Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 28 May 2003
Accepted 29 July 2003

Automatic model building based on flexible fragment formalism. The case of high-resolution protein structures

Frantisek Pavelcik

Department of Inorganic Chemistry, Faculty of Natural Sciences, Comenius University in Bratislava, 84215 Bratislava, Slovak Republic, and The University of Texas, Southwestern Medical Center at Dallas, Department of Biochemistry, 5323 Harry Hines, Dallas, TX 75235, USA. Correspondence e-mail: pavelcik@fns.uniba.sk

A concept of flexible fragments has been developed for automatic building of crystal structures. Six monopeptides were designed as search fragments in a phased rotation and translation function for protein building. Electron density in crystal and in molecular fragments is expanded in spherical harmonics and normalized spherical Bessel functions. A fast rotation function, which is calculated at each grid point of the asymmetric unit, is used to find the fragment orientation. Position, orientation and internal torsion angles are refined. An algorithm for chain building is simplified using generalized atoms and virtual bonds. The structure is built from molecular structure units rather than from individual atoms. A polyalanine model is built with a high accuracy at resolutions 1.2–2.1 Å.

1. Introduction

Automatic structure building is an important step in the overall automation of crystal structure determination. Inspection of interatomic distances (connectivity table) is usually faster and simpler than coding molecular connectivity (sometimes only partially known) for small molecules. On the other hand, using graph theory can speed up atomic resolution protein structure interpretation considerably (Oldfield, 2002*b*). Computer graphics can help in assignment of atomic types and in removing a few false atoms (*e.g.* Pavelčík *et al.*, 1992). Beyond atomic resolution, detailed stereochemical knowledge has to be introduced in some way to improve the observation/parameter ratio. There are two broad areas of X-ray crystallography where using stereochemical information is vital: powder diffraction and biomacromolecular crystallography.

This paper is oriented towards a systematic use of molecular fragments, as building units, in the interpretation of protein electron-density maps. It is an open challenge to develop a method that can build structure models automatically with minimal user interference. In protein crystallography, the input stereochemical information can be as simple as the protein sequence. Other tasks can be done by an expert system.

Several procedures and computer programs have been developed for building protein structure models. Most of them are based on the connectivity of the polypeptide chain or on the presence of α-helices and β-strands. Greer (1974, 1985) devised a rapid procedure for tracing the path of the polypeptide chain. Greer's skeletonization is still the most common aid for tracing the main chain in a given electron-density map. This procedure was extended by Swanson (1994) to allow threshold-independent skeletonization and Leherte *et al.* (1994), which involves analysing critical points. Jones & Thirup (1986), Kleywegt & Jones (1997*a*,*b*) fitted the electron density with fragments from known protein structures (*ESSENS*). Oldfield (2002*a*, 2003) described a method for automated model building that began by identifying helices and strands and then extending these segments to trace a polypeptide chain (*QUANTA*). Cowtan (1998, 2001) used FFT-based approaches to identify the location of helices, β-strands and other structures in an electron-density map by template matching (*FFFEAR*). Holton *et al.* (2000) and Ioerger & Sacchettini (2002) used machine-learning techniques to identify protein backbone and side chains in a map (*TEXTAL*, *CAPRA*). McRee (1999) has described a semi-automated method for building main-chain models in a map, beginning with the identification of C^α and fitting fragments from a main-chain library and continuing with using a rotamer library to fit side chains (*XtalView*). Levitt (2001) used a stepwise approach to model building, beginning with the skeleton of Greer to identify helices and strands and extending them one residue at a time, using (φ, ψ) angles from tables of allowed values (*MAID*). Turk (2001) iterates model building with refinement (*MAIN*). This feature is typical for the most widely used automated model-building procedure, *ARP*/*wARP* (Lamzin & Wilson, 1993; Perrakis *et al.*, 1999; Morris *et al.*, 2002). This procedure is different from all those described above because it is based on an interpretation of the difference electron density in terms of individual atoms ('small-molecule approach'), iteratively followed by atomic refinement and an interpretation of the atomic coordinates in terms of a polypeptide chain. The requirement for peaks limits the application of the method to electron-density maps at a resolution of about 2.3 Å. Terwilliger (2001, 2003) published a version of the FFT-template method, with extension of regular secondary structure by tripeptide matching (*SOLVE*/*RESOLVE*). The method works at a resolution as low as 3.5 Å. Pavelcik *et al.* (2002, referred to as PZO below) described a phased rotation and translation function (*PROTF*) for matching of arbitrary fragments of the structure to an electron-density map. It was shown that relatively small rigid molecular fragments, with a proper conformation, could be fitted into the electron density with a good accuracy.

The biopolymer can be built from rigid fragments but the number of fragments has to be large to cover a substantial part of a conformation space. Terwilliger (2003) used for example about 10000 rigid tripeptides. In this paper, a small set of fragments with added conformational flexibility is proposed to reduce the number of fragments significantly. The fragment selection is a compromise between its size and number of variable internal torsion angles. Fragments with many torsion angles would lead to an exponential conformation catastrophe. A small rigid fragment may not be found in the electron density.

