9.5 The interpretation of the electron density maps and the refinement of the model

9.5.1 The interpretation of the electron density maps

Once a phase angle estimate for the protein structure factors is available, the calculation of an electron density map is straightforward, using (see Section 2.9):

$$\rho(\boldsymbol{r}) = \sum_{\boldsymbol{h}} F_{\mathrm{P}}(\boldsymbol{h}) \exp(-\mathrm{i}\varphi_{\mathrm{P}}) \exp[-2\pi\mathrm{i}(\boldsymbol{h}\cdot\boldsymbol{r})], \tag{9.84}$$

where the Fourier coefficients are usually weighted by their figure of merit. The initial interpretation of the map is in general not easy, unless very good phases at high resolution are available. Therefore, the strategy generally adopted is to calculate first an electron density map at low resolution, say 5–6 Å: these maps allow to identify the contours of the molecules in the crystal cell, and to distinguish between solvent regions and protein. Eventually, some elements of secondary structure can be identified: α-helices will appear as cylindrical rods of diameter of about 4–6 Å. β-sheets are more difficult to distinguish, and in any case single β-strands are not visible.

When the position of the molecule has been located in the unit cell, a map at medium resolution, say 3.5–2.5 Å resolution, is calculated and an attempt to trace the polypeptide chain is made. Chain trace at this resolution is made easier, and is sometimes possible, if the amino acid sequence is known. Mistakes are quite common in the interpretation at medium resolution: the connections among secondary structure elements are often difficult to recognize and amino acids can be positioned along the chain shifted from their correct position by one or more residues.

Higher resolution phases, say 2 Å or more, allow us to correct for this kind of mistakes and to locate more accurately the amino acid side chains. Unfortunately, MIR phases very seldom extend to this resolution, and high-resolution maps can be obtained using calculated or combined phases, as will be discussed later, or starting from lower resolution phases and extending and improving phases using one (or more) of the methods described in the previous chapter. If high-resolution native diffraction data are available, maps can undergo automatic chain-tracing.[96] This procedure works, in combination with phase improvement, only if resolution is close to the atomic one, that is, 1.5 Å or less.

9.5.2 Interactive computer graphics and model building

Low-resolution maps were traditionally drawn on a small scale on transparent sheets and are known among protein crystallographers as 'minimaps': they are very useful in giving a global view of the electron density of the unit cell. Such maps can also be used at medium resolution, since they allow a preliminary, approximate tracing of the polypeptide chain.

The complete building of the molecular model in the old days was performed using an apparatus, generally home-made, called an optical comparator or 'Richard's box', from the name of its inventor.[124] Nowadays, the interpretation of the electron density can be entirely performed on a graphic display: the map is shown on a video and the operator is allowed to fit a piece of chain into the density. For this purpose, the modelling software most popular among protein crystallographers is O.[125,126] The principle of the program is that objects displayed on the screen are divided in two categories: those that can be manipulated, that is, the atomic models, and those that cannot be modified interactively, like the electron density map. Maps and models can be seen superimposed on each other. Program O contains a dictionary of stereochemical information on natural amino acids and groups often occurring in protein molecules or nucleic acids. From this dictionary, building of a part of a polypeptide chain or a portion of a macromolecule in the preferred conformation is straightforward.

The electron density, which does not need to be modified, is displayed as a background in a 'chicken wire' representation or similar, and the atomic model, or a part of it, as a foreground object. Figure 9.45 illustrates a portion of a model and the corresponding map, as they appear on a graphic workstation at different resolution. Atoms can be easily manipulated, by

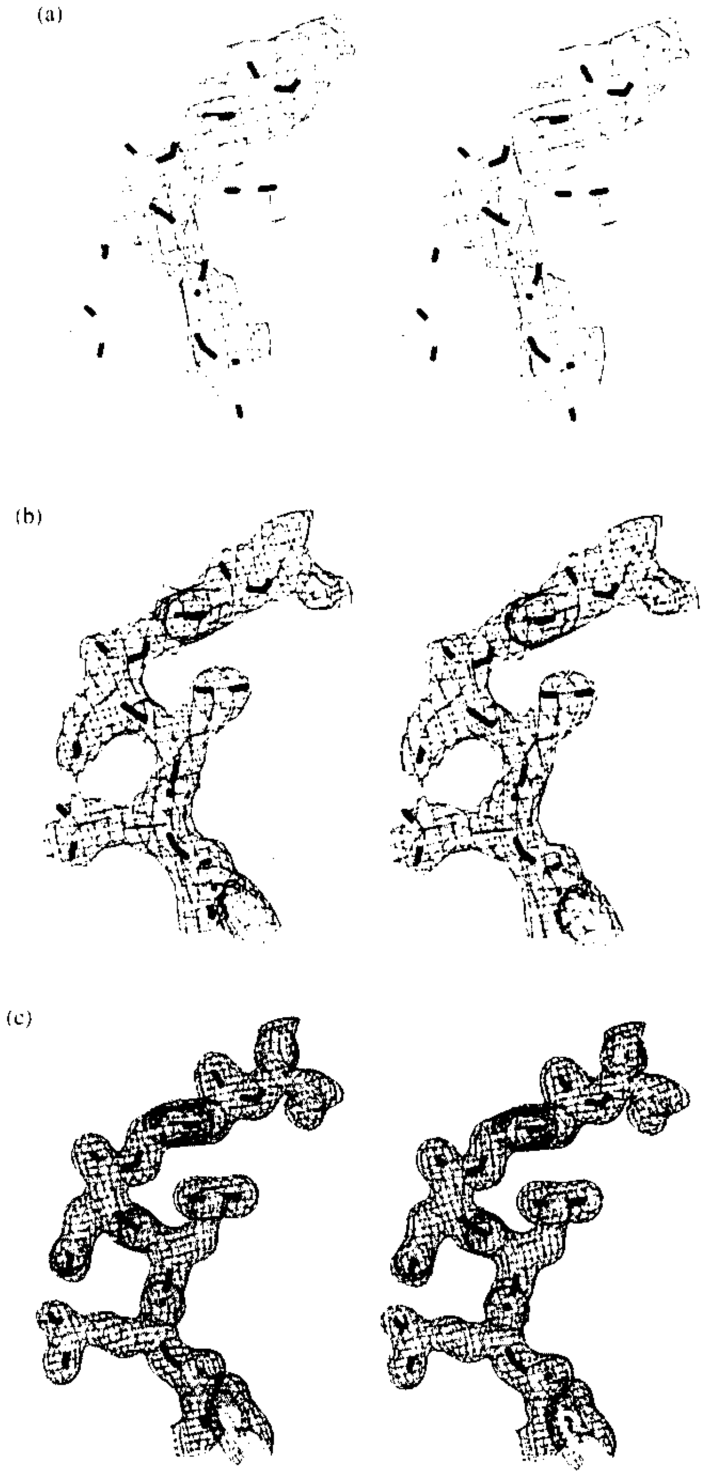

Fig. 9.45
'Chicken wire' representation of the electron density of a portion of a chain. Maps are calculated at a resolution of (a) 5 Å, (b) 3 Å, (c) 1.4 Å. In all cases coefficients $2F_{\mathrm{obs}} - F_{\mathrm{calc}}$ and phases calculated from the model were used. Drawing was produced using program O.[126] See colour plate section.

a simple rotation of a dial or the movement of the mouse: they can be moved, alone or in groups, rotation around dihedral angles performed or bonds stretched. The fitting of the built model into the electron density can be fast, if the starting phases are good enough; otherwise, a lot of time can be spent in trials. A convenient method to build the initial model of a protein has been devised,[127] based on the idea that short elements of secondary structure can be taken from a data base consisting of coordinates of a restricted set of well-refined protein structures: once some α-carbons (say 10–15) are roughly positioned into the density, the piece of model from the data base that better fits these atoms is searched for and used.

