig. 8.28 Electron density at the end of the cholesteryl side chain of HS650 (molecule *B*). The density is smeared out, and at least two stereochemically sensible positions for the  fragment can be fitted to the density, as indicated

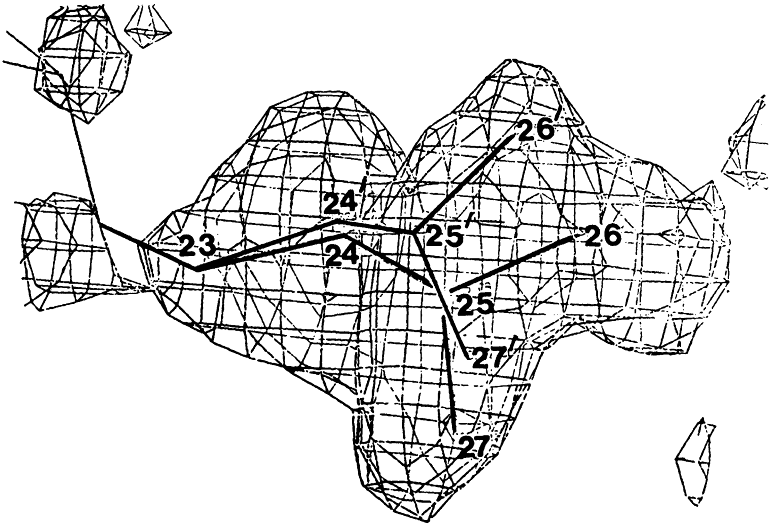


Fig. 8.29 As for Fig. 8.28, but with molecule *A* of HS650, in which the disorder in the side chain is more extensive and encompasses C23–C27. In this structure the side chains are loosely held, having little contact with neighboring molecules in the crystal

θ-value during the course of a protein structure analysis, depending on the particular stage reached. Corresponding to a given maximum θ-value there is a minimum d-value, $\lambda/2\sin\theta_{\max}$ and it is customary to speak of the $d_{\min}$ of an analysis as the nominal resolution. The protein analysis may proceed through stages of progressively higher resolution, for example, 6, 3, 2.5, 2, and 1.5 Å, the electron density image undergoing gradually improved mathematical focusing in the process.

In addition to the large quantity of data associated with protein analysis, there is a further technical problem, which limits the quality of most studies. During the process of crystallization from solution, solvent molecules, typically 40–60% by weight, are trapped in the structure: the protein molecules almost float in a solvated crystalline state [56], and consequently many regions of electron density in the protein structure may be subject to the type of disorder described above. Even in a good high-resolution analysis, protein data rarely extend beyond 1.5 Å, whence individual atoms in the protein molecule may not be revealed. A protein structure refinement involves the use of both least-squares analysis and geometrically constrained positioning of groups in order to produce a plausible model.

We conclude this section with an example of electron density determined in the high-resolution X-ray analysis of the enzyme ribonuclease [57], a small protein of molecular weight 13700. In the first example, a tyrosyl residue, Fig. 8.30, is seen at 0.85 Å extremely high resolution as a hollow ring of density and, although the individual atoms are not resolved, the shape of the density image is strikingly good. As would be expected, the sulfur atom of a methionine residue, Fig. 8.31, is quite outstanding, but it does not swamp the rest of this slender aliphatic side chain. At this resolution, the high quality of the refined analysis is evident in the appearance of resolved, solvent (water) molecules, as shown in Fig. 8.32. Further practical details of protein analysis are discussed in detail in Chap. 10.

8.10 Computer Prediction of Crystal Structures

Recent work has involved computer programs that fit structures to a given molecular conformation and minimize the lattice energy of the chosen structure model so as to obtain an energetic "best fit." As an example, we consider the known crystal structure of 5-azauracil, Figs. 8.33 and 8.34.

8.10.1 Crystal Structure of 5-Azauracil [58]

A successful modeling procedure requires: (1) an accurate model of the molecule; (2) a formula for the intermolecular force function, and (3) a method for generating close-packed structures. In this example, the molecular structure model was obtained through an energy optimization of an SCF 6-31G** wavefunction, using the program CADPAC [59].

The model for the electrostatic contribution to the lattice energy involved calculating sets of atomic multipoles derived by a distributed multipole analysis [60] of the MP2 6-31G** wavefunction for 5-azauracil. Other intermolecular forces were represented by an isotropic atom–atom repulsion and attraction (dispersion) potential. Thus the lattice energy U was written as

$$U = \sum_{i\in 1, j\in 2} (A_{pp}A_{qq})^{1/2}\exp[-(B_{pp}B_{qq})R_{ij}/2] - (C_{pp}C_{qq})/R_{ij}^{6} \tag{8.121}$$

where atom i in molecule 1 is of type p, atom j in molecule 2 is of type q, and R_{ij} is the i–j intermolecular distance; the best known values [60, 61] of the parameters A, B, and C for the C, H, O, and N atoms were used. This form of the potential function has been found to be successful in representing other small, rigid C, H, O, N molecular species.

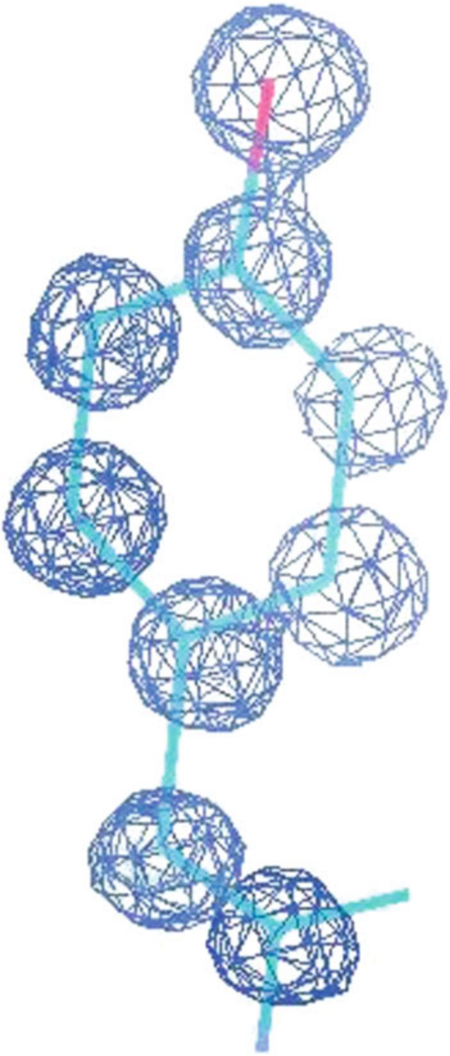


Fig. 8.30 Figures 8.30–8.33 show extracts from the electron density map (COOT, Sect. 10.8.3) of ribonuclease-*A* at 0.85 Å extremely high resolution. The maps are calculated with the coefficients $F_o + (F_o - |F_c|) = 2F_o - |F_c|$ and thus show features of both the electron density and the difference electron density. This figure shows a tyrosyl residue, $HOC_6H_4CH_2CH<$, with the hole of the phenyl ring and the –OH group (at the top) clearly indicated. [Data collected at the Diamond Light Source synchrotron (Sect. 3.1.6) at 100 K and 0.77 Å wavelength.]

