Rotation and translation functions and the molecular replacement method

Sometimes it may happen that we know the three-dimensional structure of a molecule and we are interested in solving the same structure in a different

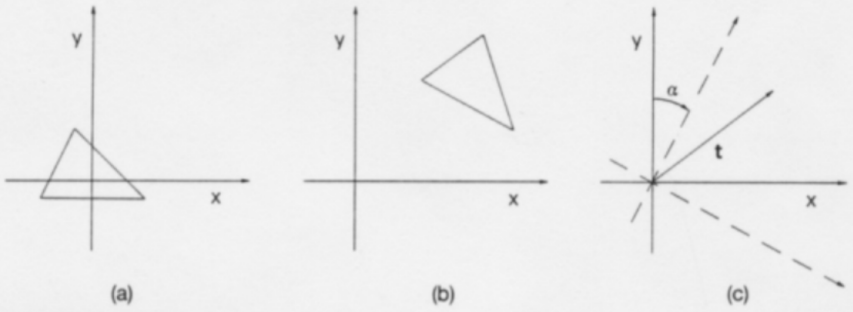

Fig. 8.14. A simplified two-dimensional illustration of (8.53). In (a) and (b) the same 'molecule' is represented in different positions with respect to the same reference system. By a rotation of an angle α and a translation of $\boldsymbol{t}$, the object in (a) can be superimposed to that in (b).

space group. At other times we have reasons to believe that the conformation of a protein is quite similar to that of another that has been previously solved, which is often the case for the same protein from different species. In all the cases mentioned above, six variables, three rotational and three translational, will approximately describe the transformation from one set of coordinates to the other. In fact, if we call $\boldsymbol{X}$ the set of vectors† representing the atoms of the original molecule and $\boldsymbol{X}'$ the transformed ones, the transformation is simply described by:

$$\boldsymbol{X}' = [\mathbf{C}]\boldsymbol{X} + \boldsymbol{t} \tag{8.53}$$

where $[\mathbf{C}]$ is a matrix that rotates the coordinates $\boldsymbol{X}$ into the new orientation and $\boldsymbol{t}$ is a translation vector. Equation (8.53) is illustrated for a two-dimensional situation in Fig. 8.14, where a 'molecule' formed by three point atoms can be superimposed to an identical molecule in a different orientation by the translation of a vector $\boldsymbol{t}$, after the rotation of an angle α.

As mentioned in Appendix 5.B (p. 000), the technique of positioning a molecule or a fragment of known structure in a crystal cell is called **molecular replacement**. In principle it is possible to simultaneously search for the six variables which minimize the difference between F_{obs} and F_{calc}, but in practice this is a very hard task, even for the fastest computer.‡ The solution of the problem was pioneered by Rossmann and Blow,[45] who explored the possibility of finding the orientation of similar subunits in a crystal cell without any knowledge about the translation $\boldsymbol{t}$, making use of the Patterson function. After the correct orientation has been found, a search for the translation vector can be carried out (a collection of papers on molecular replacement is found in the book by Rossmann).[4] Let us first describe the methodology and the problems connected with the rotation function.

The first step in molecular replacement: the rotation function

The idea of the rotation function can be easily understood by a simple two-dimensional example. In Fig. 8.15(a) a 'molecule' of three idealized

† In the following discussion all the rotations will be performed in a Cartesian reference system. It is assumed that, if required, an appropriate orthogonalization is applied before a rotation is performed.

‡ This statement is becoming untrue, due to growing availability of computing power. Subbiah and Harrison[44] have shown, in the test case of the human histocompatibility antigen, that an exhaustive three-dimensional search at low resolution can be performed. The correct solution can also be obtained, starting from a random position, using the simulated annealing approach (see p. 569).