The only disadvantage in the use of a graphic system during electron density map interpretation is that a global view of the molecular model is practically impossible, since a drawing of all the atoms of a protein gives quite confusing results. Nevertheless, interactive graphic systems are of invaluable help at this stage of the structure determination.

9.5.3 The refinement of the structure

The refinement of protein structures, with few exceptions,[128] cannot be carried out using the classical least-squares methods. This is not due to the size of the problem, since nowadays computers are powerful enough to handle systems of equations containing thousands of variables, but to the limited number of X-ray data. It has been shown in fact at Section 2.11 that, for an accurate definition of the parameters, the system must be largely overdetermined, that is the ratio of observations to variables (atomic coordinates, thermal factors and sometimes occupancy) must be of the order of ten or so, and this is indeed the case for small molecules, where diffraction data can be collected to a spacing of 0.7 Å or even less. Protein crystals are intrinsically less-ordered: diffraction data are often measured to a resolution of 3.0–2.5 Å, sometimes 2.0 Å. A resolution of 1.5 Å can be considered quite good, and only in a limited number of cases 1.0 Å data have been collected. A typical situation is illustrated in Table 9.5, where the number of independent reflections for a medium-size protein (182 amino

Table 9.5 Number of theoretical independent reflections at different resolutions for a protein crystal with one molecule of 182 amino acids in the asymmetric unit. The solvent content is about 40 per cent. The number of parameters is 4408 (1469 atoms times 3) if an overall B factor is considered, 5876 if an individual isotropic B thermal parameter is assigned to each atom. They become 13 221 for anisotropic B factors

Resolution range (Å)	Independent reflections	Ratio obs./var. (x, y, z)	Ratio obs./var. (x, y, z, B)
40–3.0	3500	0.8	—
40–2.5	6800	1.6	1.2
40–1.9	13 500	3.1	2.3
40–1.5	29 800	6.8	5.1
40–1.2	58 800	13.3	10.0
40–1.0	81 300	18.5	13.8

acids) with a solvent content of 40 per cent are calculated at different ranges of resolution. For a ratio of observations to parameters of 10 it would be necessary to collect all possible diffraction data to a spacing of 1.2 Å, a resolution difficult to attain for such kind of a crystal. Historically, the first attempts to refine a protein structure were performed in real space.[129] The method seeks to minimize the difference between the observed electron density, ρ_{obs}, computed by eqn (9.84), and a calculated model density, ρ_{calc}, obtained by assuming a Gaussian distribution of the electron densities centred at the atomic positions of the current model:

$$S = (\rho_{\text{obs}} - \rho_{\text{calc}})^2. \tag{9.85}$$

This technique suffers from all the drawbacks of a real-space refinement procedure, the most relevant being that if poor phases are available, ρ_{obs} will result quite incorrect and the convergence of the method very slow, or it will not converge at all. There are many reasons to favour the reciprocal space refinement methods and different solutions have been proposed to overcome the problem of the underdetermination of the system: in fact, improvement of the ratio of observations to parameters can be achieved by decreasing the number of variables or by artificially increasing the number of observations. The former is called **constrained** least squares and the latter **restrained** least squares.

9.5.4 Constrained versus restrained least-squares

Constrained or rigid-body refinement[a] is a well-known and widely used technique in crystallography (see Section 2.11.6): when the geometry of a group of atoms is accurately known and there are reasons to believe that it will not be significantly modified by the environment, the entire group can be treated as a rigid entity. In the classical case of a phenyl ring, the eighteen positional variables can be reduced to only three translational and three rotational.

Bond length and valence angles in amino acids are very well known from the structures of hundreds of small peptides. In a protein, they can be held fixed to their theoretical values and only torsion angles around single bonds allowed to vary. This approach was used by Diamond[129] in real-space refinement, but it can be used in reciprocal space as well. Taking into account the fact that the peptide bond can be considered planar, only two torsion angles, called φ and ψ (see Section 9.2.4), need to be varied for the backbone chain of every amino acid: for a protein of n residues, the parameters are reduced to about $2n$ for backbone plus the torsion angles of side chains. An illustration of a possible choice of constrained parameters is reported in Fig. 9.46 for a simple dipeptide. This reduces the problem of underdetermination, but the model becomes in some way too rigid and the radius of convergence, that is the maximum displacement allowed for an atom in a wrong position to be corrected, becomes quite small.

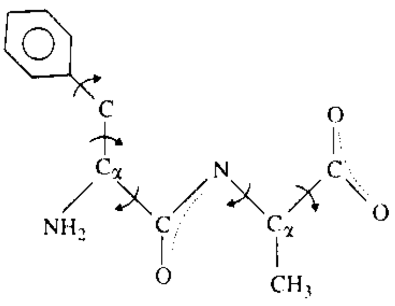

Fig. 9.46
Schematic drawing of the dipeptide phenylalanine-alanine, used to illustrate the **constrained** least-squares technique. Arrows indicate free rotation about the bond. All the bond lengths and valence angles are held fixed, and the peptide group and the phenyl ring planar. The total number of variables amount to eleven: three rotations and three translations (not indicated in figure) plus five internal torsion angles. (Some of the hydrogen atoms are indicated in the figure only for clarity, but they are usually not taken into account in the refinement.)

[a] Despite the distinction between them described in Section 2.11.6, **rigid-body** and **constrained** refinement are taken as synonyms in this chapter.

Constrained least-squares can be applied to a very different extent: the definition of rigid-body can be applied to only some group of atoms or to the entire molecule. If, for example, an approximate solution of the structure has been found using the molecular replacement technique, the first refinement can be performed by considering the entire protein (or a subunit) as a rigid group and the best position in the new crystal cell can be searched for using only three translational and three rotational variables. In that event, there is the supplementary advantage that, since the number of variables is very limited, only low-resolution data need to be included in refinement, greatly increasing the radius of convergence of the method.