The program MOLPAK [62] was used to select close-packed molecular structures, using a simple hard-sphere repulsion potential function. The lattice energies were calculated with the program DMAREL [62], and minimization of the lattice energy was carried out by a modified Newton–Raphson method that optimized the unit-cell dimensions and the rotations and translations of each molecule in the unit cell; the symmetry was generally maintained, although the minimization procedure did not enforce it. The ten structures with the lower initial lattice energies corresponded to space groups $P\overline{1}$, $P2_1$, $P2_1/c$, and $P2_12_12_1$. Structures in these space groups were minimized, and on the basis of the results MOLPAK crystal structures were generated with the less common space groups $P1$, $C2/c$, $Pna2_1$, $Pca2_1$, and $Pbca$. Table 8.25 lists the predicted unit-cell dimensions and hydrogen-bond distances, which are very structure-sensitive, for the six lower energy-minimized structures, together with the corresponding results from the X-ray study.

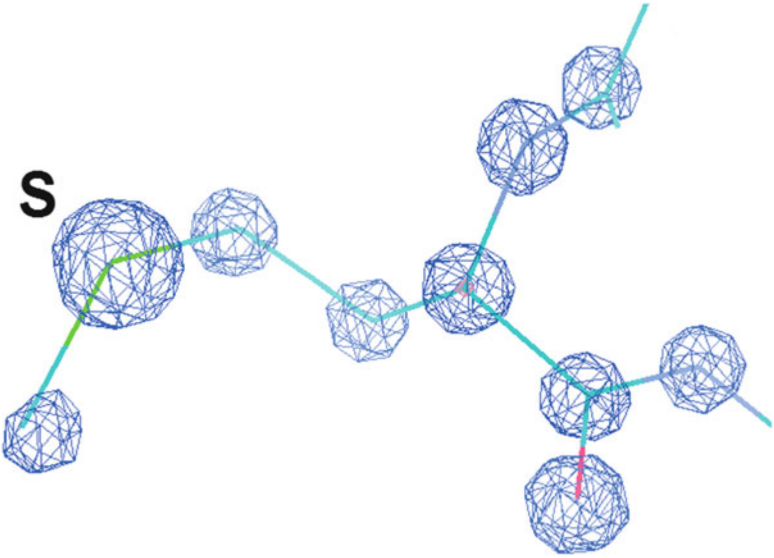


Fig. 8.31 Methionyl residue, $CH_3SCH_2CH_2CH<$, showing the outstanding electron density region around the sulfur atom

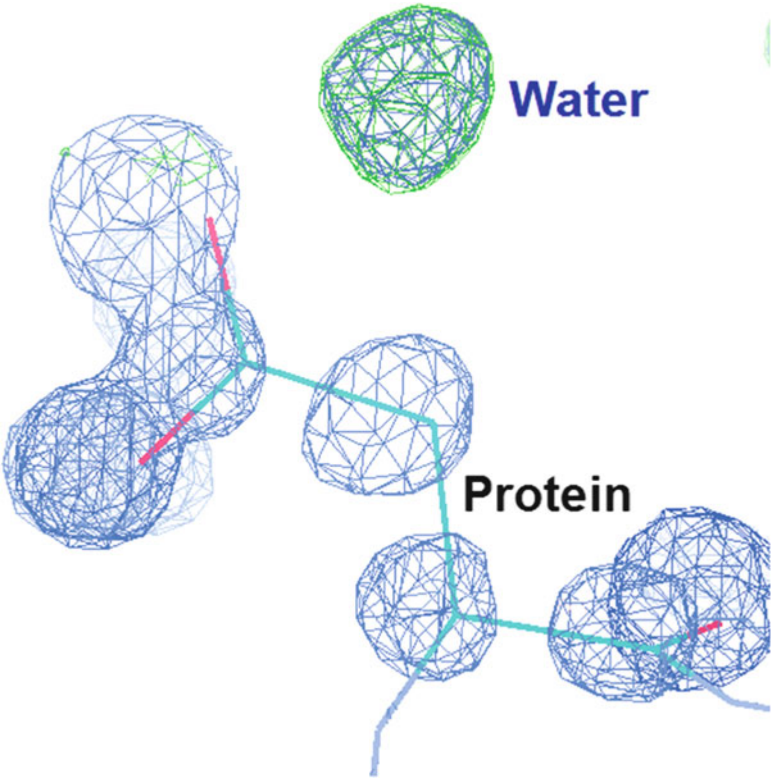


Fig. 8.32 This electron density portion shows a clearly resolved solvent molecule (water), not included in the structure factor calculation

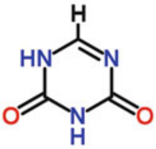


Fig. 8.33 Chemical formula of 5-azauracil, $C_3H_3N_3O_2$

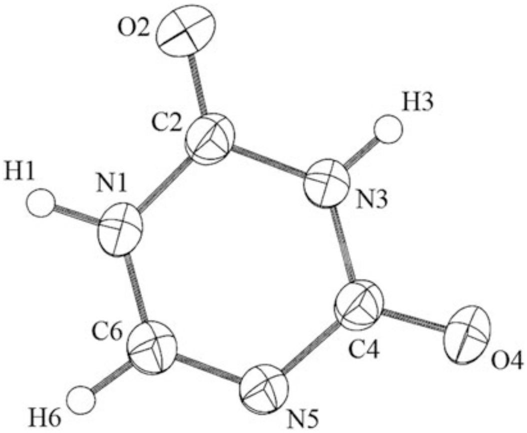


Fig. 8.34 View of the molecule of 5-azauracil molecule in the crystal structure, perpendicular to the molecular plane, showing 50% thermal ellipsoids

Conclusions

The X-ray determination [58] and the best modeled structure of 5-azauracil, labeled * in Table 8.25, are in good agreement. Both are in space group *Pbca* with unit-cell parameters all within 0.3 Å, the modeled structure requiring a unit-cell transformation $\mathbf{a}' = \mathbf{b}$, $\mathbf{b}' = \mathbf{c}$, $\mathbf{c}' = \mathbf{a}$ in order to conform to the X-ray setting. The small energy differences, Table 8.25, of the order of 1–2 kJ mol^{-1}, between the best solution and other close agreements, mean that at this stage in the development of structure prediction it may be necessary to obtain final confirmation by X-ray methods. In fact both the crystal structure and the best predicted structure in this case are very similar, as we have seen, with respect to unit cell and space group and this also applies to the structures in the details of packing and the rippled sheet H-bond formation, Fig. 8.35. Two strong hydrogen bonds, $N_1H_1...O_4$ and $N_3H_3...N_5$, and several weaker intermolecular interactions combine to produce this crinkled sheet structure.

This crystal structure was independently predicted by a search for minima in the lattice energy, as calculated using an ab initio optimized molecular structure and a distributed multipole model for the electrostatic interactions. The global minimum in the search thus corresponded to the same *Pbca* space group, with rms errors in the cell lengths of about 3.7%. There is a larger energy gap separating the observed hydrogen-bonding motif structure from alternative structures, with different hydrogen bonds and connectivity.