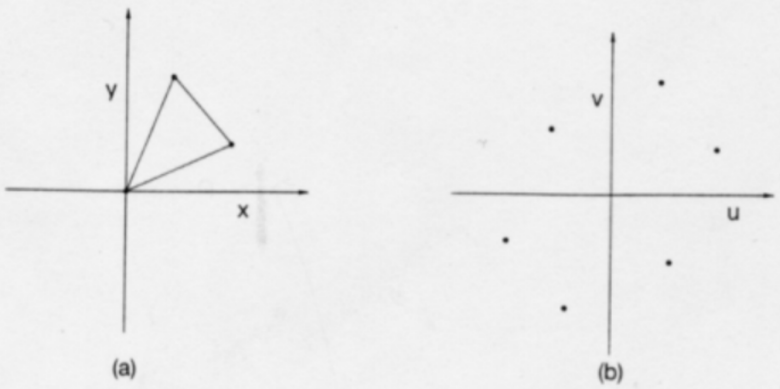

Fig. 8.15. An isolated, simplified 'molecule' of three atoms. Its self-convolution is shown in (b): since atoms are considered as points, it is everywhere 0, except when two points superimpose exactly.

point atoms is represented isolated, in a orthogonal reference system. Let us imagine a two-dimensional lattice of similar 'molecules' in a different orientation. In the lattice of Fig. 8.16(a) the unit cell is made by two of such molecules, related by a twofold axis, denoted by 1 and 2. Maxima of its idealized Patterson function, shown in Fig. 8.16(b), can be divided in two categories: those arising from intramolecular vectors, or self-vectors, and those from intermolecular or cross-vectors. Maxima belonging to the first class are indicated in the figure by circled points and are confined to a short distance from the origin. It is easy to see that by a simple rotation of 112° anti-clockwise the isolated molecule can be superimposed to molecule 1 of Fig. 8.16(a), after an appropriate translation, or, by a rotation of 292°, to molecule 2. The self-convolution function of the isolated molecule (Fig. 8.15(b)) can also be superimposed to the Patterson of the crystal if we perform the same rotation. Let us define a function $R(\mathbf{C})$:

$$R(\mathbf{C}) = \int_V P_{\text{cryst}}(\boldsymbol{u}) P_{\text{mol}}(\mathbf{C}\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \tag{8.54}$$

where $\mathbf{C}$ is a matrix that rotates the coordinates of the model molecule with respect to the reference system of the crystal, $P_{\text{cryst}}(\boldsymbol{u})$ is the Patterson

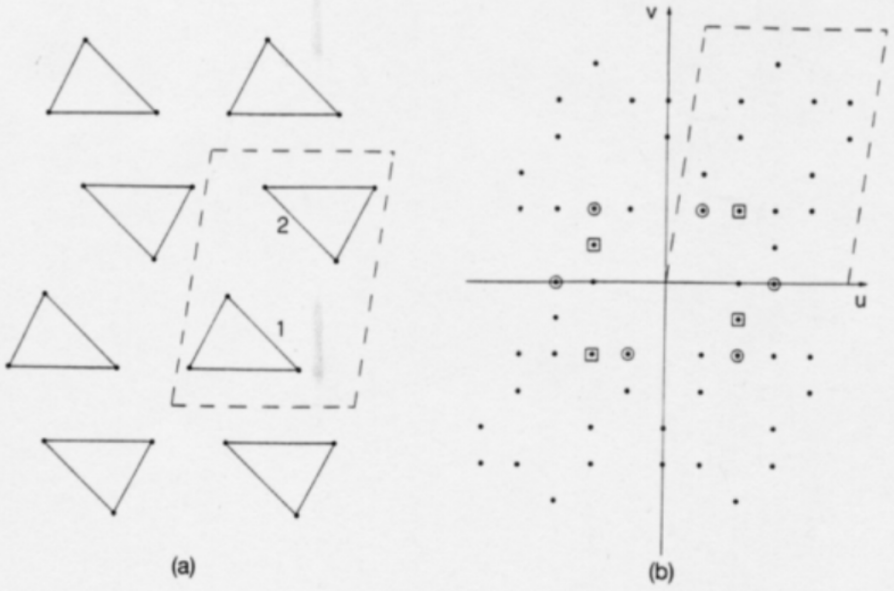

Fig. 8.16. (a) A portion of a two-dimensional lattice of a molecule identical to that of Fig. 8.15 (the unit cell is dashed). (b) Its corresponding Patterson map. Circled points indicate self-vectors. Squared points are cross-vectors close to the origin: some of the points of Fig. 8.15(b) accidentally superimpose to them during rotation, giving rise to false maxima in the rotation function.

function of the crystal and $P_{mol}(\mathbf{C}\boldsymbol{u})$ is the self-convolution function of the isolated molecule, rotated by $\mathbf{C}$. The function $R(\mathbf{C})$ will have a maximum when the peaks of the two functions superimpose, at least partially. The calculation of function $R(\mathbf{C})$ for all the possible values of the rotational variables will allow us to determine the orientation of the known molecule in the reference system of our crystal.

