Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Accurate automatic protein models

Frantisek Pavelcik

Department of Inorganic Chemistry, Faculty of Natural Sciences, Comenius University in Bratislava, 842 15 Bratislava, Slovak Republic

Correspondence e-mail: pavelcik@fns.uniba.sk

Received 23 October 2003
Accepted 22 June 2004

A method for automatic building of protein structures has been developed. The method is based on the concept of flexible structure units. A structure unit is a fragment (group) of about ten atoms that is positioned in the electron density by a phased rotation and translation function. The positions, orientations and internal torsion angles of all structure units are refined by a phased flexible refinement. Individual structure units are connected into polyalanine chains. The sequence is aligned by combined marker and rotamer methods. The side chains are built either by the marker method and a full conformational search or by the rotamer method. Side chains are independent structure units. The structure unit represents a generalized atom and the group model can be refined by the least-squares method. The protein model is built with an accuracy of about 0.2 Å at resolutions of 1.2–1.9 Å. Partial results can be obtained at resolutions of between 2.0 and 2.3 Å.

1. Introduction

Automatic model building is an open challenge in protein crystallography. Three principally different approaches have been developed so far.

(i) *ARP/wARP* (Lamzin & Wilson, 1993; Perrakis *et al.*, 1999; Morris *et al.*, 2002) is based on an interpretation of the difference electron density in terms of oxygen globs, iteratively followed by atomic refinement and interpretation of the atomic coordinates in terms of a polypeptide chain.

(ii) Greer (1974, 1985), Swanson (1994) and Leherte *et al.* (1994) devised procedures for tracing the path of the polypeptide chain and subsequent localization of C^α atoms. Skeletonization has been used, for example, by Levitt (2001), Oldfield (2002, 2003) and Turk (2001).

(iii) Positioning of protein fragments in the electron density was pioneered by Jones & Thirup (1986), Kleywegt & Jones (1997*a*,*b*), Cowtan (1998, 2001) and Terwilliger (2003). They fitted electron density with large rigid fragments from known protein structures in order to identify the location of helices, β-strands and other structures. The localization of fragments (*e.g.* helices) is an initial step in model building using computer graphics.

An advanced methodology (phased rotation and translation function, *PROTF*) for positioning molecular fragments in electron density has been described by Friedman (1999) and Pavelcik *et al.* (2002). A description of crystal structures in terms of flexible molecular fragments has recently been proposed by Pavelcik (2003). The position, orientation and internal torsion angles of molecular fragments can be refined in a space of spherical harmonics Bessel expansions. This refinement will be referred to as a phased flexible refinement (PFR). The method is different from other methods of protein

DOI: 10.1107/S0907444904015252

building and also has the potential to build nucleic acids and other polymer structures very accurately.

The method has previously been applied to the building of polyalanine chains in high-resolution protein structures (Pavelcik, 2003). The protein structure is built from a small set of carefully designed flexible units (groups) of about ten atoms. A structure unit (SU) is the principal building block and each SU behaves like a generalized atom. The group is described by a set of generalized coordinates (fractional coordinates x, y, z, Euler angles α, β, γ and torsion angles τ_1, $\tau_2 \ldots$). Bond lengths and angles within the SU are fixed and only the torsion angles are variables. The number of parameters describing the structure is significantly reduced compared with the atomic description. This reduction in dimensionality is in accordance with the reduction of observations at lower resolution compared with near-atomic resolution. The size of the group should follow the resolution change. Generalized atoms connected by virtual bonds form the flexible structure model (FSM). The SU expressed as a set of Cartesian coordinates in the PDB format is called a PDB structure unit. Atoms of two or more PDB SUs can be combined together and SU of a new type can be created. The description of the protein structure by the group model is general and different group models can be mutually interconverted.