Increasing the number of observations is another possible solution of the underdetermination problem in macromolecular refinement (see Section 2.11.6). Information from other sources can, in fact, be introduced and treated in a way similar to that used for observations coming from X-ray diffraction. The use of geometrical restraints has been proposed by Konnert and Hendrickson,[130,131] following a procedure devised by Waser[132] for small molecules. In addition to the classical quantity minimized in crystallographic least-squares:

$$S_1 = \sum_i w_i (F_{i(\text{obs})} - F_{i(\text{calc})})^2, \qquad (9.86)$$

where the summation is extended to all the i reflections, other observational functions can be added. Since distances and valence angles of amino acids are well known and they are not expected to deviate significantly from the ideal value, instead of considering them as fixed, we can also minimize them:

$$S_2 = \sum_j w_j (d_{j(\text{ideal})} - d_{j(\text{calc})})^2. \qquad (9.87)$$

$d_{j(\text{ideal})}$ is the ideal value for the specific distance we are considering, $d_{j(\text{calc})}$ is that calculated from our present model, and w_j is usually chosen as the reciprocal of the standard deviation of the distribution expected for the distances of type j. Notice that since $d_{j(\text{calc})}$ is a function of the atomic coordinates, eqn (9.87) does not increase the number of variables. The total number of equations like (9.86) is equal to the distances that are restrained: bond lengths, the distances between one atom and the next-nearest-neighbour (which is equivalent to restraining valence angles) and the first-to-fourth atom distances, where the dihedral angle described by the four atoms is in some way fixed (this is for example, the case of the planar peptide bonds). An example of the number of distances that can be restrained for a simple dipeptide is illustrated in Fig. 9.47. Other possible restraints in the Hendrickson and Konnert formulation are:

$$S_3 = \sum_k \sum_i w_m (\boldsymbol{m}_k \boldsymbol{r}_{i,k} - d_k)^2, \qquad (9.88)$$

$$S_4 = \sum_l w_l (V_{l(\text{ideal})} - V_{l(\text{calc})})^2, \qquad (9.89)$$

$$S_5 = \sum_n w_n (d_{n(\text{ideal})} - d_{n(\text{calc})})^4. \qquad (9.90)$$

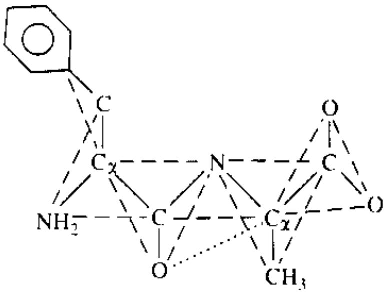

Fig. 9.47
The same dipeptide of Fig. 9.46 illustrates the **restrained** least-squares technique. The coordinates of any atom are allowed to vary, but the stereochemistry is preserved by applying restraints on bond distances (full lines), bond angles (dashed lines), torsion angles (dotted lines) and planarity. Non-bonded contacts are not shown in the figure. (Adapted from Sussman, Ref. [3], Vol. 115, p. 274.)

S_3 represents the sum of the deviations of the atoms i from the plane k, which is defined by its unit normal $\boldsymbol{m}_k$ and by the origin to plane distance d_k; $\boldsymbol{r}_i$ is the vector that defines a point i whose distance from the plane k has to be minimized.[133] S_4 restrains the volume of chiral atoms, defined for an α-carbon by the product of the interatomic vectors of the three atoms bound to it:

$$V_{C\alpha} = (\boldsymbol{r}_{\mathrm{N}} - \boldsymbol{r}_{C\alpha})[(\boldsymbol{r}_{\mathrm{C}} - \boldsymbol{r}_{C\alpha})\boldsymbol{x}(\boldsymbol{r}_{C\beta} - \boldsymbol{r}_{C\alpha})]. \tag{9.91}$$

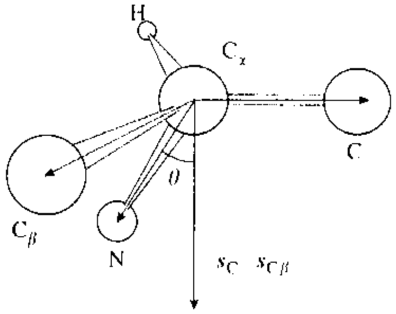

Fig. 9.48
The chiral volume for an α-carbon atom. The central atom is chosen as the origin of the coordinate system, and vectors $(\boldsymbol{r}_C - \boldsymbol{r}_{C\alpha})$, $(\boldsymbol{r}_{C\beta} - \boldsymbol{r}_{C\alpha})$, and $(\boldsymbol{r}_{\mathrm{N}} - \boldsymbol{r}_{C\alpha})$ (9.91) are denoted s_C, $s_{C\beta}$, and s_{N} respectively. The cross-product $s_C \times s_{C\beta}$ is a vector perpendicular to the plane $CC_\alpha C_\beta$. If it is on the same side of vector s_{N}, as in figure, that is if angle θ is less than 90°, the dot product between the two vectors is positive. If s_C and $s_{C\beta}$ are reversed, that is if the wrong configuration is chosen for the α-carbon, the vector $s_C \times s_{C\beta}$ points in the opposite direction and the value of $V_{C\alpha}$ becomes negative.

Since the sign of $V_{C\alpha}$ depends upon the handedness, S_4 restraints the chiral centres to their correct configuration (Fig. 9.48). S_5 is applied to all non-bonded atoms (except those taken into account in S_2) and avoids too close contacts. Other kinds of restraints can be considered, that is, on isotropic thermal parameters, occupancy and non-crystallographic symmetry. It may sometimes happen, particularly during the first stages of refinement, that some parts of the structure are poorly determined and the model 'blows up'. In that event, a restrain on the excessive shifts can be applied:

$$S_6 = \sum_k w_k(\boldsymbol{r}_k - \boldsymbol{r}_0)^2, \tag{9.92}$$

where $\boldsymbol{r}_k$ and $\boldsymbol{r}_0$ are the atomic vectors of the target and the initial model, respectively. Using eqns from (9.87) to (9.92), the number of observational functions is now greatly increased from the original number, represented by eqn (9.86). Equation (9.92) has effect only on the diagonal terms of the normal matrix. The number of restrained parameters for the example described in Table 9.5 is shown in Table 9.6.

Table 9.6 Number of restraints, following Hendrickson and Konnert,[131] for a protein molecule of 1469 atoms (excluding H) in the asymmetric unit. The example is relative to the case of Table 9.5

Number of distances	
Bond distances	1460
Angle distances	1974
Planar 1 4 distances	523
Planes	256
Chiral centres	201
Torsion angles	892
Possible contacts	
Contacts due to single torsion	496
Contacts due to multiple torsion	967
Possible H-bonds	102

It must be remembered that in protein crystallography 'experimental' phases are very often available. They can be included in least-squares as an additional information that imposes another restraint:[134,135]

$$S_7 = \sum_i w_p(\varphi_{i(\mathrm{obs})} - \varphi_{i(\mathrm{calc})})^2. \tag{9.93}$$

φ_{obs} is the estimate of the phase angle from isomorphous and anomalous data and φ_{calc} is the phase calculated from the model. Weights for eqn (9.93) must take into account the cyclic nature of phase angles.