The dominant part of the main chain of the protein is composed of three basic types of secondary structures: α-helix, β-strand and γ-turn. These conformations cover more than 90% of all conformations in a typical protein structure. While an α-helix is relatively rigid, β-strands and γ-turns are flexible in a broad range of (φ, ψ) angles. The β-region seems to have two distinct broad local energetic minima, further designated as β_1 (sheet region) and β_2 (random coil region). A recent paper (Hovmöller *et al.*, 2002) was used as a guideline for selection of fragments.

Each fragment represents a building unit. Virtual bonds can describe the connectivity of building units. The virtual bond concept is well established in biomacromolecular chemistry. We selected rather small fragments of 9–11 atoms for the protein building. A fast method for connecting fragments into a polypeptide chain was developed. The protein structure is built from generalized atoms rather than from individual atoms and a significant reduction of dimensionality was achieved.

2. Methods

2.1. Flexible structure model (FSM)

The crystal structure can be described as composed of groups of atoms (in stereochemical terms these are molecular fragments). These can be for example a peptide group, a phenyl group, one turn of an α-helix, a phosphate group *etc.* Suitable selected molecular fragments can be used as a building unit. Ideally, the building unit is a rigid body. The number of group types can be relatively small in polymer structures. Bond lengths and bond angles are usually assumed to be known (and tabulated) in organic chemistry. Under these conditions, only torsion angles are variables. To keep building units sufficiently large, conformational flexibility should be added to the building units. Such a flexible group of atoms will be called a structure unit (SU) in this paper. The structure unit can be treated as a generalized atom. The SU is characterized by its name, position, orientation and internal torsion angles. The SU has its internal connectivity, numbering scheme of atoms, torsion bonds and torsion groups. The position of the SU is given by the position of its geometrical centre in fractional coordinates. Orientation of the SU is conveniently given by a rotation matrix, three Euler angles (α, β, γ) or quaternion. SUs can have further attributes: symmetry code, connectivity to other SUs, sequence number, chain letter *etc*. For each type of SU, a standard SU should be defined. Orientation (and sometimes also conformation) is given with respect to this reference standard. A standard SU is a set of atomic Cartesian coordinates, atomic names, atomic types, atomic numbers and residue types in a protein. The geometrical centre of the standard SU is at origin (0,0,0).

The whole structure can be regarded as composed of SUs connected by virtual bonds. This representation of the structure will be called a flexible structure model. A FSM has virtual geometry (virtual bond lengths, virtual bond angles, virtual torsion angles). In a description of the structure with structure units, some structure atoms are defined more than once because of group overlap. The advantage of using SUs is that the number of parameters for the description of the structure is significantly reduced, *e.g.* a SU containing 10 atoms and 2 torsion rotations (τ_1, τ_2) is described by 30 Cartesian coordinates but only eight generalized coordinates ($x, y, z, \alpha, \beta, \gamma, \tau_1, \tau_2$). A rigid part of a heme described by only six generalized coordinates represents 99 Cartesian coordinates.

2.2. Molecular fragments for high-resolution protein building

From a practical point of view, the choice of fragments is from monopeptides, dipeptides and tripeptides because the peptide group is a rigid group of five atoms. For automatic interpretation of high-resolution protein structures by the phased rotation and translation function, the selected fragments are of two types: peptide centred (P0) fragments

$$\mathrm{N-C^{\alpha}(C^{\beta})-C({=}O)-N-C^{\alpha}(C^{\beta})-C}$$

and C$^{\alpha}$-centred (A0) fragments

$$\mathrm{C^{\alpha}-C({=}O)-N-C^{\alpha}(C^{\beta})-C({=}O)-N-C^{\alpha}.}$$

The radius of these fragments is about 3.0 Å. The related radius of the electron-density expansion was fixed at 3.7 Å. This selection was not arbitrary. In the sphere of radius 3.7 Å, two peptide groups can be accommodated (with the exception of a fully extended conformation). These fragments are the smallest ones from which a polyalanine chain (C$^{\beta}$ present) can be built. Fragments contain an element of chirality, which is useful for unique fitting. This radius is also a compromise for eventual building of side chains (Pavelcik, in preparation). Fragments have only two torsion bonds (φ and ψ torsion angles) and cover the main regions of the Ramachandran plot. Fragment names and main-chain torsion angles are given in Table 1.

Fragments are of standard geometry and were built by molecular modelling. A geometrical centre of the fragment is at the origin of the coordinate system. Fragments are used in two ways: as rigid bodies in *PROTF* and as flexible fragments for refinement and model building.

A total of six rigid fragments are used in *PROTF*. AlphaP0, Beta1P0, Beta2P0, GammaP0 and BridgP0 are peptide-centred fragments. Fragments have the same connectivity (atom numbering) and bond geometry. The difference is only in conformation. AlphaP0 is related to the α-helix. Beta1P0 is a β-strand from the sheet region. Beta2P0 is a β-strand from the random coil (proline) region. The GammaP0 fragment is of helical conformation near α_L and γ-turn conformations.

The general *cis*-peptide bond was neglected, but the most frequent *X*-Pro *cis*-peptide bond is considered. The probability of occurrence of a non-Pro *cis*-peptide bond is about 1:3000 (Pal & Chakrabarti, 1999). A special fragment cisPro0 was modelled for this purpose. The fragment is the same as AlphaP0, but contains also C^γ and C^δ pyrolidine ring atoms (total 11 atoms). ω is 0°. The pyrrolidine ring is planar. C^γ and C^β atoms are not connected and ring puckering can be refined (see later).