The right-hand side of (8.54) can be Fourier transformed[45,46] and reduced, neglecting a constant, to:

$$R(\mathbf{C}) = \sum_{p} \sum_{h} F_{mol}(\boldsymbol{p})^2 F(\boldsymbol{h})^2 G_{\boldsymbol{h},\boldsymbol{h}'}. \tag{8.55}$$

$F(\boldsymbol{h})$ are the Fourier coefficients of the crystal and $F_{mol}(\boldsymbol{p})$ the coefficients of the Fourier transform of the isolated molecule, rotated by $\mathbf{C}$ ($\boldsymbol{h}$, $\boldsymbol{h}'$ are used here to indicate different terms of (hkl) values, $\boldsymbol{p}$ represents a point in reciprocal space of a continuous transform). $G_{\boldsymbol{h},\boldsymbol{h}'}$ is an interference function whose magnitude depends on $\boldsymbol{h}$, $\boldsymbol{h}'$, and the volume used in the integration of (8.54).

The function $R(\mathbf{C})$ can be evaluated in real space using (8.54) or in reciprocal space, using (8.55). In both cases the computing time is strongly dependent on the sampling chosen, which in turn is related to resolution. In real space P_{cryst} and P_{mol} must be sampled finely enough for the resolution selected (generally this means a value around 1/2 or 1/3 the d spacing). The volume of integration is a sphere whose radius depends on the size of the isolated molecule, and this value determines the steps of the angular variables used in evaluating $R(\mathbf{C})$. In reciprocal space problems are quite similar, since $F_{mol}(\boldsymbol{p})$ is a continuous function, defined over all the reciprocal space. The isolated molecule can be put in an artificial cell, generally a cube whose edges can be about two to three times the size of the molecule, and the continuous function evaluated with a sampling appropriate to the resolution used. In practice, since (8.55) is dominated by large Fourier coefficients, it is possible to limit the numbers of $F(\boldsymbol{p})$ used.

A faster but more complex approach in evaluating the rotation function, the so-called fast-rotation function, has been devised by Crowther.[47] If we express the Patterson function in terms of spherical polar coordinates, (r, θ, φ), for a rotation $\mathbf{C}$, corresponding to the three angles α_1, α_2, α_3, the rotation function can be written:

$$R(\mathbf{C}) = \int P_{cryst}(r, \theta, \varphi) \mathrm{R}P_{mol}(r, \theta, \varphi) r^2 \sin\theta \, \mathrm{d}r \, \mathrm{d}\theta \, \mathrm{d}\varphi \tag{8.56}$$

where $\mathrm{R}P_{mol}$ is P_{mol} after a rotation $\mathbf{C}$. Equation (8.56) can be expanded using Bessel functions, more appropriate to a rotation group than a Fourier series, well suited for translation operations. The use of Bessel functions requires a lot of difficult mathematics, outside the scope of this book, but the final result is that $R(\mathbf{C})$ can be evaluated as a summation of two terms, one of them independent of the rotation itself. The computation time is consequently greatly reduced with respect to the use of (8.54) or (8.55).

The rotation matrix C and the choice of variables

Rotation is usually performed with respect to an orthogonal system, making use of different rotational variables. Quite common are the Eulerian

rotation angles θ_1, θ_2, and θ_3, illustrated in Fig. 2.3(a): θ_1 is the rotation angle about the z axis and is positive when the rotation is clockwise looking from the origin; θ_2 is a rotation about the new x axis and θ_3 a rotation about the new z axis. The matrix **C** describing such a rotation is given in (2.32b).† An appropriate rotation for the three angles will cover all the space (see p. 72), but if the Patterson map presents some rotational symmetry, the rotation function will also have symmetry and a partial rotation will be sufficient.