Phase information is also used by Lunin and Urzhumtsev.[136] They suggest that only differences among crystallographic quantities be minimized, that is structure factor amplitudes and phases. Since phase probability distribution may be represented by eqn (9.38), they assume an analogous probabilistic distribution for the module of the structure factor F for reflection i of the form:

$$P(F_i) \approx \exp[-(F_i^2 - F_{i(\mathrm{obs})}^2)^2/2\sigma^2], \tag{9.94}$$

and if structure factors moduli and phases are assumed to be mutually independent, the joint probability distribution will be given by the product of eqns (9.38) and (9.94). The most probable model will consequently be that which minimizes:

$$S = \sum_i \{(1/2\sigma^2)[F_i^2 - F_{i(\mathrm{obs})}^2]^2 - [A\cos\varphi_i + B\sin\varphi_i + C\cos 2\varphi_i + D\sin 2\varphi_i]\} \tag{9.95}$$

Using (9.95) the multimodality of the phase distribution is taken into account.

A different approach to using restraints has been proposed by Jack and Levitt:[137] instead of restraining stereochemistry, they minimize:

$$S = E + D, \tag{9.96}$$

where D represents the difference among observed and calculated structure factor amplitudes given by eqn (9.86) and E is a potential-energy function:[138]

$$\begin{aligned} E = {} & \sum \mathrm{k_b}(\mathrm{b}_{j(\mathrm{calc})} - \mathrm{b}_j^0)^2 + \sum \mathrm{k}_\tau(\tau_{j(\mathrm{calc})} - \tau_j^0)^2 \\ & + \sum \mathrm{k}_\theta\{1 + \cos(m\theta_\mathrm{k} + \delta)\} + \sum(\mathrm{A}r^{-12} + \mathrm{B}r^{-6}). \end{aligned} \tag{9.97}$$

The four terms on the right side describe bond, valence angle, dihedral torsion angle, and nonbonded interactions, respectively. $\mathrm{k_b}$ is the bond stretching constant and k_τ the bond angle bending force constant; k_θ is the torsional barrier and m and δ the periodicity and the phase of the barrier. A and B are the repulsive and the long-range nonbonded parameters. The summation extends to the j bonds, the l valence angles, the θ torsion angles and the n nonbonded interactions between all pairs of atoms separated by at least three bonds. Despite the apparently very different approach, the energy minimization and the geometrically restrained least-squares are not too different in practice, since the final effect of eqn (9.97) is to impose restraints on the model.

Whatever method is used, special care is needed about the weights applied to the different functions. We are in fact dealing with non-homogeneous quantities, like structure factor amplitudes and interatomic distances, so the weights of the relative observational functions must be chosen in such a way that everything is put on the same scale: an overestimate of geometric restraints will in fact produce a stereochemically perfect model associated with a very high crystallographic R factor; on the contrary, an underestimate of the same weight will result in a good R factor with unreasonable bond lengths and angles.

9.5.5 *Restrained* and *constrained* least-squares

The two methods described above, **restrained** and **constrained** least-squares, can be combined:[139] the molecule(s) is(are) considered as made up of rigid groups, and restraints are applied to distances among such groups. The quantity minimized, S, is the sum of three terms:

$$S = w_\mathrm{F}\mathrm{DF} + w_\mathrm{D}\mathrm{DD} + w_\mathrm{T}\mathrm{DT}, \tag{9.98}$$

where DF is eqn (9.86), DD restrains the stereochemistry, analogously to eqns (9.87)–(9.91), and DT restrains the structure from moving away from the starting set of coordinates (9.92). All of the terms of (9.98) are functions of the atomic coordinates, generally referred to an orthogonal reference system. If a subset of these atoms is considered as a rigid group, S can be expressed, for that particular group of atoms, as a function of six rigid-body parameters, three rotational and three translational, and an arbitrary number of torsion angles, that is,

$$S = S(\boldsymbol{t}_i, \boldsymbol{R}_i, \Psi_{il}, \ldots, \Psi_{im}, B_{il}, \ldots, B_{in}), \tag{9.99}$$

where $\boldsymbol{t}_i$ and $\boldsymbol{R}_i$ are the translation vector and the rotation matrix of the entire group i, $\Psi_{il}, \ldots, \Psi_{im}$ are the m torsion angles and $B_{il}, \ldots, B_{in}$ the n temperature factors of group i.

Since the definition of rigid group is left to the user, the entire molecule or a portion of it can be constrained, or eventually some subunits. The restrained constrained approach was originally devised for nucleic acid refinement, but it has been successfully used in refinement of protein structures too.[140,141] A computationally quite efficient method of combining sterical restraints and rigid-body refinement has been also recently proposed.[142]

9.5.6 Crystallographic refinement by molecular dynamics

The development of vectorial and parallel computers offers nowadays the possibility of performing molecular dynamics calculations on complex systems, including proteins in the crystal state. The application of molecular dynamics calculations to macromolecules is a quite widespread technique,[144] but its introduction in the crystallographic refinement of protein structures has been proposed only recently.[145,146]

Molecular dynamics of free atoms consists in solving the classical Newton equation of motion:

$$m_i\,\mathrm{d}^2x_i(t)/\mathrm{d}t^2 = -\mathrm{grad}_x E_{\mathrm{tot}}. \tag{9.100}$$

To take into account the effect of the medium and the approximations used to calculate the total energy, dynamical effects can be better represented by a set of Langevin equations:

$$m_i\,\mathrm{d}^2x_i(t)/\mathrm{d}t^2 = -\mathrm{grad}_x E_{\mathrm{tot}} + f_i(t) - m_i b_i\,\mathrm{d}x_i(t)/\mathrm{d}t, \tag{9.101}$$

where b_i is a frictional coefficient, used to prevent atoms from moving away too much from their original positions, k_{B} is the Boltzmann's constant, T_0 the temperature, and $f_i(t)$ a random force with Gaussian distribution and properties:

$$\begin{aligned} \langle f_i(t)\rangle &= 0, \\ \langle f_i(t) f_i(0)\rangle &= 2k_{\mathrm{B}}T_0 b_i m_i \delta(t). \end{aligned} \tag{9.102}$$

The simulation starts from an initial set of coordinates. To each atom is assigned a velocity, usually at random from a Maxwellian distribution corresponding to the temperature selected, and eqn (9.100) or (9.101) is integrated at a given temperature for a given time.[v] New velocities are then assigned, eventually at a new temperature, and the calculation continued. The simulation is normally performed for a short period of time, usually of the order of few picoseconds.