AlphaA0 is a C^α-centred fragment of the helical conformation. It is not used primarily in the rotation function to build the polypeptide chain but is used in extending the chain, refining the structure and building the C^β position. An accurate C^β is needed later for building side chains.

When the polyalanine structure is built, fragments are connected into a larger chain. There is an overlap of four atoms for two P0 fragments and even larger overlap for connecting P0 and A0 fragments together. The P0 fragment represents almost two residues (one oxygen missing); the larger A0 fragment uniquely represents only one residue.

The properties of virtual bonds for P0 fragments were analysed on several arbitrarily selected high-resolution protein structures. SU positions can be calculated simply by selecting a group of atoms from the PDB file and by calculating the group geometrical centre. Virtual bond lengths and virtual angles can be derived from these positions. Parameters for typical secondary structures were calculated from a hexapeptide chain build by molecular modelling with all torsion angles equal to torsion angles given in Table 1. Results are given in Table 2.

In special cases, we allowed further freedom to fragments. This is a rigid rotation about the peptide bond from the *trans* to the *cis* configuration. The search fragments are also standard structure units.

2.3. Dual model representation

In the process of model building, two representations of the structure are used: Cartesian representation (or PDB representation) and FSM representation. These representations are equivalent and mutually interconvertible if bond lengths and angles in the structure and SUs are the same. Even if the bond angles and lengths are not the same (*e.g.* not fixed during refinement of the high-resolution crystal structure), the description of the crystal structure by the FSM is within reasonable accuracy, acceptable for practical use. Dual representation is used in calculating fragment overlap and in converting SUs to different types, *e.g.* two AlphaP0s can be converted into one AlphaA0. One AlphaA0 can be converted into two AlphaP0s if two external torsion angles are specified. cisPro0 can be converted into *cis*-AlphaP0 and *e.g.* one *cis*-AlphaP0 and one *trans*-AlphaP0 fragment can be converted into *cis*1-*trans*2-AlphaA0. The SU (generalized atom), expressed as the set of PDB coordinates, will be called PDB SU. Parameters of a SU can be converted into a PDB SU by simple matrix algebra. The rigid fragment is characterized only by its position and orientation and we can write

$$\mathbf{x}_{ic} = \mathbf{R}_p \mathbf{x}_{is} + \mathbf{O}\mathbf{x}_p \qquad (1)$$

$\mathbf{x}_p$ is the position vector of the SU, $\mathbf{R}_p$ is the rotation matrix of the SU, $\mathbf{x}_{ic}$ is the position vector of the *i*th atom of the PDB SU, $\mathbf{x}_{is}$ is the position vector of the Cartesian atom of the standard SU (search fragment) and $\mathbf{O}$ is an orthogonalization matrix. If the fragment is flexible, the standard SU is rotated first about torsion bonds into the required conformation. After a change of conformation, the modified standard SU is centred on the new geometrical centre and further treated as a new standard in equation (1).

Opposite transformation from Cartesian coordinates (PDB SU) to SU is not so straightforward because it requires calculation of orientation. The orientation is relative and is related to the orientation of the standard SU. Another problem is that the standard orientation is changed with the change of conformation of the fragment. This change is not unique and is associated with a definition of torsion groups (which part of the fragment is rotated and which part is fixed). The geometrical centre of the fragment, as was already mentioned, gives the position. Various methods can be used to calculate the rotation matrix. A simple scheme is used for P0 fragments in which fragment atoms are rotated to move both C^α atoms on the *x* axis. This is done for standard SU and PDB SU. At first, the fragment is shifted and one C^α atom is moved to a new origin, then the fragment is rotated around the *z* axis to move a second C^α atom to the *xz* plane and then the

Table 1
Fragments used for building protein structures, *r* is fragment radius.

Fragment	φ (°)	ψ (°)	*r* (Å)
AlphaP0	−64	−41	3.00
Beta1P0	−121	128	3.01
Beta2P0	−66	136	3.21
GammaP0	50	25	2.86
BridgP0	−105	65	3.03
cisPro0	−77	138	3.08
AlphaA0	−64	−41	3.21

Table 2
Virtual parameters for the P0 fragment in selected structures.

α is the conformation related to the AlphaP0 fragment, $\beta1$ is related to the Beta1P0 fragment and $\beta2$ is related to the Beta2P0 fragment. Distances (d) are in Å and angles (a) in °.

Code	$d_{\min}$	$d_{\max}$	$\langle d\rangle$	$\sigma(d)$	$a_{\min}$	$a_{\max}$	$\langle a\rangle$	$\sigma(a)$
α	2.64	2.64	2.64	0.0	103.6	103.6	103.6	0.0
$\beta1$	3.37	3.37	3.37	0.0	171.2	171.2	171.2	0.0
$\beta2$	3.07	3.07	3.07	0.0	146.9	146.9	146.9	0.0
2trx	2.24	3.99	2.97	0.38	93.6	179.4	131.7	30.2
2fdn	2.38	3.92	2.97	0.37	94.4	176.2	139.2	25.8
1rbg	2.36	3.83	3.00	0.36	95.6	175.3	140.9	26.9
1mfm	2.18	4.16	3.12	0.40	92.3	178.3	146.1	22.7
1igd	2.49	4.14	3.08	0.40	97.7	176.9	142.1	29.1
1bpi	2.52	3.69	3.03	0.32	98.1	176.2	140.2	25.1
1ab1	2.44	3.80	2.89	0.33	92.9	174.4	127.8	28.1
193l	2.17	3.81	2.86	0.32	95.0	176.1	126.4	26.0
3lzt	2.25	3.81	2.86	0.32	94.9	176.7	126.7	26.2

fragment is rotated around the y axis to move the second C^{α} onto the x axis. The last rotation is about the x axis to overlap both peptide oxygen atoms. The orientation matrix can be calculated as

$$\mathbf{R} = \mathbf{R}_{vz}^{T}\mathbf{R}_{vy}^{T}\mathbf{R}_{fx}\mathbf{R}_{fy}\mathbf{R}_{fz}. \tag{2}$$