8.10.2 Developments in Computer Crystal Structure Prediction

The state of the art in the computer prediction of organic crystal structures has been the subject of several authoritative reviews [63, 64]. Many programs have been written that attempt to predict structures, and a list of them is given in Table 8.26. The assumption is made that the true crystal structure will correspond to the global minimum lattice energy among the predicted structures. Since

Table 8.24 Crystal data for 5-azauracil, $C_3H_3N_3O_2$

Color/shape	Colorless/needles
Temperature (K)	289(2)
Crystal system	Orthorhombic
Space group	*Pbca*
a (Å)	6.5135(3)
b (Å)	13.5217(4)
c (Å)	9.5824(4)
Z	8
Diffractometer/scan	CAD4/$\omega - 2\theta$
Radiation	Cu $K\alpha$
Wavelength (Å) (graphite monochromator)	1.54178
Crystal dimensions (mm)	0.48, 0.33, 0.14
Independent/observed reflections	779/763
θ (°) range for data collection	6.54–74.178
Corrections applied	Lorentz and polarization
Absorption correction	None
Computer programs	CAD4-Express 1988
Structure solution	SHELX-86
Structure refinement	SHELXL-93
Refinement method	Full matrix least squares on $\|F\|^2$
Data/restraints/parameters	763/0/86
Goodness-of-fit on F^2	1.045
Final *R* indices [$I \geq 2\sigma(I)$]	$R_1 = 0.0337$; $wR_2 = 0.0909$
R indices (all data)	$R_1 = 0.0409$; $wR_2 = 0.1192$

Table 8.25 Crystal structure parameters for modeled structures and the X-ray structure

Space group	$-U$ (kJ mol^{-1})	*a* (Å)	*b* (Å)	*c* (Å)	β (°)	(V_c/*Z*) (Å^3)	NH…O (Å)	NH…N (Å)
*Pbca**	109.4	13.777	9.197	6.814	90	107.9	2.17	1.87
$P2_1/c$	108.2	9.545	7.293	9.710	140.3	107.9	2.51	1.83
Pbca	108.1	7.277	9.730	12.243	90	108.3	2.48	1.82
$P2_1$	106.8	4.767	9.717	4.767	80.3	108.8	2.49	1.81
Pbca	106.4	7.302	9.694	12.296	90	108.8	2.51	1.81
$Pca2_1$	105.2	13.517	3.682	9.129	90	113.6	2.09	1.86
X-ray	–	6.5135	13.5217	9.5824	90	105.5	2.03	1.99

the structure is calculated for atoms at rest, that is, at 0 K, the lattice energy will differ from the lattice-free energy at ambient temperature mainly by an entropic component, which could affect the choice of structure when several results of similar lattice energy are calculated.

8.11 Blind Structure Prediction of the Flexible Molecule 1-Benzyl-1*H*-Tetrazole

1-Benzyl-1*H*-tetrazole is a molecule with 2° of conformational freedom and has an unusual tetrazole functional group, providing a challenge to crystal structure prediction methods. There are too few structures with the C–H tetrazole fragment in the Cambridge Structural Database [65, 66] for any prior expectations of the intermolecular interactions of the tetrazole ring to be made.

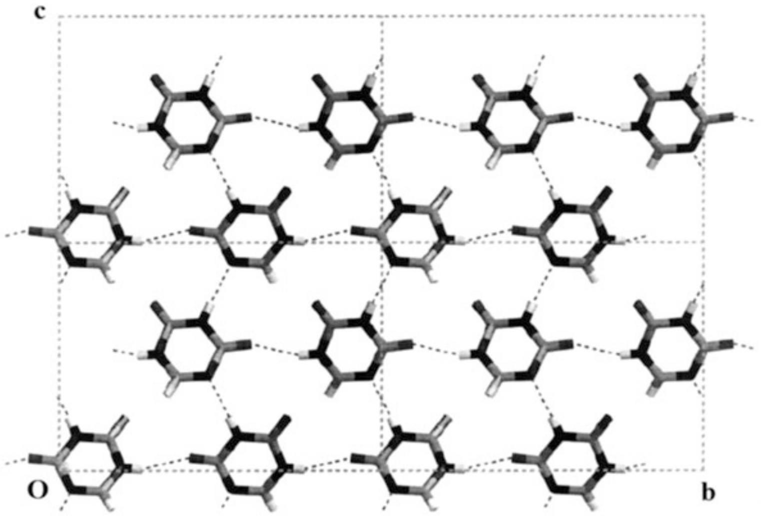


Fig. 8.35 The crystal structure of 5-azauracil: four unit cells of the hydrogen bonded sheet, viewed along the *a* axis; hydrogen bonds are denoted by dashed lines

Table 8.26 Programs for organic crystal structure prediction

Program	Type of molecule used in development	Search type
Chin	Crystal engineering: diketo-piperazines	Monte Carlo simulated annealing with hydrogen-bonding bias
CRYSTALG	Rigid organics: amides, bases	Self-consistent basin-to-deformed-basin mapping global optimization
CRYSCA	Pigments, organometallics	Random search with steepest descent
ICE9	Rigid hydrocarbons	Systematic grid search to generate close-packed structures
MDCP	Small rigid	Constant pressure molecular dynamics to find crude structures
MOLPAK	Energetic materials, rigid	Systematic search for high density structures in common coordination types
DMAREL	Rigid polar and hydrogen bonded	
MPA, extended to Mpg	Small rigid	Lattman systematic, or random generation of expanded trial unit cell
Perlstein	Moderate sized, semi-flexible organics	Aufbau search for low energy one-dimensional and two-dimensional aggregates, primarily for monolayer predictions
PMC	Hydrocarbons	Symmetry adapted grid systematic
Polymorph Predictor	Flexible organics, including pharmaceuticals	Monte Carlo simulated annealing with intermediate clustering
PROMET	Rigid hydrocarbons	Selecting cohesive dimer, ribbons, and layer substructures of partial space group
SySe and PP	Pigments	Grid-based systematic
UPACK	Sugars and alcohols	Systematic grid or random search, with intermediate clustering

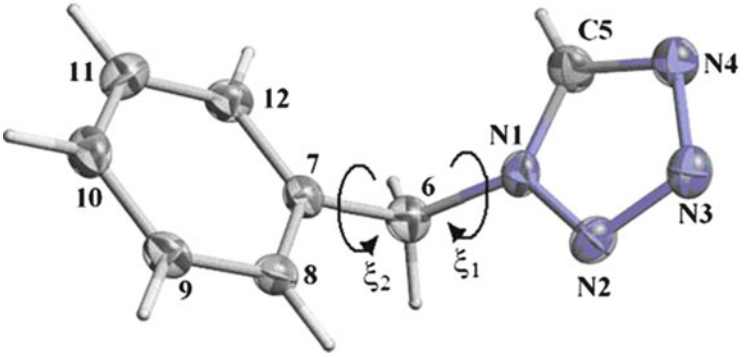


Fig. 8.36 1-Benzyl-1*H*-tetrazole showing thermal ellipsoids at 50 % probability. The torsion angles $\xi_1 =$ C(7)– C(6)–N(1)–C(5) $= -89.8(3)°$ and $\xi_2 =$ C(12)– C(7)– C(6)–N(1) $= 98.5(3)°$ are indicated (drawn with ORTEP/RASTER)

Table 8.27 Crystal data: experimental, computed structure #2, global minimum, GM

	Experimental	#2	GM
a (Å)	7.6843(5)	7.932	9.992
b (Å)	5.5794(4)	5.550	8.175
c (Å)	9.4459(7)	9.437	11.038
β (°)	100.949(4)	101.323	117.504
V_c (Å^3)	397.61(5)	407.331	799.726
Z	2	2	4
Space group	$P2_1$	$P2_1$	$P2_1/c$
ξ_1 (°)	89.91	87.38	86.07
ξ_2 (°)	−98.57	−97.88	−92.07
Density (g cm^{-3})	1.338	1.306	1.330

Crystal Structure of 1-Benzyl-1*H*-Tetrazole

The crystal structure determination and refinement of 1-benzyl-1*H*-tetrazole, $C_8H_8N_4$, has been carried out using the programs SHELXS-86 and SHELXL-97 in WinGX [22], with Mo $K\alpha$ X-ray data measured at 120(2) K [67]. The final data/restraints/parameter ratios were 977/1/113; *R* indices, $I > 2\sigma(I)$: $R_1 = 0.0426$, $wR_2 = 0.0753$; *R* indices, all data: $R_1 = 0.0636$, $wR_2 = 0.0827$. Figure 8.36 shows the molecular structure and atom numbering, and Table 8.27 list the crystal data.