Numerous computer experiments (Pavelcik, unpublished work) have shown that the principal difficulties in model building are related to disordered protein conformations. Another problematic area is accidental fitting of the main-chain search fragment into a side chain. Peaks of good height but low connectivity can hinder chain building in some cases. In this work, building of the polyalanine chain has been revised in order to improve its performance. The concept of virtual bond lengths and virtual angles was therefore abandoned, as well as stepwise model building based on α-helix and β-strand secondary structures. Instead, we introduce a concept based on flipped refinement and connectivity sorting of the *PROTF* peaks. The process of main-chain building was thus simplified and improved. Multiple-chain sequence assignment, post-sequence model building and building of side chains were developed. Here, we propose refinement of the group model by the least-squares method. The results presented suggest that lengthy analysis of electron-density maps could become avoidable in the near future.

2. Methods

2.1. Molecular fragments for protein building

The principles for fragment selection are the same as described previously (Pavelcik, 2003). The radius of an electron-density expansion was fixed at 3.7 Å. Fragment names are given in Table 1 together with fragment radii.

AlphaP0, Beta1P0, Beta2P0, GammaP0 and BridgP0 (Pavelcik, 2003) are peptide-centred nine-atom fragments,

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

Table 1
Fragments used for building protein structures.

r is the fragment radius.

Fragment	r (Å)	Fragment	r (Å)
AlphaP0	3.00	Beta1P0	3.01
Beta2P0	3.21	GammaP0	2.86
BridgP0	3.03	AlphaA0	3.21
CisPro0	3.08	SSlink	1.04
TrpM	2.98	Cys0	2.32
Asp0	2.57	Glu0	2.76
Phe0	3.59	His0	3.32
Ile0	2.71	Lys0	4.06
Leu0	2.63	Met0	2.89
Asn0	2.58	Pro0	2.11
Gln0	2.89	Arg0	4.37
Ser0	1.98	Thr0	1.90
Val0	1.96	Trp0	4.22
Tyr0	3.97	Aaa0	2.78

These fragments have the same valence geometry; they differ only in conformation. CisPro0 is analogous to the Ala-Pro fragment (11 atoms) for modelling the *cis*-peptide bond. AlphaA0 is a CA-centred fragment of helical conformation (ten atoms),

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

Side-chain fragments have a general connectivity given by the chemical formula

$$\mathrm{N-C^{\alpha}(C)-}R.$$

The main-chain atoms N, CA and C are part of the side-chain fragments and there is an overlap of four atoms with AlphaA0. Fragments were built by molecular modelling. Side-chain fragments are flexible. χ torsion angles can be changed and refined. Only two torsion angles are usually optimized in conformation searches. Aaa0 is a special fragment used in the model building of long flexible amino acids (Glu, Gln, Arg and Lys) to fit the requirement of a 3.7 Å sphere and also to avoid extensive conformation searching,

$$\mathrm{N-C^{\alpha}(C)-C^{\beta}-C^{\gamma}-C^{\delta}}.$$

The five-membered ring of Pro0 is planar and is disconnected at the CG—CD bond. χ_1 and χ_5 torsion angles can be optimized during refinement. χ_5 is usually coupled with the main-chain torsion angle φ. TrpM and SSlink are special fragments (markers) used in *PROTF* to search for the positions of electron-rich groups to help in establishing the sequence. SSlink is a disulfide bridge of two S atoms. TrpM is a rigid part of tryptophan (ten atoms, starting at the CB atom).

A general abbreviation for the peptide-centred fragments in this paper is P0, for CA-centred fragments A0 and for side chains S0.

2.2. Algorithm of the protein building

Details of individual techniques can be found in Pavelcik (2003) or Pavelcik *et al.* (2002) or are given later in this paper.