$\mathbf{R}_{vz}$ and $\mathbf{R}_{vy}$ are rotation matrices for the PDB SU, $\mathbf{R}_{fx}$, $\mathbf{R}_{fy}$ and $\mathbf{R}_{fz}$ are rotation matrices for a standard SU (search fragment). The axis of rotation is specified by the index.

The AlphaA0 SU can be created from two AlphaP0 SUs. Two AlphaP0 SUs are transformed into their PDB representations. Relevant atoms are selected (duplicate atoms are averaged) to get a new AlphaA0 PDB SU. The geometrical mean of the PDB SU is calculated to get the position of the SU and the orientation matrix is calculated with respect to the standard AlphaA0. Eventually, torsion angles are calculated if the conformation is not standard. In fact, not all PDB SU atoms are required in this transformation. Three non-linear atoms and internal torsion angles are enough to calculate position and orientation of SU. This is utilized in various conformation searches.

2.4. Model building

2.4.1. Input information. The data necessary for model building are only: cell dimensions, space-group symmetry, structure factors (h, k, l, F_o or E_o and phase) and the number of residues in the asymmetric unit. The model building is divided into well defined steps.

2.4.2. Expansion of the electron density and rotation function. The method is described in PZO. The electron density is expanded in orthonormal spherical harmonics–normalized Bessel functions:

$$\begin{aligned}\rho(r,\theta,\varphi) &= \sum_{n=1}^{n_{\max}}\sum_{l=0}^{l_{\max}}\sum_{m=-l}^{l} a_{nlm}S_{nlm}(r,\theta,\varphi)\\ S_{nlm}(r,\theta,\varphi) &= g_l(k_{nl}r)Y_l^m(\theta,\phi).\end{aligned} \tag{3}$$

a_{nlm} are expansion coefficients. S_{nlm} are basis functions. $Y_l^m(\theta,\phi)$ is a spherical harmonics function. $g_l(k_{nl}r)$ is a normalized spherical Bessel function of order l, k_{nl} is such that $k_{nl}a = x_n$, where x_n are zero Bessel values, a is the radius of the chosen sphere of expansion.

An optimal number of expansion coefficients for radius 3.7 Å was found to be restricted by $n_{\max} = 5$ and $l_{\max} = 7$. The electron-density expansion is calculated directly from structure factors (utilizing both amplitude and phase) by FFT. Expansion coefficients are sorted for each grid point of the asymmetric unit and normalized. Coefficients represent several gigabytes of data, depending on the FFT grid. The grid step was 0.4 Å.

Fragment expansion coefficients are calculated for each of six P0 fragments given in Table 1. Two methods for calculation of the expansion coefficients of the fragments were used. For ‘Dirac fragments’, the expansion coefficients are calculated directly from Cartesian coordinates. For ‘F_c fragments’, equation (16) of PZO is applied. The structure factors are calculated for the same list of reflections as used for the electron-density expansion.

Molecular fragments are positioned by comparing crystal electron-density expansion and fragment electron-density expansion at different positions and orientations. This is done by the fast rotation function in the asymmetric unit of the unit cell. The search is limited to high-density regions by applying a cut-off. This step is computationally the most demanding. A rotation time depends on the number of grid points used for calculation. A 6D map is analysed by a peak-picking procedure. The peak picking was separated into a 3D search in an angular space (top orientation is saved) and a 3D search in (x, y, z) space (in PZO, we used a full 6D search). Positional and angular parameters are interpolated. Peaks are sorted.

2.4.3. Flexible and flipped refinement of Dirac fragments. Top peaks on the peak list are refined. We extended refinement of Dirac fragments by refinement of its internal torsion angles. The general strategy of the refinement is the same as for refinement of rigid fragments (details are in PZO). The torsion angle (τ) is changed by a fixed step. Cartesian coordinates of the fragment are recalculated for a new conformation and new coefficients of fragment expansion are evaluated. The target function is calculated at three points ($\Delta\tau$ = +10, 0 and −10°). The first and second derivatives are calculated by a numerical method to obtain the shift of the torsion angle. Several iterations are performed. If it is necessary to increase the radius of convergence of the refinement, the torsion angles are changed by 30° (greater change was not found useful) and the refinement procedure is repeated. The best refinement is accepted. The radius of convergence is about 50° for P0 fragments. During torsion refinement, the radius of the fragment is changed and the geometrical centre of the fragment should be recalculated. The result of refinement is a set of x, y, z, α, β, γ and $\Delta\varphi$, $\Delta\psi$ parameters. Refinement of torsion angles is also useful for establishing puckering of the five-membered ring in the cisPro0 fragment. In this case, the χ_4 torsion angle is refined.