The symmetry of the rotation function is a combination of the symmetries of the two Patterson functions, P_{cryst} and P_{mol}.[48] The Eulerian angles make easy to describe the symmetry of the rotation function.[49] Any triplets of angles θ_1, θ_2, and θ_3 can be considered as a point of a three-dimensional system, whose unit cell has dimensions 2π in all directions: a rotation α is in fact equivalent to $\alpha + 2\pi$. The resulting rotation space groups are some of those described in the *International tables for x-ray crystallography*.‡

A disadvantage in using θ angles is that when θ_2 is small, θ_1 and θ_3 represent a rotation about nearly the same axis, and maxima will resemble strips rather than maxima. The distortion effect can be avoided if a combination of Eulerian angles is used instead:[50]

$$\theta_+ = \theta_1 + \theta_3, \qquad \theta_- = \theta_1 - \theta_3 \qquad \theta_2 = \theta_2. \tag{8.57}$$

A different possibility is the use of spherical polar angles, φ, ψ, and χ (Fig. 2.3(b)). Angles φ and ψ define a spin axis, and a rotation of χ around this axis is performed. Polar angles are very useful when a particular direction has to be exploited or when a defined rotation has to take place, as is sometimes the case for self-rotation (see p. 558).

Translation functions

Once the orientation of a known molecule in an unknown cell has been found, the next step is the determination of its absolute position. Only when one molecule is present in space group P1 is this problem non-existent, since in this case the origin of the crystal cell can be chosen arbitrarily with respect to all three axes. In all the other cases, when the reference molecule, exactly oriented, is translated in the unknown cell, symmetry-related molecules move accordingly and all the intermolecular vectors change: only when all the molecules in the crystal cell are in the correct position, the calculated Patterson cross-vectors superimpose to those of the observed Patterson (intramolecular vectors are insensitive to translation). Figure 8.17 illustrates the method. Molecule 1 is positioned in the crystal cell of the unknown structure in the correct orientation: $\boldsymbol{s}_1$ is the vector defining its position with respect to the origin. Since we do not know yet the correct position of the molecule in the cell, $\boldsymbol{s}_1$ is arbitrarily chosen. Molecule 2 is generated by the twofold axis, and its position is defined by vector $\boldsymbol{s}_2$. The correct solution is shown in Fig. 8.17(a), where the correct origin of molecule 1 is indicated by $\boldsymbol{s}_1^0$. As vector $\boldsymbol{s}_1$ varies, all the intermolecular vectors among symmetry-related atoms will change: they will coincide with

† In Chapter 2 the rotation matrix **C** is called $\mathbf{R}_{Eu}$ for Eulerian angles and $\mathbf{R}_{sp}$ for spherical polar angles.

‡ The reader must be warned that the Eulerian rotation matrix is not Hermitian, that is reversing the order of the Patterson functions does not produce the same rotation-equivalent positions.