(i) Input. The basic inputs for the protein building are unit-cell parameters, space-group symmetry, sequence and structure-factor amplitudes and phases. Sequence alignment is essential for side-chain building; the sequence is provided in single-letter code with 52 amino acids per line. Chain letters (such as A, B) and the total number of residues in a chain are given in a similar fashion to that used in the PDB format; for example, for 1a75,

```
SEQV  A  108  AFAGILADADCAAAVKACEAADSFSYKAFFAKCGLSGKSADDIKKAFVFIDQ
SEQV  A  108  DKSGFIEEDELKLFLQVFKAGARALTDAETKAFLKAGDSDGDGAIGVEEWVA
SEQV  A  108  LVKA
SEQV  B  108  AFAGILADADCAAAVKACEAADSFSYKAFFAKCGLSGKSADDIKKAFVFIDQ
SEQV  B  108  DKSGFIEEDELKLFLQVFKAGARALTDAETKAFLKAGDSDGDGAIGVEEWVA
SEQV  B  108  LVKA
```

(ii) Electron-density expansion. The radius of the electron-density expansion was fixed at 3.7 Å. The maximal indices for spherical harmonics and Bessel functions are $n_{\max}$ = 5 and $l_{\max}$ = 7. A grid step in the FFT transform is 0.4–0.5 Å. Coefficients of the expansion are stored.

(iii) *PROTF* and PFR. The phased rotation and translation function is calculated for AlphaP0 (1.5*N*), Beta1P0 (1.0*N*), Beta2P0 (1.0*N*), GammaP0 (0.2*N*), BridgP0 (0.2*N*) and CisPro0 (0.025*N*) fragments. *PROTF* peaks are refined by phased flexible refinement. The number of accepted peaks for each fragment type is given in parentheses; *N* is the number of protein residues.

(iv) Polyalanine-model building. *PROTF* peaks are refined for the second time by the flipped PFR. Peaks are sorted on the basis of peak height and connectivity (overlap). Peaks are connected into chains. The AlphaP0 structure units are transferred into AlphaA0 SUs. The FSM of polyalanine chains is created. Details are given in the next section.

(v) Marker *PROTF*. The *PROTF* peaks are calculated for TrpM, SSlink and Phe0. The number of accepted marker peaks depends on the sequence.

(vi) Sequence alignment. Rotamer side chains and marker fragments are used for the calculation of scoring tables. Each FSM chain is aligned with each sequence and the best solution is accepted.

(vii) Post-sequence modelling. The model is corrected in order to conform to the known sequence. This step includes a rebuilding of incorrect conformations, AlphaA0 refinement, sequence-driven chain extension, small loop building and new chain connections.

(viii) Side-chain building. Side chains can be constructed by the marker and full conformation search or by the rotamer method. Multiple side chains at both ends of the main chain are built. Conformations at the chain-end positions are optimized.

(ix) Flexible refinement of side chains. This includes PFR of side chains and single-point refinement of large side chains (Lys, Arg or Trp).

(x) Refinement of group temperature factors by the least-squares method.

(xi) Creation of a coordinate file of the protein structure.

2.3. Polyalanine model

The whole concept was redesigned compared with that described in Pavelcik (2003). The model refinement was the last step in the previous algorithm. The flipped refinement, which was a marginal technique in the refinement, has become one of the principal aspects of the present method. It is introduced at the very beginning of the connecting process as a way of improving (φ, ψ) conformations and thus the quality of fragment overlaps. Building of the main chain can be divided into four well defined steps as follows.

(i) Refinement and sorting. *PROTF* peaks from all P0 search fragments are combined together into one 'pool'. Duplicate peaks are removed. All remaining peaks are recalculated for the AlphaP0 type and refined by flipped refinement. The flip of (φ, ψ) torsion angles is 120 and 240°. Inter-peak distances are calculated. For two peaks at a short distance, DCA and FIT are evaluated. DCA is the CA—CA distance (DCA = *X*, *n* = 1). FIT is an overlap of two fragments at the CA, N, C and CB atoms (FIT = *X*, *n* = 4).

$$X = \left(\frac{\sum d^2}{n}\right)^{1/2}. \tag{1}$$

d is the distance between two related atoms. Refined peaks are sorted on basis of peak height and peak connectivity. Peaks with two good DCAs and two good FITs are given preference in sorting. The overlap contribution is calculated by

$$w = \exp\left[-\frac{\mathrm{DCA}}{K\sigma(\mathrm{DCA})}\right]\exp\left[-\frac{\mathrm{FIT}}{K\sigma(\mathrm{FIT})}\right], \tag{2}$$

where σ(DCA) and σ(FIT) are estimated from all overlaps. *K* is an empirical parameter (*K* = 4). CisPro0 fragments are transformed into Cis-AlphaP0, but are not used in sorting.