Experience showed that one property of the refinement is that it moves fragment atoms to atomic positions in the crystal structure, with eventually one atom in the 'vacuum', rather than moving them to some mean position. In P0 fragments, the conformations of N and C^{β} (or C^{β}, C for right end) are sometimes incorrect with C^{β} occupying the N position and *vice versa* (the second atom at the H position). It is useful to make a second refinement in which φ and ψ torsion angles are changed by120 or 240° before refinement (flip) and to perform an independent refinement for all nine combinations. Large changes of torsion angles remove differences between fragments called AlphaP0, Beta1P0 *etc*. It is useful to recalculate torsion angles on one P0 SU; the AlphaP0 was selected for this purpose.

2.4.4. Sorting refined peaks. Various FOMs that are more sophisticated (sum, product, minimum) can be used in the sorting stage because it is possible to use an experimental electron-density map. We use CC (correlation between crystal electron density and atomic number of the fragment atom) at calculated atomic positions in addition to the overlap integral or correlation coefficient calculated by the rotation function (H). Refined peaks are characterized by a FOM describing how well the fragment fits into the electron density. Various sorting functions were tested. In current use is the formula

$$\begin{aligned}\mathrm{FOM} &= 0.35H + 0.65\mathrm{CC}(1 + 0.2\mathrm{SRO})\\ \mathrm{SRO} &= (1/n)\sum_i \rho_i/\rho_{\max}.\end{aligned} \qquad (4)$$

SRO is a scaled mean electron density. n is the number of atomic positions. Each peak of the rotation and translation function represents a potential group of protein atoms.

2.4.5. Connecting fragments. Connecting fragments is a principal step of the model building. The procedure is similar to procedures for building a 'single molecule' in small-molecule crystallography. The generalized atom (SU) is used instead of a normal atom. SUs and their symmetry equivalents are used for calculating virtual distances. Virtual bond distances should be in the ranges given in Table 2. For two SUs, which satisfy these criteria, overlap is calculated on the fragment level. SUs are converted into PDB SUs in Cartesian coordinates. Two important criteria are calculated. The first criterion is the distance between two C^{α} atoms. The second criterion is a mean distance for all four overlapping atoms in two P0 fragments. This overlap is defined as

$$\mathrm{FIT} = \left[\left(\sum d^2\right)/4\right]^{1/2}. \qquad (5)$$

d is the distance between related atoms (from the left end of the first fragment and from the right end of the second fragment). Whereas the C^{α} criterion is a strong indicator that two SUs could be connected, the second criterion tells us whether their conformations are correct. These overlaps also show which SU is closer to the N end of the chain and which is closer to the C end.

The total number of refined peaks from all rotation searches is several times more than the number of residues. In an ideal case, each peak has only two neighbours (or one neighbour for the end of the chain). In practice, the SU often has several SUs on virtual distances, some of them false. To make the process of construction of the structure model safe, a stepwise approach was developed.

In the first stage, there are two independent searches for basic secondary structures: α-helix and β-strand. These structures are well defined by virtual parameters. SUs, from the rotation function based on AlphaP0 fragment, are used for building the α-helix. SUs are connected if virtual distances are in the range 2.64 (30) Å and virtual angles are in the range 104 (20)°. Small pieces of α-helices are created. The chain is accepted only if it contains three or more residues.

For building β-structure, SUs for Beta1P0 and Beta2P0 are combined together in one 'pool'. Equivalent SUs are eliminated and from SUs on very short virtual distances only the SU with higher FOM is retained. Virtual distances are calculated and SUs are considered as connected if virtual distances are in the range 3.22 (40) Å and virtual angles in the range 160 (30)°. Chains of β strands contain SUs of both Beta1P0 and Beta2P0 fragments. Chains created for α and β structures are seeds for the building of a whole polypeptide chain and a base of the FSM.

In the third stage, $\alpha + \beta$ SUs in the FSM and SUs from all search fragments are combined together in one pool. Again, duplicate peaks are removed. SUs are sorted on the basis of their FOMs. SUs already in the FSM are given higher FOMs. Chains of the FSM are extended at both ends with any peak that is in a range of virtual distances 3.0 (9) Å, virtual angles in the range 135 (45)°, SUs have good C^{α} overlap (less than 1.0 Å) and good fragment overlap (FIT less than 1.7 Å).

In the fourth stage, the requirement for the group overlap (FIT) is dropped and only C^{α} overlap is used. In the fifth stage, all SUs that are a short distance from the FSM are deleted. Only peaks that are within 4.0 Å of an end of the chain are retained in the pool. At this moment, the pool peaks are refined by flipped (120°) refinement before connecting, to increase the number of correct conformations. All SUs are also converted into AlphaP0 type during the refinement. The limits on connecting are further liberalized to allow loop building. The polypeptide chain is usually formed with an exception of distorted (double) conformations. Each accepted SU is given a symmetry code for transformation from refined peak coordinates to 'single-molecule' coordinates.