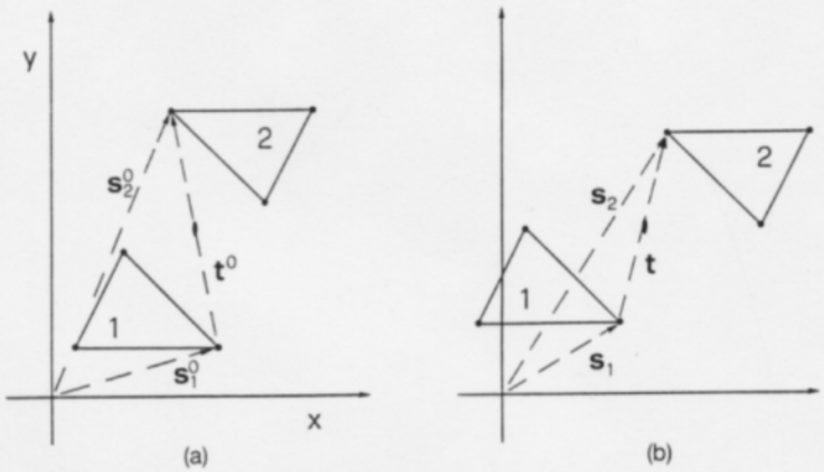

Fig. 8.17. (a) 'Molecules' 1 and 2 correctly oriented and positioned with respect to a symmetry element. s_1^0 is the vector from the origin to an arbitrary point of molecule 1, and s_2^0 the corresponding one for molecule 2, which is generated from 1 by twofold rotation. $t^0 = s_2^0 - s_1^0$ is the translation vector from molecule 1 to 2. (b) Molecule 1 has been translated (but not rotated, so that orientation is unchanged). Molecule 2 has moved accordingly, and now all vectors defining the molecular positions in the cell are changed. Only when $t = t^0$ the two models, and consequently cross-vectors of Patterson maps, superimpose. Determination of s_1^0 from t^0 is straightforward.

those of Fig. 8.17(a) only when $s_1 = s_1^0$, that is when the local origin is correctly defined with respect to the symmetry element. The determination of the translation vector is performed by comparing two Patterson maps, just as before in the rotation case, except that now we are interested in maximizing the superposition of a different class of peaks. The problem described above is in practice quite difficult, since the function representing the superposition is generally very noisy and with many small maxima. Several translation functions have been proposed, and some of them are briefly summarized in Appendix 8.B. To illustrate the principles of translational search, only the T function of Crowther and Blow[51] will be described here, following the treatment of Latman.[52] In the case illustrated in Fig. 8.17, the set of cross-vectors of the calculated Patterson from molecule 1 to molecule 2 can be written as:

$$P_{12}(\boldsymbol{u}) = \int_V \rho_1(\boldsymbol{x})\rho_2(\boldsymbol{x}+\boldsymbol{u})\,\mathrm{d}\boldsymbol{x} \tag{8.58}$$

where ρ_1 and ρ_2 represent the electron density of the two molecules. If molecule 1 is now translated, a new vector s_1 will define its origin. At the same time molecule 2 will move into the cell, and a new function P_{12} can be calculated for every value of s_1. Since we are looking at intermolecular vectors, it is more useful to define the translation as a function of vector $t = s_2 - s_1$, which defines a local origin with respect to a symmetry element. If $P_{\text{obs}}(\boldsymbol{u})$ is the value of the observed Patterson at point $\boldsymbol{u}$, the T translation function is defined as:

$$T(\boldsymbol{t}) = \int_V P_{\text{obs}}(\boldsymbol{u})P_{12}(\boldsymbol{u}, \boldsymbol{t})\,\mathrm{d}\boldsymbol{u}. \tag{8.59}$$

Function T will have a maximum when the two Pattersons superimpose. In reciprocal space, (8.59) can be written:[51]

$$T(\boldsymbol{t}) = \sum_{\mathbf{h}} I_{\text{obs}}(\boldsymbol{h})\boldsymbol{F}_1(\boldsymbol{h})\boldsymbol{F}_1^*(\boldsymbol{h}\mathbf{A}) \exp(-2\pi\mathrm{i}\boldsymbol{h}\boldsymbol{t}) \tag{8.60}$$

where $\boldsymbol{F}_1(\boldsymbol{h})$ is the Fourier transform of the model molecule 1 and $\boldsymbol{F}_1(\boldsymbol{h}\mathbf{A})$ the calculated structure factor of molecule 1 after application of symmetry

operation **A**. Looking at Fig. 8.17, since $\boldsymbol{t} = \boldsymbol{s}_2 - \boldsymbol{s}_1$:

$$T(\boldsymbol{t}) = \sum_{\boldsymbol{h}} I_{\text{obs}}(\boldsymbol{h}) \boldsymbol{F}_1(\boldsymbol{h}) \boldsymbol{F}_1^*(\boldsymbol{h}\mathbf{A}) \exp\left[-2\pi i \boldsymbol{h}(\boldsymbol{s}_2 - \boldsymbol{s}_1)\right]. \tag{8.61}$$

The T function will have a peak at position $\boldsymbol{s}_2 - \boldsymbol{s}_1 = \boldsymbol{t}_0 \equiv \boldsymbol{s}_2^0 - \boldsymbol{s}_1^0$ (or $\boldsymbol{s}_1 - \boldsymbol{s}_2 = -\boldsymbol{t}_0$). The determination of $\boldsymbol{s}_1^0$ from function T is equivalent to solving a Patterson map with only two atoms in the crystal cell. In general the T function does not need to be evaluated for the entire cell: if for example two molecules are related by a screw axis along z, the maximum $\boldsymbol{t}_0$ will have coordinates $(2x, 2y, 1/2)$, that is it will be confined only to section $z = 1/2$.