In the crystallographic refinement of macromolecules by molecular dynamics, the X-ray information is used to restrain the energy of the system. The total potential energy is, in fact, considered as the sum of two terms:[147]

$$E_{\text{tot}} = E_{\text{emp}} + E_{\text{eff}}. \tag{9.103}$$

E_{emp} represents an empirical potential energy, analogous to that defined by eqn (9.97), E_{eff} is a sort of 'experimental' potential energy, and is considered as the sum of three terms:

$$E_{\text{eff}} = E_{\text{xray}} + E_{\text{P}} + E_{\text{NB}}. \tag{9.104}$$

E_{xray} in eqn (9.104) describes the difference between observed and calculated structure factor amplitudes:

$$E_{\text{xray}} = (w_{\text{A}}/N_{\text{A}}) \sum_h w_h [F_{(\text{obs})} - F_{(\text{calc})}]^2. \tag{9.105}$$

w_{A} is a factor which puts E_{xray} on the same scale as the empirical potential energy term and N_{A} is given by $\sum w_h \, (F_{\text{obs}})^{[2]}$, to ensure that w_{A} is independent of the resolution range used. The terms E_{P} and E_{NB} can be included to take into account experimental information about MIR phases and crystal packing, respectively.

Molecular dynamics simulation can be performed at ambient temperature,[145] or at higher temperature, as in the version called **simulated annealing**.[144,146] The latter consists in starting the simulation at room temperature, say 300 K, and heating up the system (e.g. at 2000–5000 K) and subsequently cooling down to the initial value. The advantage of going to high temperatures,[w] unreasonable from the biological point of view, is that the model can come out of the local minimum, and the radius of convergence of the method is increased with respect to classical least-squares.

The result of molecular dynamics calculations is a family of conformations, but the constraints imposed by X-ray data restrict these conformations to all those with the lower crystallographic R factor.

9.5.7 The strategy of the refinement of protein structures

The initial model of a protein structure is, very often, not good enough to allow for a fully automated refinement. Indeed, if some serious errors are present in the model, for example, the polypeptide chain is positioned more

[v] Numerical integration can be performed using for example, the Verlet algorithm.

[w] The term temperature must be regarded cautiously here: it does not indicate a physical temperature, but rather a parameter controlling the refinement. The simulated annealing is in fact virtually equivalent to the Metropolis algorithm (Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, A., and Teller, E. (1953). *Journal of Chemical Physics*, **21**, 1087–92).

or less correctly, but the amino acids are shifted one residue or more along the chain, an automatic procedure will hardly recover from that error. Besides, the radius of convergence of constrained or restrained least-squares methods can be evaluated to be around a half the resolution of the data used, that is not more than 1–1.5 Å. At the beginning of the refinement, to speed up convergence, medium-resolution data (3.0–2.5 Å) can be employed. Since the number of observations at that resolution is quite low, an overall temperature factor for all the atoms is used. Afterwards, the resolution is gradually extended, solvent molecules included, and isotropic individual B factors applied.

The same seems not to be true using the simulated annealing technique, which allows a more rapid and automatic convergence: the heating makes it easier to get out of a false minimum without manual intervention. Some more experience nevertheless must be accumulated. A fully automated refinement was possible in the test case of the enzyme aspartate aminotransferase, refined with data at 2.8 Å resolution starting from MIR coordinates.[147] A careful comparison between the model of myohemerytrin refined, starting from the same model, in one case with several cycles of restrained least-squares and manual rebuilding and in the other case with the 'simulated annealing' technique without manual intervention has been reported.[148] The two structures compare quite well, but molecular dynamics procedure could not bring the refinement to completion in a fully automated way: manual intervention is still necessary to correct for gross errors (say more than 3–5 Å in the main chain) and to include solvent molecules. Nevertheless, simulated annealing can save a lot of human effort, at the expenses of a quite long computational time.

For the above mentioned reasons, some cycles of automatic minimization are usually followed by recalculation of the electron density maps and manual adjustment or rebuilding of the model.

In recalculating electron density maps, a major problem is the choice of the phases and the coefficients to be used. MIR phases suffer from all the errors described in Section 9.4.8, and the isomorphism of heavy-atom derivatives does not extend generally beyond 3 Å or so: high-resolution electron density maps are seldom achieved with MIR phases. On the other hand, calculated phases tend to reproduce the model used in calculating them, and an electron density map obtained with calculated phases may be strongly biased. For these reasons, phases coming from independent sources, for example, the phases from isomorphous derivatives data and those calculated from the model, can be combined to produce an improved electron density map.[149,150] The probability distribution of calculated phases, $P_{\mathrm{calc}}(\varphi)$, can be evaluated by using a procedure due to Sim[59] (see Appendix 6.G) and can be used, along with the 'experimental' probability $P_{\mathrm{tot}}(\varphi)$, to obtain a combined probability distribution:

$$P_{\mathrm{comb}}(\varphi) = P_{\mathrm{tot}}(\varphi) P_{\mathrm{calc}}(\varphi). \qquad (9.106)$$

The new figure of merit, m_{comb}, obtained by (9.42) can be used to calculate a best combined electron density map. If only calculated phases are available, the Sim formula can be used to weight the Fourier coefficients. A scheme of the possible refinement procedure is illustrated in Fig. 9.49.

The coefficients and phases more commonly used for Fourier syntheses are listed below, but other combinations of them are possible:

$$mF_{\text{obs}}\exp(i\varphi_{\text{MIR}}), \tag{9.107}$$

$$w_{\text{comb}}F_{\text{obs}}\exp(i\varphi_{\text{comb}}), \tag{9.108}$$

$$w_{\text{sim}}|F_{\text{obs}} - F_{\text{calc}}|\exp(i\varphi_{\text{calc}}), \tag{9.109}$$

$$w_{\text{comb}}(2F_{\text{obs}} - F_{\text{calc}})\exp(i\varphi_{\text{comb}}), \tag{9.110}$$

$$w_{\text{sim}}(2F_{\text{obs}} - F_{\text{calc}})\exp(i\varphi_{\text{calc}}). \tag{9.111}$$

Equation (9.107) gives the coefficients of a classical observed Fourier synthesis. In principle, they could be used during all stages of the refinement, but MIR phases can be improved by phase combination. Furthermore, very often they do not extend to high resolution, and calculated phases must be used instead, when a reasonable atomic model is available. To reduce bias, a Fourier map with combined phases can be calculated. Coefficients (9.109) correspond to Fourier-difference maps with calculated phases. If they are calculated from a partial model, they are known as **omit maps** and can be useful in positioning portions of the molecule that did not appear clearly in the MIR maps. Coefficients (9.110) and (9.111) correspond to a combination of a Fourier electron density map and a difference-Fourier: if F_{obs} and F_{calc} are very similar, the magnitude of the coefficients approaches to (9.108); otherwise, terms with F_{obs} greater than F_{calc} will have a higher weight. The practical result is an enhancement of the regions of density where severe errors are present in the model. The weight used for map calculation can be a modification of the Sim scheme or the weight obtained by the combination procedure.