In the crystal structure, the benzyl and tetrazole rings are essentially planar. The individual rings are each coplanar with the inter-ring link atom C(6), and H(5) on C(5) of the five-membered tetrazole ring is also coplanar with its ring. The dihedral angle between the two rings is 68.52(5)°.

Crystal Packing: Weak Hydrogen Bonding

The mode of packing of the molecules in this crystal structure is unusual and unexpected. Molecules such as 1-benzyl-1*H*-tetrazole, based on linked delocalized rings, might be expected to form crystal structures involving pi. . .pi ring stacking bonding. In this case, however, no such interactions occur. Instead the structure is held together through a large number of weak intermolecular hydrogen bonds: [68, 69] twelve of them are of the type CH. . .N; and three are of the weaker type CH. . .C, involving atoms in the phenyl ring, Table 8.28. The result is a structure composed of infinite *S*-shaped layers, as illustrated in Figs. 8.37 and 8.38. Another interesting and unusual by-product of the hydrogen bonding

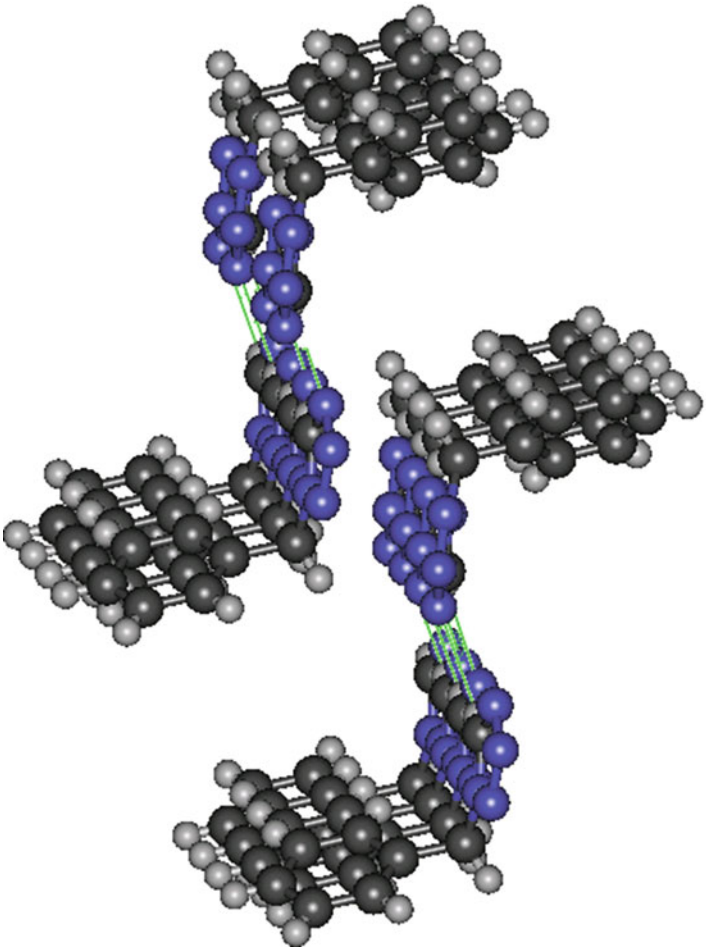
Table 8.28 Hydrogen bonds for 1-benzyl-1*H*-tetrazole

D-H. . .*A*	*d*(*D*–H) (Å)	*d*(H. . .*A*) (Å)	*d*(*D*. . .*A*) (Å)	∠*DHA* (°)
C(8)–H(8). . .C(8)#1	0.95	2.93	3.743(3)	144.3
C(9)–H(9). . .C(6)#1	0.95	2.93	3.668(4)	135.4
C(10)–H(10). . .N(3)#2	0.95	2.91	3.723(3)	143.9
C(10)–H(10). . .N(4)#2	0.95	2.78	3.674(4)	157.2
C(12)–H(12). . .N(2)#3	0.95	2.78	3.617(4)	148.1
C(6)–H(6B). . .N(3)#3	0.99	2.65	3.598(4)	159.7
C(6)–H(6B). . .N(3)#3	0.99	2.93	3.766(4)	142.4
C(5)–H(5). . .N(3)#3	0.90(3)	2.70(3)	3.490(4)	146(2)
C(6)–H(6A). . .N(3)#4	0.99	2.75	3.245(3)	111.5
C(5)–H(5). . .N(4)#5	0.90(3)	2.64(3)	3.292(4)	130(2)
C(12)–H(12). . .N(3)#3	0.95	3.14	3.991(3)	150.5
C(6)–H(6B). . .N(3)#3	0.99	2.93	3.766(4)	142.4
C(9)–H(9). . .N(4)#6	0.95	3.00	3.664(3)	127.8
C(10)–H(10). . .N(4)#6	0.95	3.04	3.679(4)	126.4
C(10)–H(10). . .C(5)#6	0.95	3.15	3.882(4)	135.0

Symmetry transformations used to generate equivalent atoms:
#1: $-x + 1, y - \frac{1}{2}, -z + 2$; #2: $-x + 2, y + \frac{1}{2}, -z + 2$; #3: $x, y + 1, z$
#4: $-x + 1, y + \frac{1}{2}, -z + 1$;#5: $-x + 2, y + \frac{1}{2}, -z + 1$; #6: $x, y, z + 1$

Additionally there is an intermolecular CH. . .pi interaction: C(8)–H(8). . .benzyl centroid #1 = 2.859 Å

Fig. 8.37 1-Benzyl-1*H*-tetrazole : view of the infinite S-shaped layers held together by weak CH...N and CH...C hydrogen bonds (drawn with Accelrys Discovery Studio 3 with coordinates generated in MERCURY)



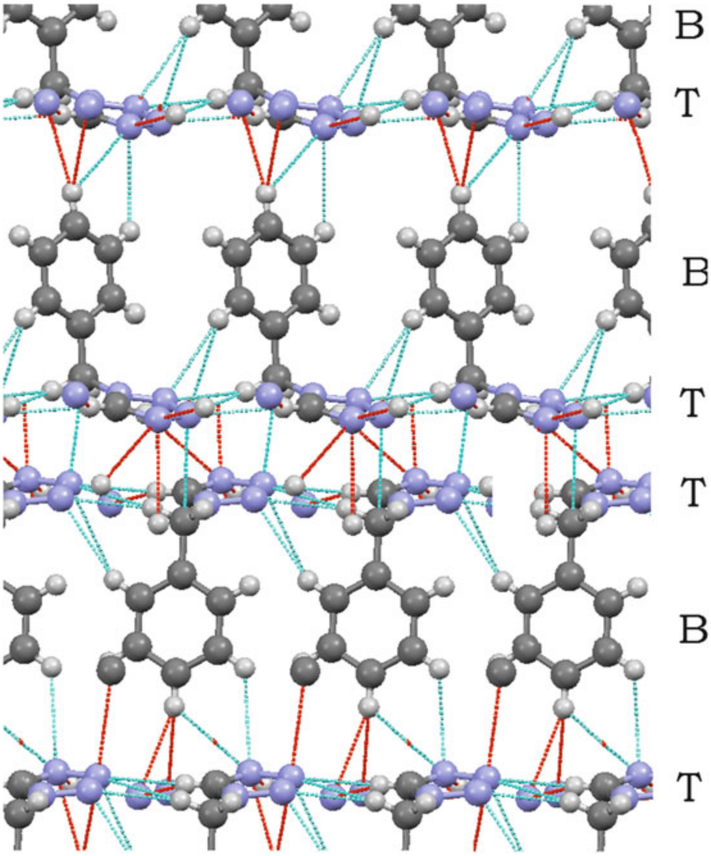


Fig. 8.38 Partial view of the layer structure 1-benzyl-1*H*-tetrazole showing alternative benzyl (*B*) and tetrazole (*T*) layers. Note the repeating layer sequence BTBT | TBTB (drawn with MERCURY)

is the formation of tetrazole clusters, Fig. 8.39, each having a central tetrazole ring coordinated by six other tetrazoles. There is in addition one intermolecular CH. . .pi interaction, but no pi. . .pi interaction.