Other types of translation functions have been developed, and some of the more commonly used in protein crystallography are summarized in Appendix 8.B. A general review on translation functions is reported by Beurskens *et al.*[53]

Self-rotation and self-translation functions: improving the electron density maps

Sometimes more than one molecule is present in the crystallographic asymmetric unit. If we assume that they are identical, or at least very similar, we can take advantage of the independent information present in the structure factors. In fact, if we are able to identify the non-crystallographic symmetry elements relating the independent molecules, the electron density map can be averaged and substantially improved. The presence of three molecules in the asymmetric unit has allowed the solution of the structure of the haemaglutinin of the influenza virus using only SIR phases.[54]

The self-rotation function is very similar to the general rotation function defined in (8.54):

$$R(\mathbf{C}) = \int_V P_1(\boldsymbol{u}) U(\boldsymbol{u}) P_1(\mathbf{C}\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \tag{8.62}$$

where $P_1(\mathbf{C}\boldsymbol{u})$ is the $P_1(\boldsymbol{u})$ Patterson function rotated by matrix $\mathbf{C}$, and $U(\boldsymbol{u})$ is a function which is 1 inside a sphere and 0 elsewhere. The function U is necessary since both Patterson maps extend to all space, but we are interested only in the superposition of self-vectors, confined to a region around the origin of the cell. The sphere defined by U generally has a diameter slightly larger than the maximum supposed molecular dimension.

The choice of polar rotation angles is quite common for self-rotation and deserves a brief comment. Quite often the non-crystallographic symmetry is represented by a rotation axis, in a direction different from the crystallographic ones. In that event, the use of polar angles reduces the search for the position of the axis from a three-dimensional problem to a two-dimensional one: a twofold axis, for example, will correspond to a rotation of 180° around the polar axis χ, and a two-dimensional map (calculated for φ from 0° to 360°, ψ from 0° to 360°, and $\chi = 180°$) will show the presence of the axis. A clear example of that is presented by Evans *et al.*[55]

The definition of the translational component of the non-crystallographic symmetry represents the last and possibly the more difficult step. Let us for example assume that the direction of a twofold non-crystallographic axis is

known. It has been shown[56] that only the component of the translation vector $\boldsymbol{t}$ in the direction of the axis can be determined precisely. The other component of vector $\boldsymbol{t}$, that is that perpendicular to the axis, is intrinsically an imprecise parameter, unless the molecular structure is perfectly known, which is not the case. In general, the self-translation function (analogous to the T function previously described) is used to detect the existence of a translational component of a rotational symmetry.

The steps and the possible different pathways described in the previous paragraphs for the solution of a crystal structure of a macromolecule are summarized in the scheme reported in Fig. 8.18.