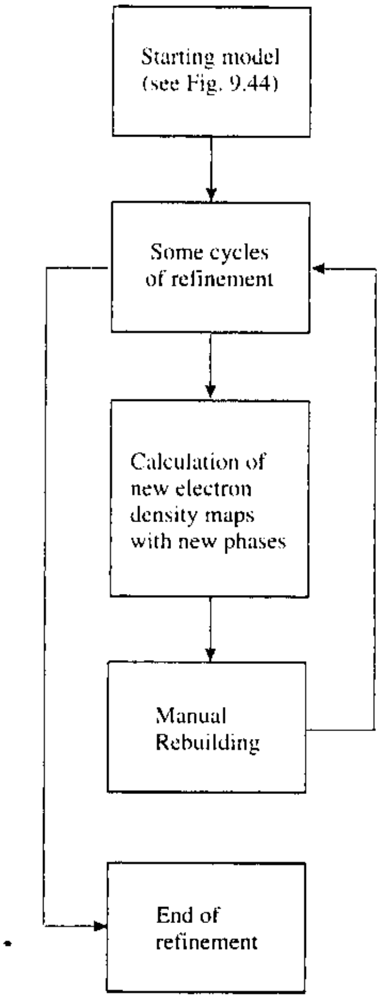

Fig. 9.49
Block diagram summarizing the phases of structural refinement. The starting model is obtained by one of the procedures schematized in Fig. 9.44. The numbers of iterating cycles necessary to reach convergence can be quite variable, depending from the quality of the initial model. If MIR phases are not available (i.e. the structure has been solved by molecular replacement techniques), only maps with calculated phases can be used. Some of the coefficients used in electron density map calculation are described by eqn from 9.107 to 9.111.

9.5.8 R factor and R_{free}: structure validation

The value of the crystallographic R factor is defined as:

$$R = \frac{\sum_h |F_{\text{obs}}(\boldsymbol{h}) - KF_{\text{calc}}(\boldsymbol{h})|}{\sum_h |F_{\text{obs}}(\boldsymbol{h})|}, \tag{9.112}$$

where F_{obs} and F_{calc} are the observed and calculated structure factor moduli, respectively, for reflection $\boldsymbol{h}$, K a scale factor and the sum extends over all the observed reflections. The numerator of eqn (9.112) is related to the negative logarithm of likelihood of the model, assuming all the observations are independent. In small molecule crystallography, eqn (9.112) generally represents a reliable indication of the correctness of the structure. In macromolecular crystallography, it must be taken with caution: owing to the limited (and sometimes very limited) ratio between observation and parameters, a low R factor alone is not sufficient. It has been shown[151] that in some limited situations wrong structures can refine to relatively low R factors. More commonly, a correct structure is often subjected to over-refinement, that is, the R factor decreases without any improvement (or even a worsening) of the accuracy of the coordinates of the molecular model. A classical situation is represented by solvent molecules positioned in peaks of the electron density of the map that do not correspond to real maxima, but simply to errors of the phases (or to series-termination errors). Despite that, the increase of the number of variables usually produces a (modest) decrease in the R factor.

An indicator of the quality of the structure virtually independent from refinement is the so-called R_{free}, defined as:[152]

$$R_{\text{free}} = \frac{\sum_{\boldsymbol{h} \in T} |F_{\text{obs}}(\boldsymbol{h}) - KF_{\text{calc}}(\boldsymbol{h})|}{\sum_{\boldsymbol{h} \in T} |F_{\text{obs}}(\boldsymbol{h})|}, \tag{9.113}$$

where T is a subset of reflections not used in refinement. In practice, a limited portion of data, randomly chosen and consisting usually of about 5–10 per cent, is left out from the refinement process and used only to calculate the R value. R_{free} is generally higher than R, but it reflects more closely than the classical R factor the real information content of the molecular model. The selection of a reference set, which should be made at the beginning of the refinement process to avoid any bias, slightly decreases the number of data available for refinement, but this is a little cost in comparison to the great advantage of having a reliable indicator that can be used to monitor the refinement process: the addition, for example, of solvent molecules or side chains in alternate conformations or the use of anisotropic thermal parameters always has the effect of reducing the R factor, but if they do not reduce the R_{free} this is a clear indication of a mathematical artifice, without correspondence to the physical reality.

It is difficult to give an exact indication of what value the R_{free} should assume in order to be able to say that the refinement has converged, since it strongly depends on the resolution of data. At 2 Å resolution the R factor of a refined structure is usually less than 0.20 (sometimes well below that value) and a good R_{free} 20–30 per cent higher. For a low resolution structure, for example, at 3 Å, the conventional R factor can drop to similar values, but R_{free} seldom reaches values lower than 0.30. This fact correlates with the overall quality of a low resolution model and it is an empirical indication of the validity of R_{free}.

Other indicators must anyhow be used to assess the quality of the structure. Most of them are geometric, that is, they are based on the comparison with the large amount of structural data available on protein and peptides. For example, deviations from ideality of bond lengths and valence angles of about 0.01–0.02 Å and 2°–3°, respectively, are considered reasonable. Automatic programs can be used that give not only an idea of the quality of the structure, but they also give direct indications of the portions of the model that eventually need corrections or that have to be checked more carefully.[153]

One of the most classical indicators is represented by the Ramachandran plot (Section 9.2.4): amino acids falling into forbidden regions must generally be corrected, and values at the border or in unfavourable zones considered with suspicion. An example of such a plot, along with other indicators, for a normally refined structure is given in Figs 9.50 and 9.51.

9.5.9 Thermal parameters and disordered structures

Special care must be used in evaluating temperature factors obtained from the refinement of a protein crystal structure. Experimental B factors include, besides thermal vibrations, the static disorder, which in such kind of crystal is particularly relevant. Owing to the high content of solvent, superficial groups of the macromolecule have the effect of partially ordering the solvent, but at the same time, as a consequence of this contact, they result very mobile. It has been shown[154] that it is possible, for very well-refined structures, to distinguish the contribution to B factors of real thermal vibration from the static disorder. Moreover, during the refinement procedure thermal parameters are usually restrained, that is their variation is in some way smoothed down. In any case, the comparison of temperature factors for the same structure, Ribonuclease A, independently refined by two different groups using different data[155] shows a quantitative agreement in the trends, that is regions with an high B factor roughly correspond in both structures. As a general rule, in well-refined structures main chain atoms present lower thermal motion than side chains, or in any case less disorder.