Structure Prediction

In view of this unusual mode of packing it was decided to initiate a blind crystal structure prediction of 1-benzyl-1*H*-tetrazole. Initially the modeling study investigated possible steric hindrance to rotation about the two torsion angles ξ_1 and ξ_2, Fig. 8.36, which indicated a high degree of flexibility in the molecule. Such a degree of flexibility, where a wide range of very different molecular shapes is possible, is required to be taken into account from the beginning of the structure search. This is a significant difference to the approach that can be used for more rigid molecules [72–79].

Those working on this prediction study were provided only with a sketch of the 1-benzyl-1*H*-tetrazole molecule and were told only that the crystal structure had one molecule per asymmetric unit. The methodology used here for the prediction of crystal structures of flexible molecules had previously been extended to a pharmaceutical-like molecule with seven torsion angles linking four aromatic rings and a peptide group [73], in the Fifth Blind Test of Crystal Structure Prediction [74]. The search using the program CrystalPredictor [75] covered the 59 most commonly occurring space groups and generated about 170000 structures. The lattice energy, $E_{\mathrm{latt}} = U_{\mathrm{inter}} + \Delta E_{\mathrm{intra}}$, where U_{inter} is the intermolecular packing energy and $\Delta E_{\mathrm{intra}}$ is the energy penalty for changing the conformation of the molecule, was minimized by varying the cell parameters and torsion angles ξ_1 and ξ_2.

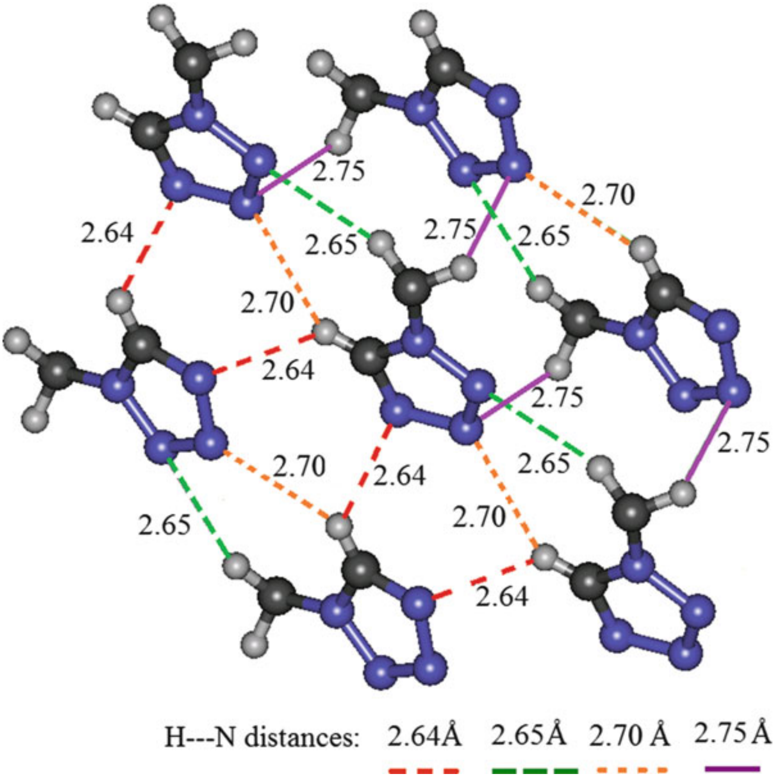


Fig. 8.39 1-Benzyl-1*H*-tetrazole packing, showing an unusual tetrazole cluster. (drawn with Accelrys Discovery Studio 3 from coordinates generated in MERCURY)

At this stage, the lattice energy was calculated crudely by using a grid of ab initio calculations for ΔE_{intra} and an atomic point charge model and empirical repulsion-dispersion model for U_{inter}. This reduced the search to 44000 unique structures. At this stage, for each of the lowest 10000 crystal structures, the energy of the molecule and its charge distribution were calculated at the PBE0/6-31G (d,p) level of theory using GAUSSIAN [76], to provide a better estimate of ΔE_{intra}, and a more accurate representation of the charge density in terms of a distributed multipole model [77]. This was combined with an atom–atom exp–6 repulsion-dispersion potential, using parameters that had been fitted to azahydrocarbon [60] and polar crystal structures [78], which were assumed transferable to this tetrazole. The lattice energy of the crystal structure was minimized with DMACRYS [79], with the molecule held rigidly. One hundred most stable structures were then further refined by allowing the molecular conformation to adjust more accurately to the intermolecular forces using the program CrystalOptimizer [73, 80] to combine the GAUSSIAN calculations on the conformations of the isolated molecule with the crystal structure minimized by DMACRYS. Finally, the effect of the crystal environment on the conformational energies ΔE_{intra} and charge density was estimated by calculating the conformational energy and charge distribution in a polarizable continuum, with the dielectric constant ε equal to 3, a value typical of organic molecules [81], using the Polarizable Continuum Model [82] as implemented in GAUSSIAN [76] at the PBE0/6-31+G(d) level of theory.

Results

The two most stable crystal structures on the crystal energy landscape, Fig. 8.40, are separated by less than 0.1 kJ mol^{-1} but are very different from one another in terms of structural features. One of these structures, labeled #2, is in the same space group, $P2_1$, as determined experimentally and has very similar unit-cell dimensions, Table 8.27. This computed structure gives an excellent overlay with the

Fig. 8.40 Crystal energy landscape of 1-benzyl-1*H*-tetrazole. Each point represents a crystal structure that is a lattice energy minimum, using PCM ($\varepsilon = 3$) at PBE0/6-31+G level of theory molecular calculations. The structure that matches the experimental crystal structure, optimized with the same lattice energy model, is labeled with its energy rank #2. The global minimum structure (#1) is labeled GM