2.4.6. Removing branches in a connectivity tree. During the process of connecting fragments, there are undesirable branches in the connectivity tree. Some of them are natural, *e.g.* double conformations, but many of them are branches having partial fits into side chains. The decision-making process is not an easy task and the algorithm used currently may not be the best or final one. In general, the protein is a linear polypeptide (S—S bridges form large cycles through cysteine side chains). SUs, which are causing cycling of the chain, are deleted. Accidental fitting into side chains may be detected by the connectivity index (unless short hydrogen-bonded side chains simulate a long chain). In cases of searching for secondary structures (steps 1–2), inclusion of the virtual angle is very discriminative. In a general case, information on the peak height (FOM), C^α overlap, group FIT, virtual angle and connectivity index are combined together in a combined figure of merit. Mean values and standard deviations are calculated from all virtual atoms. The CFOM is a sum (or weighted sum) of all criteria:

$$\mathrm{CFOM} = \sum_X (X - \langle X \rangle)/\sigma(X). \qquad (6)$$

The connection with the best CFOM is selected for a continuation in the chain building.

2.4.7. Model extension. Experience showed that the chain formed by connecting fragments ends either at a true chain end or in a side chain. The error occurs usually at the C end. One side of the fragment has a good overlap with the main-chain atoms and the second half of the fragment is at the position of the side chain. Double conformations are also problematic. A series of shorter chains is formed instead of one large chain. These chains should be extended and connected. The AlphaA0 fragment is used for extensions. Extension starts from $(n - 1)$ AlphaP0 SU because the last (n) is supposed to be corrupted. The AlphaA0 fragment is oriented in such a way that one (C^α—CO—N—C^α) peptide group is overlapped with the peptide group of the AlphaP0. Then a full 2D (φ, ψ) search is carried out for the second peptide group of the AlphaA0, with a step of 10°. The electron density in calculated atomic positions is used as a figure of merit:

$$\mathrm{FOMRO} = 0.4\mathrm{CC} + 0.6\mathrm{SRO}. \qquad (7)$$

CC and SRO are the same as in equation (4). Because a different fragment is used and because the search is carried out for a whole Ramachandran space, there is a great probability for getting different solutions. The conformation with the best FOMRO is refined in the expansion coefficient space. This step can be regarded as a correction step. The procedure is repeated with another AlphaA0 fragment, at the end of the previously built AlphaA0, to extend the chain. The peptide groups of AlphaA0 are converted into AlphaP0 SUs and again refined. SUs, generated at each chain end, are combined with FSM and the procedure for fragment connecting is repeated, resulting in chain extension and new chain connections. The independent connecting of all generated SUs also has some features of autocorrection of previously built models. The whole procedure is repeated several times and is finished (in an ideal case) when all chains are connected or when generated SUs are of very low FOM.

2.4.8. Chain overlap. Testing a chain overlap is another procedure for connecting smaller chains to a large chain. Virtual distances are calculated for all SUs, including symmetry equivalents. If the virtual distance is shorter than 4.0 Å, overlaps of individual atoms are calculated (also antiparallel β-strands are near that virtual distance). If the $C^\alpha \cdots C^\alpha$ distance is close to 3.8 Å, then AlphaP0 is inserted into an empty space and refined.

2.4.9. Building the AlphaA0 chain and refinement of the model. Polyalanine chains built from AlphaP0 fragments are affected by some incorrect conformations. These wrong conformations are introduced in the final stages of chain connecting, when C^α overlap is acceptable but total overlap is not very good. Positions of some (N, C^β or C^β, C) atoms may not be correct in P0 fragments. Some $C^\alpha \cdots C^\alpha$ connections, particularly in loop regions and in distorted groups, are too long. After connecting SUs, the positions of N or C atoms can be taken from neighbouring SUs and main-chain torsion angles can be defined more precisely. At the connection of two peptide groups, the AlphaA0 is generated with approximately correct torsion angles and orientation. The AlphaA0 is refined in the space of expansion coefficients. From refined AlphaA0 fragments, again AlphaP0s are generated and the structure is refined in this way. The final structure model can be represented by a scheme

$$\begin{array}{ccccccccccccc} & A & & A & & A & & A & & A & & A & \\ P & & P & & P & & P & & P & & P & & P. \end{array}$$

P is AlphaP0 SU and A is AlphaA0 SU. There are two independent chains. The A chain is shorter. A0 SUs are in spaces between two AlphaP0 and *vice versa*. These two chains represent two independent refinements of the protein structure. Because of a half-residue shift and because of overlapping peptide groups, we called this a *DOMINO* scheme. Correctness of the model building is increased by a mutual consistency of these two chains.

2.4.10. Creating the PDB file. In the final stage of the protein structure building, the SUs are converted into Cartesian coordinates. Atoms from overlapped groups are averaged. C^β atoms are taken from AlphaA0 groups. Atoms, at the start and the end of the chain are accepted only if atoms of the AlphaP0 and the AlphaA0 chains are in good agreement. The PDB file is created. This file can be input into visualization or refinement programs.

3. Results and discussion

A principal question in the early stages of the development of the method was 'which fragments should be used and how many (simple peptide group, monopeptide, dipeptide, tripeptide or larger)?'. Preliminary calculations showed that monopeptides and dipeptides are sufficient for building high-resolution protein structures and that fragments of about 20 atoms can be used at a resolution of 3.0 Å to locate rigid

Table 3
Results of automatic model building of protein structures.