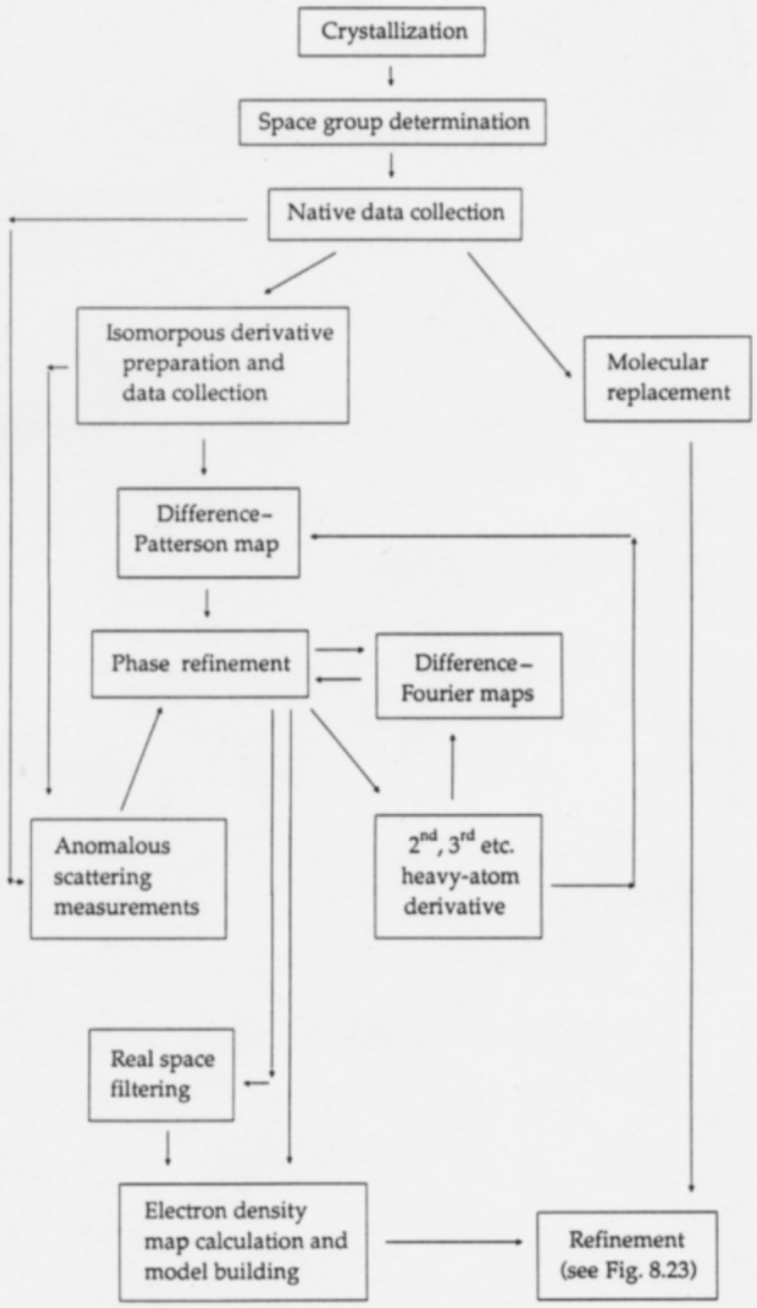

Fig. 8.18. Simplified scheme summarizing some of the possible steps in the determination of the structure of a macromolecule. Some of them can be used alternatively or combined, depending on the specific problem, that is size of the protein, previous knowledge of the structure, number of molecules in the asymmetric unit, and so on.

Practical hints in molecular replacement

It is difficult to underestimate the importance that nowadays the molecular replacement holds in protein crystallography: as explained in the introduction, the number of possible folds of globular proteins is somehow limited and it is expected than in a short period all the possible folds (or a large majority of them) will be known. At that point it should be possible, at least in principle, to solve the structure of any new crystal by molecular replacement, assuming a model similar enough is available. But here comes the most tricky question: how similar the starting model has to be to the actual structure for the method being successful? There is no unambiguous answer to this question, except that based on practice and previous experiences: it is said among crystallographer that if the unknown model presents an identity in primary sequence of more than about 60 per cent with the unknown one, it is expected that the method will be successful. But of course, there are a lot of cases where a smaller similarity has not prevented the achievement of the correct solution.

It is needless to say that, for very difficult cases, special tricks can be used. One of them is represented by the possibility of testing, with very fast computing programs, a very large number of possible solutions. It has for example, been found that quite often the correct orientation for the rotation function corresponds to a peak which is quite far from the maximum, but not lower than the 50 per cent of the highest maxima of the correlation function. So the possibility of fast testing of several solutions with different parameters, as with the software AmoRe,[111] is important in itself.

Another quite successful tool in molecular replacement is represented by the so-called PC-refinement (Patterson-correlation refinement),[112] particularly useful in the presence of macromolecules made up by flexible parts or domains: every solution of the rotation function, before translation, undergoes a special optimization, where the model is divided in parts and each of them is refined separately in order to maximize the correlation.

The steps and the possible different pathways described in the previous paragraphs for the solution of a crystal structure of a macromolecule are summarized in the scheme reported in Fig. 9.44.