When looking at crystallographic results, it must be kept in mind that a very high B factor for some residues could be either due to an intrinsic disorder of that part of the molecule, or an indication of a misinterpretation of the electron density. In some cases, small parts of the structure, often at

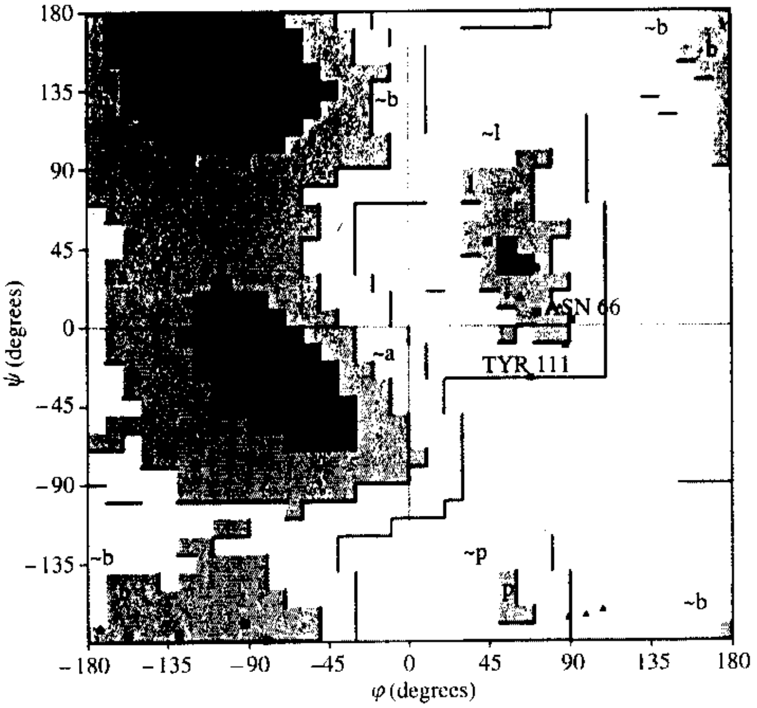

Fig. 9.50
Ramachandran plot for the model of pig retinol binding protein, orthorhombic form, $P2_12_12_1$ refined to a crystallographic R factor of 0.184 (R_{free} 0.237) up to a resolution of 1.65 Å.[31] Small triangles represent glycine residues, squares all the others. The origin of the diagram is in the centre. Red, yellow, and light yellow correspond to most favoured, additionally allowed and generously allowed regions. Excluding glycines, 91, 8, and 1 per cent of the residues fall in the first, second, and third zone, respectively. None in the forbidden areas. The two residues that present torsion angles in a relatively unfavoured region are Tyr 111 and Asn 66: the former presents a similar conformation in all the structures of the other proteins of the same family till now determined, and the second is located in a flexible and not well ordered loop. (Plot produced using the program PROCHECK.[153])

the beginning or at the end of the polypetide chain, or some loops protruding towards the solvent regions, are very mobile and cannot be seen at all in the map. At higher resolution, more than one conformation can be individuated for some residues.

The disorder can also have some functional role, as is sometimes the case for allosteric proteins, where two conformational states are present, one of them characterized by a portion of a disordered chain that becomes ordered in the other state.[156,157] In any case, this is a case in which the

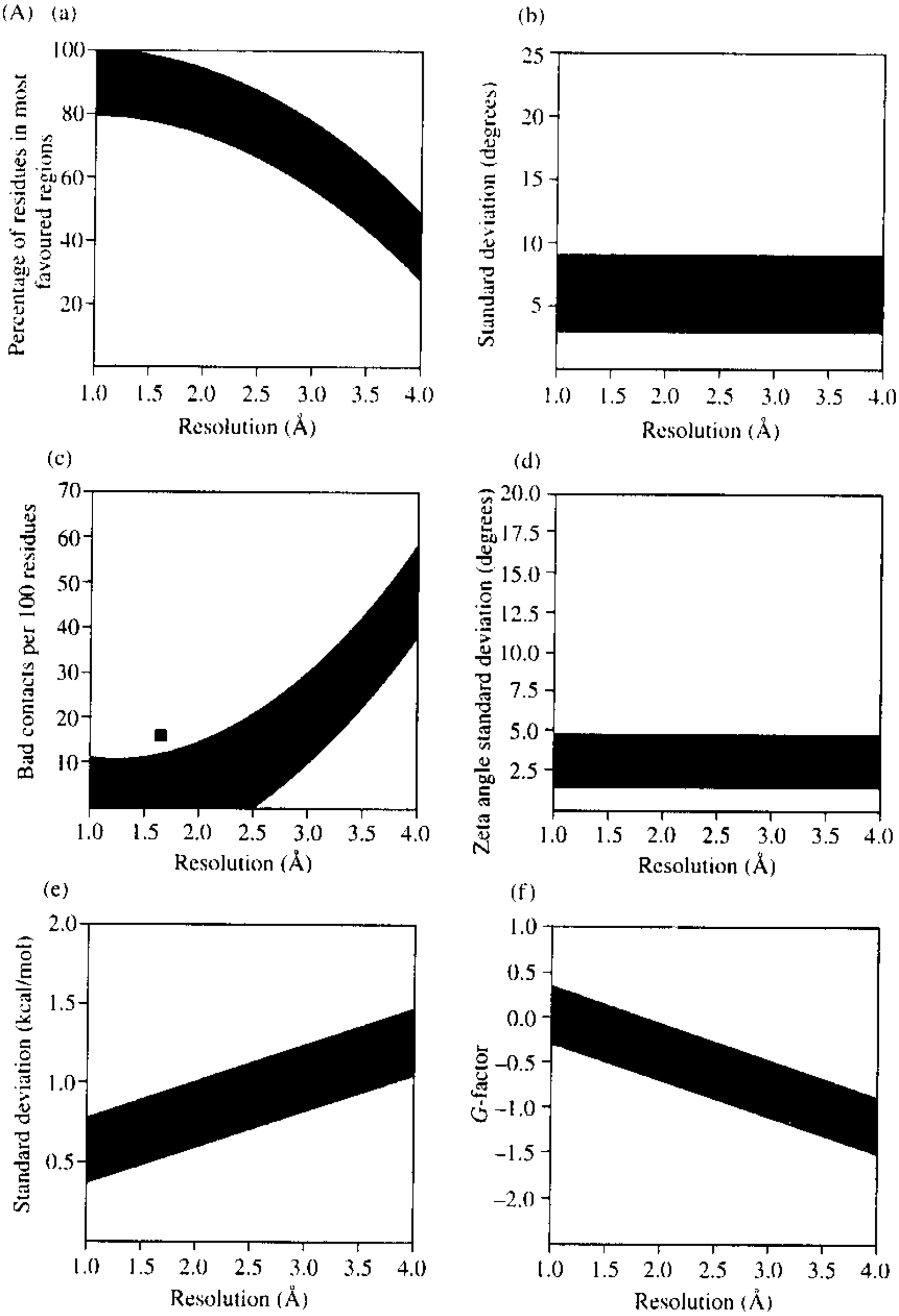

Fig. 9.51 (A)