Code is the PDB code or a structure code used in structure determination. Resid is the number of residues in the protein. Npdb is number of residues in the PDB file of the refined structure. Nres is number of residues found by the method described in this paper. % is percentage of residues found. c.m.p. is defined in equation (8). Resolution and c.m.p. are given in Å. R is published or final R factor.

Code	Resolution	Space group	R	Resid/Npdb	Nres	%	c.m.p.	References
1pen	1.1	$P2_1$	0.13	16/16	16	100	0.10	pdb
1ab1	0.9	$P2_1$	0.15	46/46	46	100	0.14	pdb
2fdn	0.9	$P4_32_12$	0.10	55/55	55	100	0.12	pdb
1rb9	0.9	$P2_1$	0.06	52/52	52	100	0.15	pdb
1g7a	1.2	$R3$	0.17	204/201	196	97.5	0.14	pdb
gibr	1.3	$I222$	0.10	387/386	386	100	0.11	(*a*)
9rnt	1.5	$P2_12_12_1$	0.14	104/104	104	100	0.21	pdb
1a75	1.9	$P2_1$	0.21	216/214	211	98.6	0.20	pdb
ica3	2.3	$P2_12_12_1$	0.17	–/584	435	74.5	0.32	(*b*)
tp47	2.3	$P3_221$	0.21	830/815	574	71.8	0.51	(*c*)

References: pdb Protein Data Bank; (*a*) Borek (2002); (*b*) Borek (2001), the exact number of residues is unknown because of autoprotolysis; (*c*) Tomchick (2001), DM data at resolution 2.3 Å, current PDB code is 1o75.

groups and α-helices. A single peptide group (five atoms) can be used for the electron-density-map interpretation just beyond atomic resolution. Unfortunately, the peptide group has the same geometry as atomic groups in His, Phe, Tyr and Trp. Asymmetric carbon is useful for unique fitting, for establishing chirality and for determination of the chain orientation. Preliminary calculations also showed that refined rigid fragments of β-strands were not accurate enough to make reliable connections of fragments. Refinement of internal torsion angles was proposed to solve the problem. Peptide-centred fragments (P0) are less sensitive to the actual conformation in the rotation function than A0 fragments (at the beginning also AlphaA0, Beta1A0 and Beta2A0 fragments were used in the rotation function). The P0 fragment has two chiral groups. The basic algorithm was finally designed for P0 fragments.

There are many false peaks in the rotation function originating in accidental similarity of conformations of side chains and search fragment, fragments fitted partially to main chain and partially to side chain, and many incomplete fits. An overlap with heavy atoms (sulfur, chlorine and metals) was almost avoided using a figure of merit based on the correlation coefficient rather than on the simple product function (product of the crystal electron density and the fragment density). Determination of structure is reduced to a problem of selecting the correct structure units among many false ones. Calculation of virtual bond lengths and angles is a tool to do this. In Table 2, virtual parameters for secondary structures (α-helices, β-sheets) are presented. Regular secondary structures have characteristic virtual bond lengths and angles. These virtual parameters are well separated. Average virtual parameters are rather conservative in protein structures. Particularly important for model building is the fact that the virtual angle is never smaller than 90°.

The model-building method was tested on several protein crystal structures. Structure factors and PDB coordinates were taken either from the PDB database or directly from the original authors. PDB coordinates were used to calculate phases. The R factors were usually slightly higher than in published structures because details of the refinement cannot be reproduced from the PDB file alone (*e.g.* solvent model). Nevertheless, this was considered as a useful (for test purposes) disturbance of otherwise very good phases. The test structures are given in Table 3. Three test files were experimental (gibr, ica3 and tp47) and phases were exported from mtz files. Only tp47 phases are not related to the final refined structure. The R factor was used as a criterion of the quality of data. Electron-density maps were not inspected because the philosophy is to build structures fully automatically. The method was evaluated on the basis of the number of residues found and connected. Only chains longer than four residues were considered. Another criterion is the accuracy of reproducing refined protein structure. Output of the model building is a PDB file of the protein structure model. The PDB file of the original structure was used for comparison and for calculating c.m.p:

$$\text{c.m.p.} = \left[\left(\textstyle\sum d^2\right)/n\right]^{1/2}. \tag{8}$$

d is the distance between the related model atom and the atom in the original PDB file, n is the number of atoms used.

1pen and gibr were dominant structures for development of the method and for debugging the relevant computer program (Pavelcik, work in progress). On the relatively large structure of gibr, the stepwise process of successive model building is demonstrated. The total number of refined peaks resulting from all rotation searches was 2707. The number of residues built at each step is reported. Step 1: α structure, 229. Step 2: $\beta1 + \beta2$, 96; $\alpha + \beta$, 325. Step 3: good C^α overlap and good FIT, 333. Step 4: good C^α overlap and any FIT, 376. Step 5: flip refinement and loop building, 380. Step 6: extension, 386 residues. Because limits on virtual distances and angles were not very strict, conformations close to α-helix were also built in the first step. In steps 1–2, a lot of smaller chains were formed. Principal steps for connecting chains into larger units were 4 and 5. Step 5 was designed also for refinement and improving the quality of the structure model before extending.