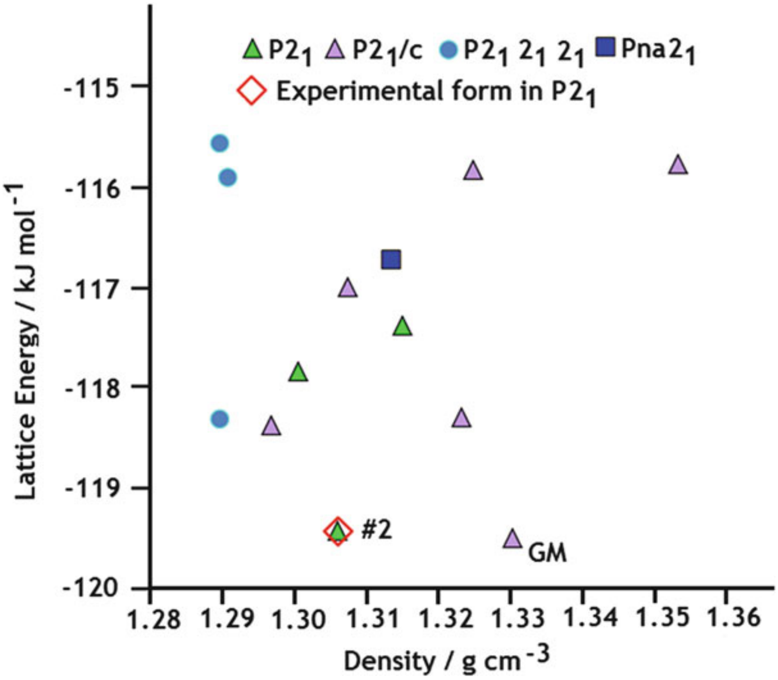
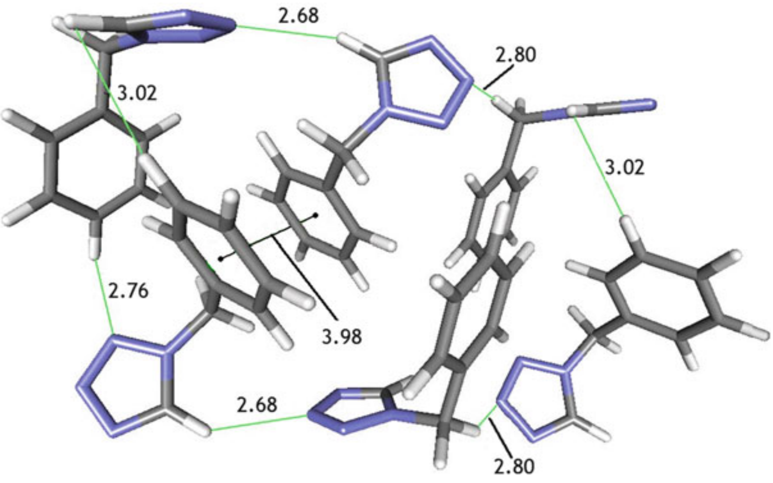


Fig. 8.41 Global minimum predicted structure of 1-benzyl-1*H*-tetrazole: the packing reveals some of the weak CH. . .N and CH. . .C interactions and the weak pi. . .pi interaction; distances are in Å. There is a strong possibility that this structure may exist as a polymorphic form (drawn with Accelrys Discovery Studio 3 from coordinates generated in MERCURY)



experimental structure using the COMPACK [83] facility as implemented in MERCURY [84], with a root mean square difference in the positions of the non-hydrogen atoms in a 15-molecule cluster of only 0.148 Å. This model #2 provided a starting model that was easily refined in SHELXL-97 against the experimental $F_o(hkl)$ data, which is further proof that it is a genuine reproduction of the crystal structure.

Lowest Energy Predicted Crystal Structure

It is of interest to note that the structure corresponding to the global minimum in lattice energy, Table 8.28, is in space group $P2_1/c$ and presents a very different spatial arrangement as shown in Fig. 8.41. This structure also has dominant weak hydrogen bonds: eight of type CH. . .N; and one

CH. . .C. There is also one weak pi. . .pi interaction (benzyl. . .benzyl) which is indicated in Fig. 8.41. Thermodynamically, this structure could be a polymorph of 1-benzyl-1*H*-tetrazole, but no polymorph screening had been undertaken when it was synthesized.

Conclusions

1-Benzyl-1*H*-tetrazole shows an unusual crystal packing with segregated layers of phenyl and tetrazole interactions. This was predicted blindly from a structure that modeled accurately the electrostatic forces arising from molecular charge density, including the anisotropic forces from the lone pairs and pi-electrons, but otherwise had not been tailored to tetrazole. . .tetrazole interactions. The successful prediction of the experimental structure as one of the two distinct most stable structures shows that the unusual layers do present an optimal compromise between the many different weak hydrogen bonds and other intermolecular interactions.

In structures that can exist in polymorphic modifications, the minimum energy conformation should correspond to the thermodynamically most stable form. In practice however, various factors, such as temperature, rate of crystallization, and nature of solvent, could affect the form that is actually observed. It seems probable that improvements in the precision of the prediction method will involve incorporating more accurate forms of the intermolecular potential into the program, as well as allowing for the possibility of kinetic control of crystallization. While computer prediction clearly has a part to play in crystal structure determination, it will probably remain that X-ray diffraction will be needed for confirmation of the structure and particularly for obtaining accurate molecular geometry.

8.12 Problems

8.1. Choose three of the following reflections to fix an origin in space group $P\bar{1}$, giving reasons for your choice.

hkl	|E|	*hkl*	|E|
705	2.2	$6\bar{1}\,\bar{7}$	3.2
$42\bar{6}$	2.7	203	2.3
$4\bar{3}\,\bar{2}$	1.1	$8\bar{1}\,\bar{4}$	2.1

Are there any triplets which meet the vector requirements of the Σ_2 formula?

8.2. The geometric structure factor formulae for space group $P2_1$ are

$$A = 2\cos 2\pi(hx + lz + k/4)\,\cos 2\pi(ky - k/4)$$
$$B = 2\cos 2\pi(hx + lz + k/4)\,\sin 2\pi(ky - k/4)$$

Deduce the amplitude symmetry and the phase symmetry for this space group according to the two conditions $k = 2n$ and $k = 2n + 1$.

8.3. In space group $P2_1/c$, two starting sets of reflections for the application of the Σ_2 formula are proposed:

	Origin-fixing	Symbols
(a)	041, 117, $\bar{1}\,23$	242, $\bar{1}62$
(b)	223, 012, $13\bar{7}$	111, 162

Using just this information, which starting set would be chosen in practice? Give reasons. What modification would have to be made to the starting set if the space group is $C2/c$?

8.4. The following values of $\ln\left[\sum_j f_j^2(hkl)/\overline{|F_o^2(hkl)|}\right]$ and $\overline{(\sin^2\theta)/\lambda^2}$ were obtained from a set of three-dimensional data for a monoclinic crystal. Use the method of least squares (program LSLI) to obtain values for the scale K (of F_o) and temperature factor B by Wilson's method.

$\ln\left[\sum_j f_j^2(hkl)/\overline{\vert F_o^2(hkl)\vert}\right]$	$\overline{(\sin^2\theta)/\lambda^2}$
4.0	0.10
5.6	0.20
6.5	0.30
7.9	0.40
9.4	0.50

What is the value of the root mean square atomic displacement corresponding to the derived value of B?

8.5. An orthorhombic crystal contains four molecules of a chloro-compound in a unit cell of dimensions $a = 7.210(4)$ Å, $b = 10.43(1)$ Å, $c = 15.22(2)$ Å. The coordinates of the Cl atoms are

$$\tfrac{1}{4}, y, z; \quad \tfrac{3}{4}, \bar{y}, z; \quad \tfrac{1}{4}, \left(\tfrac{1}{2} + y\right), \left(\tfrac{1}{2} + z\right); \quad \tfrac{3}{4}, \left(\tfrac{1}{2} - y\right), \left(\tfrac{1}{2} + z\right)$$

with $y = 0.140(2)$ and $z = 0.000(2)$. Calculate the shortest Cl. . .Cl contact distance and its estimated standard deviation.