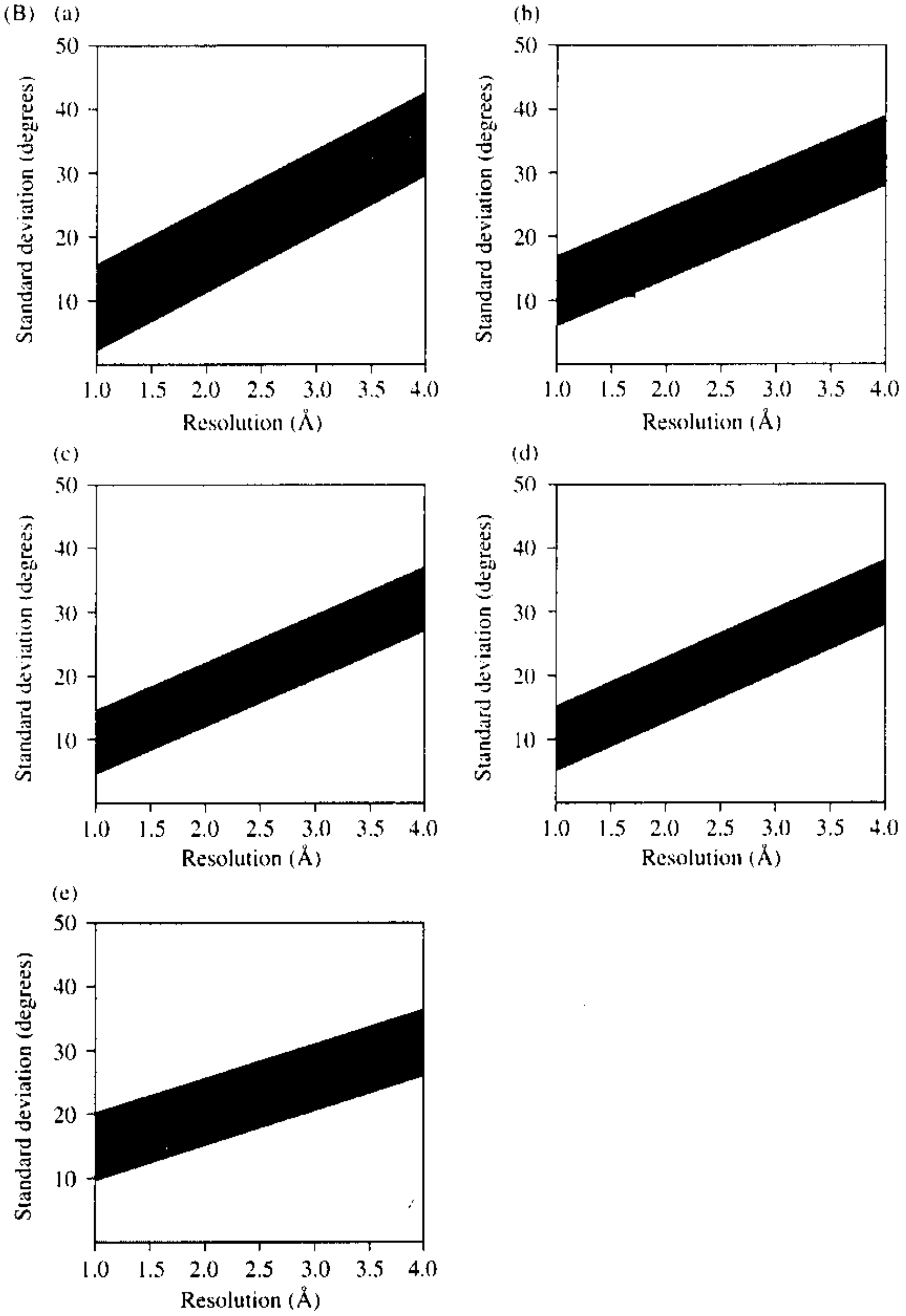

Fig. 9.51
Statistics for some relevant geometric parameters for the same protein model of Fig. 9.50, as detected from PROCHECK. Each graph represents a property of (a) main chain or, (b) side chains. The blue band indicates the results for well-refined structures: some properties are resolution-dependent, that is, a higher resolution implies a better definition of the parameters, others are not, like the planarity of peptide bond or the chirality of C_α atoms.
(A) From top to bottom and from left to right: (a) Ramachandran plot quality, that is, the percentage of residues falling in the most favoured regions, (b) planarity of peptide bonds, as measured by the standard deviation of the ω torsion angle, α-carbon tetrahedral distortion, (c) number of bad contacts per 100 residues, (d) distortion of C_α atoms from the correct chiral volume, (e) standard deviation of the hydrogen-bond energies for main-chain hydrogen bonds, (f) overall *G*-factor.
(B) Standard deviation of the side-chain torsion angles, $\chi 1$ and $\chi 2$, from some well-defined conformations: (a) Chi-1 gauche minus, (b) chi-1 trans, (c) chi-1 gauche plus, (d) chi-1 pooled standard deviation, (e) standard deviation of chi-2 *trans* angle.

crystallization process introduces a bias in the results, since less ordered or disordered proteins are likely to be more difficult or even impossible to crystallize.

9.5.10 The organization of solvent

A high portion of the solvent contained in a crystal cell can be considered not relevant for the macromolecule, but it is simply there to fill the channels produced by molecular contacts in the crystal. And indeed, this unordered solvent cannot be seen in an X-ray diffraction experiment. Water molecules closely bound to the protein, on the contrary, can be considered as part of the structure of the macromolecule itself: a protein cannot be completely dehydrated without a complete crash of the architecture of its three-dimensional structure. Tightly-bound solvent molecules in the crystal are identified during the process of refinement, and can be distinguished in three groups:

1. Water molecules making hydrogen bonds with hydrophilic side-chains on the surface of the protein, where they often take part in tetrahedral or trigonal network of hydrogen bonds. Ordered waters are substantially on the first shell of coordination around the protein, or eventually in the second shell, bound to the water molecules of the first one.
2. Water molecules that serve as a bridge among parts of the main chain or other structural elements that are too far apart to form hydrogen-bonds: for example, if two strands of a β-sheet diverge slightly, a water molecule can make a H-bond in the middle, filling the gap. This kind of solvent molecules is essential in stabilizing the protein structure.
3. Solvent molecules located in the internal cavities of the protein, where sometimes they do not form very stable interactions, but simply fill the vacuum.

It should be noted that the arrangement of the solvent structure around a protein determined by X-ray analysis is strongly influenced not only by the crystal packing, but also by the pH and the solvent used in the crystallization, and it cannot be considered fully representative of the situation *in vivo*.

9.5.11 The influence of crystal packing

The possibly most often asked question since the beginning of protein crystallography can be summarized as follows: how is the structure in the crystal representative of the 'real' *in vivo* structure? Proteins are quite stable, but highly flexible molecules: the same protein obtained in different crystal forms, and consequently subjected to completely different packing forces, presents in general the same fold, with some differences, usually small, in the regions of contact among molecules in the crystal. Figure 9.52 shows the α-carbon chain trace of the enzyme rhodanese from two crystal forms, monoclinic and orthorhombic.[158,159] The fact that molecules crystallized in different conditions of pH and precipitants keep the same overall conformation is an indirect evidence of the stability of protein conformation, and of the validity of the structure obtained using the crystallographic technique. On the other hand, the local variations suggest that great care has to be taken in drawing specific conclusions on functional aspects from details of the structure.