1pen and 1ab1 structures can be built by connecting (no extension needed) and 1pen can be built using only the AlphaP0 fragment. Small structures at atomic resolution (2fdn, 1rb9, 1g7a) were built without significant problems. The problems met during model building were connected usually with disordered conformations. Only one chain is built by the automatic procedure in this case. Disordered residues (25–28) of 2fdn have no corresponding peaks in *PROTF* (this is a significant limitation of the method because the rotation function is shape sensitive). These residues were built by the extension procedure, including the Asp28/Arg29 unusual peptide bond (modelled by a standard *trans*-peptide). The extension procedure is sensitive to some internal limits in the program, *e.g.* minimal FOM of the peak to be accepted and maximal allowed $C^{\alpha}\cdots C^{\alpha}$ distance. If these limits are strict then a limited number of disordered chains is built. If these criteria are liberated then false chain extensions are produced. Only connecting fragments can create a *cis*-peptide bond. The current extension procedure cannot create a *cis*-peptide bond.

Dirac and F_c fragments gave the same results for small atomic resolution structures. F_c fragments are slightly better than Dirac fragments in rotation function (and also better than E_c fragments used in PZO) for structure beyond atomic resolution. Dirac fragments are based on point scatterers. Scattering curves and temperature factors are used for the F_c fragments.

Structures 9RNT and 1A75 were used only to test performance of the method at resolution 1.5–1.9 Å. One can estimate from these tests and from results on ica3 that the practical limit to build a 'whole structure' is somewhere about 2.1 Å.

The accuracy of the model building is high, c.m.p. being about 1.5 of a standard uncertainty resulting from a structure refinement. The model building presented here can be regarded as a method of protein structure refinement in the real (electron-density) space.

At a resolution of 2.3 Å, only partial results were obtained. In tp47, experimental density-modified (DM) phases were used. Only 70% of the structure was built. In addition to that, 574 residues were distributed in 26 chains. The largest chain had 80 residues. There were also chains running in the opposite direction (from the C end to the N end). It seems clear that larger fragments would be required for safe model building at lower resolutions.

Refinement of torsion angles is one of the principal aspects of the method and a method of building accurate protein models. The number of fragments in rotation searches was reduced considerably. The flexible fragment concept may bring some changes into low-resolution crystallography. Structure units may be refined by a (rigid-body) least-squares program and each SU can be assigned temperature-factor tensor(s). Difference structure factors can be calculated and a new search can be done for missing SUs. The electron density may be inspected only for final checking, for correction of the model and for building non-peptide structure elements. A clone of the algorithm developed here can be used also for building nucleic acid polymers and some inorganic polymers like silicates.

The author is indebted to the University of Texas, Southwestern Medical Center at Dallas for financial support. The research was supported by grants NIH GM 53163 (USA) and VEGA MS 1/7264/20 (SR). The author is grateful to Dr Z. Otwinowski for helpful discussions and to Drs D. Borek and D. R. Tomchick for supplying data prior to publication.

References

Borek, D. (2001). PhD thesis, A. Mickiewicz University, Poland.
Borek, D. (2002). Private communication.
Cowtan, K. D. (1998). *Acta Cryst.* D**54**, 750–756.
Cowtan, K. D. (2001). *Acta Cryst.* D**57**, 1435–1444.
Greer, J. (1974). *J. Mol. Biol.* **84**, 279–301.
Greer, J. (1985). *Methods Enzymol.* **115**, 206–224.
Holton, T., Ioerger, T. R., Christopher, J. A. & Sacchettini, J. C. (2000). *Acta Cryst.* D**56**, 722–734.
Hovmöller, S., Zhou, T. & Ohlson, T. (2002). *Acta Cryst.* D**58**, 768–776.
Ioerger, T. R. & Sacchettini, J. C. (2002). *Acta Cryst.* D**58**, 2043–2054.
Jones, T. A & Thirup, S. (1986). *EMBO J.* **5**, 819–822.
Kleywegt, G. J. & Jones, T. A. (1997*a*). *Acta Cryst.* D**53**, 179–185.
Kleywegt, G. J. & Jones, T. A. (1997*b*). *Methods Enzymol.* **227**, 208–230.
Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.
Leherte, L., Fortier, S., Glasgow, J. & Allen, F. H. (1994). *Acta Cryst.* D**50**, 155–166.
Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.
McRee, D. (1999). *J. Struct. Biol.* **125**, 156–165.
Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 968–975.
Oldfield, T. (2002*a*). *Acta Cryst.* D**58**, 487–493.
Oldfield, T. (2002*b*). *Acta Cryst.* D**58**, 963–967.
Oldfield, T. (2003). *Acta Cryst.* D**59**, 483–491.
Pal, D. & Chakrabarti, P. (1999). *J. Mol. Biol.* **294**, 271–288.
Pavelčík, F., Sivy J., Rizzoli C. & Andreetii G. D. (1992). *J. Appl. Cryst.* **25**, 328–329.
Pavelcik, F., Zelinka, J. & Otwinowski, Z. (2002). *Acta Cryst.* D**58**, 275–283.
Perrakis, A., Morris, R. M. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Swanson, S. M. (1994). *Acta Cryst.* D**50**, 695–708.
Terwilliger, T. C. (2001). *Acta Cryst.* D**57**, 1755–1762.
Terwilliger, T. C. (2003). *Acta Cryst.* D**59**, 38–44.
Tomchick, D. R. (2001). Private communication.
Turk, D. (2001). *Methods in Macromolecular Crystallography*, edited by D. Turk & L. Johnson, pp. 148–155. Amsterdam: IOS Press.