8.6. The following data give phase indications for the reflection 771 ($|\mathrm{E}_{\mathbf{h}}| = 2.2$, $\varphi_{\mathrm{calc}} = -14°$) in a crystal of space group $P2_12_12_1$. Determine $\varphi_{\mathbf{h}}$ by both (8.25) and (8.28). For simplicity, let $w_{\mathbf{h}}$ in (8.28) be taken as unity.

k	$\varphi_{\mathbf{k}}$ (°)	**h − k**	$\varphi_{\mathbf{h-k}}$ (°)	$\vert\mathrm{E}_{\mathbf{k}}\vert\vert\mathrm{E}_{\mathbf{h-k}}\vert$
12,10	0	$\bar{5}61$	−37	4.4
$7\bar{1}4$	177	$08\bar{3}$	−180	5.1
$12,0\bar{1}$	90	$\bar{5}72$	−144	4.5
12,01	90	$\bar{5}70$	−90	3.3
613	102	$16\bar{2}$	−64	2.7
$\bar{1}45$	−79	$83\bar{4}$	92	3.7

8.7. Figure 8.13a, b shows unit cells for a hypothetical search model (S) and a target structure (T) respectively. Assuming these structures to be correct, which of the intermolecular vectors indicated in Fig. 8.13b will not actually occur in the Patterson for the search molecule? Explain your answer.

8.8. (a) Why would you not use a molecular graphics package alone to generate coordinates for all atoms in the coumarin derivative shown in Fig. 8.16a?

(b) The z coordinates of all atoms in the coumarin model built with Chem-X in Table 8.16 are all 0.0000. Why is this so?

(c) The unit cell for the coumarin model in Table 8.15 has three sides each of 100 Å and all angles 90°. Why do you think this is so? (Hint: Apparently this is not a real unit cell.)

8.9. Determine graphically the X, Y, and Z Cartesian coordinates in Å of the six atoms of a (planar) benzene ring to be used in Patterson search. The bond lengths are all 1.40 Å and angles 120°. [Hint: Construct a regular hexagon with one side parallel to the X (horizontal) axis and its center at the origin. The X, Y coordinates of atoms 1 and 2 are then obvious. The positions of the other atoms may be generated by applying symmetry. All Z coordinates are of course 0.0]. Check your results with the program INTXYZ, Sect. 13.6.6. If the unit cell shown has the dimensions $a = 2.959$ Å, $b = 5.741$ Å, $c = 10.0$ Å, $\alpha = \beta = \gamma = 90°$, determine the x, y, z fractional coordinates of the six atoms.

8.10. From the definition of |E|, show how a Patterson function with $|(E^2 - 1)|$ values as coefficients leads to a sharpened Patterson function with the origin peak removed.

8.11. When employing Patterson Search methods for structure analysis, under what circumstances would you expect the search molecule to be (a) very similar in size to, and (b) much smaller than, the target molecule? Discuss your answer in some detail.

8.12. A third order Karle–Hauptman determinant for a centrosymmetric crystal may bewritten in the form

$$\begin{vmatrix} E(0) & E(\mathbf{h}) & E(\mathbf{k}) \\ E(-\mathbf{h}) & E(0) & E(-\mathbf{h}+\mathbf{k}) \\ E(-\mathbf{k}) & E(-\mathbf{k}+\mathbf{h}) & E(0) \end{vmatrix} \geq 0$$

If $\mathbf{k} = \mathbf{h}$, form and evaluate the determinant. If $E(0) = 3$ and $|E(\mathbf{h})| = |E(2\mathbf{h})| = 2$, determine the sign of $E(2\mathbf{h})$.

8.13. Consider the three triplets 101, $21\bar{2}$, $\bar{1}\,3\bar{3}$; 101, $21\bar{2}$, $\bar{1}\,\bar{1}3$; and 101, $21\bar{2}$, $\bar{3}\,\bar{1}1$. Decide whether each triplet is a structure invariant, a structure seminvariant, or neither, and give reasons. Can any of these triplets be used to specify an origin in space group $P1$? Explain your conclusions.

8.14. Consider a hypothetical crystal structure with a single atom at $x = 0.3$, $y = 0.2$, and $z = 0.1$. Assume that $f(103) = 1.0$, and calculate $|F(103)|$ and $\varphi(103)$ (a) in space group $P2_1$ (b) in space group $P2_12_12_1$ when the 2_1 axis parallel to y (i) passes through the point 0, 0, 0, and (ii) when it is placed in the standard orientation for this space group. Comment on the results.

References

1. Barrett AN, Palmer RA (1969) Acta Crystallogr B25:688
2. Palmer HT (1973) Private communication
3. Sayre D (1952) Acta Crystallogr 5:60
4. Lonsdale K (1929) Proc R Soc A123:494
5. Hauptman H, Karle J (1953) Solution of the phase problem, 1. The centrosymmetric crystal, American Crystallographic Association monograph. Polycrystal Book Service, New York
6. See Bibliography, Woolfson
7. See Bibliography, Chapter 1
8. Karle J, Karle IL (1966) Acta Crystallogr 21:849
9. Barrett AN, Palmer RA, loc. cit.
10. Karle J, Hauptman H (1950) Acta Crystallogr 3:181
11. Germain G, Main P, Woolfson MM (1971) Acta Crystallogr A27:8
12. Main P et al (1980) MULTAN 80: a system of computer programs for the automatic solution of crystal structures from x-ray diffraction data. University of York, England
13. Rogers D (1980) In: Ladd MFC, Palmer RA (eds) Theory and practice of direct methods in crystallography. Plenum, New York
14. Stroud RM (1973) Acta Crystallogr B29:60
15. Sheldrick GM (1990) Acta Crystallogr A46:467
16. Hauptman H (1972) Crystal structure determination: the role of the cosine seminvariants. Plenum, New York
17. Hauptman H (1972) Acta Crystallogr B28:2337
18. Sheldrick GM (1982) In: Sayre D (ed) Computational crystallography. Clarendon Press, Oxford
19. Müller P (2006) Crystal structure refinement. Oxford University Press, New York

20. Sheldrick GM (2008) Acta Crystallogr A64:112
21. The SHELX-97 manual. http://shelx.uni-ac.gwdg.de/SHELX/shelx.pdf
22. Farrugia LJ (1999) J Appl Crystallogr 32:837
23. http://shelx.uni-ac.gwdg.de/SHELX/applfrm.htm
24. http://www.chem.gla.ac.uk/~louis/software/wingx/
25. Giacovazzo C (1974) Acta Crystallogr A30:631
26. Palmer RA et al (2010) Med Chem Commun 1:45
27. Rossman MG, Blow DM (1961) Acta Crystallogr 14:641
28. Braun PB, Hornstra J, Leenhouts JI (1969) Philips Res Rep 42:85
29. http://www.ccp4.ac.uk/
30. See Bibliography, Chapter 7
31. http://download.cnet.com/ACD-ChemSketch-Freeware/3000-2054_4-10591465.html
32. Lisgarten JN et al (2003) J Chem Crystallogr 33:149
33. Gavuzzo E et al (1974) Acta Crystallogr B30:1351
34. http://www.ccdc.cam.ac.uk/
35. Tanczos AC et al (2004) Comput Biol Chem 28:375
36. EPSRC Data Collection Service, University of Southampton, UK
37. http://www.ccp14.ac.uk/tutorial/snb/
38. Miller R et al (1994) J Appl Crystallogr 27:613
39. DeTitta GT et al (1994) Acta Crystallogr A50:203
40. idem. ibid. (1994) A50:210
41. Hauptman H (2002) Z Kristallogr 217:406 and references therein
42. Deacon AM et al (1998) Proc Natl Acad Sci U S A 95:9284
43. Mendham AP et al (2011) J Chem Crystallogr 41:1323
44. http://www2.lv.psu.edu/jxm57/irp/chisquar.html
45. Fisher RA, Yates F (1974) Statistical tables for biological agricultural and medical research, 6th edn. Oliver & Boyd, Edinburgh
46. Duax WL et al (1975, 1984) Atlas of steroid structure, vols 1 and 2. IFI/Plenum, New York
47. http://www.ccdc.cam.ac.uk/index.php
48. http://home.rcsb.org
49. http://cds.dl.ac.uk/cds/datasets/crys/icsd/llicsd.html
50. http://www.nist.gov/index.html
51. http://cds.dl.ac.uk/cds/datasets/crys/cweb/cweb.html
52. http://www.bath.ac.uk/library/subjects/chem/core/daresbury.html
53. El-Shora AL et al (1984) J Crystallogr Spectrosc Res 14:89
54. Husain J et al (1982) Acta Crystallogr B38:2845
55. Moss DS, Morffew A (1981) Comput Chem 6:1–3
56. Bibliography, Chapter 6
57. Palmer RA et al (2013) (to be published)
58. Potter BS et al (1999) J Mol Struct 485–486:349
59. Amos RD et al (1992) The Cambridge analytical derivatives package, 5th edn
60. Williams DE, Cox SR (1984) Acta Crystallogr B40:404
61. Cox SR et al (1981) Acta Crystallogr A37:293
62. See Appendix D
63. Price Sarah L (2003) In: Atwood J, Steed J (eds) Encyclopedia of supramolecular chemistry. Marcel Dekker, New York
64. Beyer T et al (2001) CrystEngComm 3:178
65. Allen FH (2002) Acta Crystallogr B58:380
66. van de Streek J (2006) Acta Crystallogr B62:567
67. Spencer J, Patel H, Deadman JJ, Palmer R, Male L, Coles S, Uzoh O, Price S (2012) CrystEngComm 14(20), 6441
68. Desiraju GR (2002) Acc Chem Res 35:565
69. Herrebout WA, Suhm MA (2011) Phys Chem Chem Phys 13:13858
70. Potter BS et al (1999) J Mol Struct 486:349
71. Nowell H et al (2006) Acta Crystallogr B62:642
72. D'Oria E et al (2010) Cryst Growth Des 10:1749
73. Kazantsev AV et al (2011) Int J Pharm 418:168
74. Bardwell DA, Adjiman CS, Arnautova YA, Bartashevich E, Boerrigter SX, Braun DE, Cruz-Cabeza AJ, Day GM, la Valle RG, Desiraju GR, van Eijck BP, Facelli JC, Ferraro MB, Grillo D, Habgood M, Hofmann DW, Hofmann F, Jose K, Karamertzanis VPG, Kazantsev AV, Kendrick J, Kuleshova LN, Leusen FJ, Maleev AV, Misquitta AJ,

Mohamed S, Needs RJ, Neumann MA, Nikylov D, Orendt AM, Pal R, Pantelides CC, Pickard CJ, Price LS, Price Sarah L, Scheraga HA, van de Streek J, Thakur TS, Tiwari S, Venuti E, Zhitkov IK (2011) Acta Crystallogr B 67:535

75. Karamertzanis PG, Pantelides CC (2007) Mol Phys 105:273
76. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery J, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski J, Ayala Morokuma SPY, Voth K, Salvador GA, Dannenberg P, Zakrzewski JJ, Dapprich VG, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian Inc., Wallingford
77. Stone AJ (2005) J Chem Theory Comput 1:1128
78. Coombes DS et al (1996) J Phys Chem 100:7352
79. Price SL et al (2010) Phys Chem Chem Phys 12:8478
80. Kazantsev AV et al (2010) In: Adjiman CS, Galindo A (eds) Molecular system engineering, vol 6. Wiley-VCH Verlag GmbH, Weinheim, p 1
81. Cooper TG et al (2008) J Chem Theory Comput 4:1795
82. Cossi M et al (2002) J Chem Phys 117:43
83. Chisholm JA, Motherwell S (2005) J Appl Crystallogr 38:228
84. Macrae CF et al (2006) J Appl Crystallogr 39:453
85. http://www.povray.org
86. http://www.chem.gla.ac.uk/~louis/ortep3/
87. Ladd MFC, Palmer RA (1993) Structure determination by X-ray crystallography, 3rd edn. Springer, New York

Bibliography: Crystallographic Computing

Ahmed FR et al (eds) (1970) Crystallographic computing. Munksgaard, Copenhagen

Diamond R, Ramaseshan S, Venkatesan K (eds) (1980) Computing in crystallography. Indian Academy of Sciences, Bangalore

Flack HD, Párkányi L, Simon K (eds) (1993) Crystallographic *computing*, vol 6, A window on modern crystallography. Oxford University Press, Oxford

http://www.iucr.org/resources/commissions/crystallographic- computing/schools

http://crystallography.org.uk/

Isaacs MR (1988) Crystallographic computing, vol 4, Techniques and new technologies. Oxford University Press, Oxford

Mitchell DJ, Lippert EL (1965) Acta Crystallogr 18:559

Mitchell EJ (1965) PhD thesis, Vanderbilt University

Moras D, Podjarny AD, Thierry JC (eds) (1991) Crystallographic computing, No. 5: from chemistry to biology, International union of crystallography symposia. Oxford University Press/International Union of Crystallography, Oxford

Müller P (ed) (2006) Crystal structure refinement: a crystallographer’s guide to SHELXL. International Union of Crystallography, Oxford

Pepinsky R, Robertson JM, Speakman JC (eds) (1961) Computing methods and the phase problem in X-ray crystal analysis. Pergamon, Oxford

Rollett JS (ed) (1965) Computing methods in crystallography. Pergamon, Oxford

Sheldrick GM (2008) A short history of SHELX. Acta Crystallogr A64:112

Sheldrick GM (1985) Crystallographic computing, vol 3, Data collection, structure determination, proteins and databases. Oxford University Press, Oxford

Watkin D (1994) Acta Crystallogr A39:158

Direct Methods

ACORN ab initio phasing. http://www.ccp4.ac.uk/dist/ccp4i/help/modules/exptphase.html#acorn; http://www.ccp4.ac.uk/newsletters/newsletter40/07_acorn.pdf

Giacovazzo C (1980) Direct methods in crystallography. Academic, New York

Ladd MFC, Palmer RA (eds) (1980) Theory and practice of direct methods in crystallography. Plenum, New York

Rodriguez DD et al (2009) Nat Methods 6:651

Woolfson MM, Hai-Fu H (2005) Physical and non-physical methods of solving crystal structures. Cambridge University Press, Cambridge

Woolfson MM (1997) An introduction to X-ray crystallography, 2nd edn. Cambridge University Press, Cambridge

Woolfson MM (1961) Direct methods in crystallography. Clarendon Press, Oxford

Chemical Data

Allen FH et al (1987) J Chem Soc Perkin Trans II:S1–S19

Chemical Database Service (2011). http://cds.dl.ac.uk/

Goodman Group, University of Cambridge (2011). http://www-jmg.ch.cam.ac.uk/

Kitaigorodskii AI (1973) Molecular crystals and molecules. Academic, New York

Megaw HD (1973) Crystal structures. Saunders, Philadelphia

NIST Chemical Webbook (2011). http://webbook.nist.gov/

Sutton LE (ed) (1958, suppl 1965) Tables of interatomic distances and configurations in molecules and ions. The Chemical Society, London