(ii) Connection. The connection procedure (Pavelcik, 2003) was simplified to two steps. In the first step, only chains with SUs having all good FITs are accepted. Many smaller chains are created. These chains correspond to rigid parts of the protein structure and are the only seeds for further chain building. In the second step, the requirement for a good FIT is dropped and all SUs with good DCAs are connected. Cis-AlphaP0 peaks are added.

(iii) Extension. The published procedure (Pavelcik, 2003) was modified. Two conformations are generated instead of one and refined. The procedure is more powerful for extending because the extension can take place in two directions. One extension is usually in the direction of the main chain and the second extension in the direction of the side chain. One of the directions is usually correct. The *PROTF* peaks are reused after extension. A *cis*-peptide bond can also be created in the extension process based on the CisPro0 rotation peak.

(iv) Conversion. The AlphaP0 SUs are converted into AlphaA0 structure units. During this conversion, the few remaining wrong conformations are corrected (Pavelcik, 2003). AlphaA0 SUs are refined. The structure model may contain several missing residues and a larger number of polypeptide chains than expected from the protein sequence. This FSM is input into sequence alignment.

2.4. Sequence alignment

2.4.1. Scoring tables. For sequence assignment, a scoring table $M \times N$ is created, N being the number of AlphaA0 structure units in the FSM. M is number of natural amino acids ($M = 18$; Gly and Ala are not considered). Scores are related to the probability of occurrence of a particular amino acid at the given chain position. Unassigned amino acids are given scores of zero. Two methods for table creation were developed.

(i) Marker method. The position of a large side chain with limited torsional freedom can be obtained directly from *PROTF*. The special markers TrpM and SSlink, as well as Phe0, which also represents Tyr, are used. The His0 fragment also has some potential for early recognition. First, the asymmetric unit is searched. Peak positions are refined (flexible refinement for the flexible fragments) and sorted. The number of accepted peaks for sequence alignment is only slightly higher than the number of related amino acids in the sequence. Distances are calculated between AlphaA0 SUs and marker positions. If the distance corresponds to a bond distance, then the overlap of individual atoms (FIT) is calculated. For His0 and Phe0 fragments, CA and CB atoms are considered. For TrpM markers, CB (CG) or NE1 atoms are used (tryptophan can be in a pseudo-symmetric position with NE1 in place of CG). For SSlink markers, the S$\cdots$CB distance is calculated and an absolute value of difference between the C—S bond distance and standard C—S bond distance is used as the FIT.

If FIT < ERR then the appropriate place in the scoring table is assigned a value of 1 (ERR is an empirical parameter, ERR = 1.5 Å). If more than one marker contributes to the same A0 position (*e.g.* tryptophan and histidine), the marker with a higher FOM in PFR is given preference. SSlink and TrpM markers are unique. The Phe0 marker files position Phe and Tyr. All other positions in the table are zeros.

(ii) Rotamer method. All rotamers for 18 amino acids are built for each AlphaA0 SU. The electron density at calculated atomic positions is evaluated. The score is an FOM,

$$\mathrm{FOM} = 0.4\mathrm{CC} + 0.6\mathrm{SRO},$$
$$\mathrm{SRO} = \frac{1}{n}\sum_i \frac{\rho_i}{\rho_{\max}}. \tag{3}$$

CC is a correlation coefficient between the electron density and the atomic number of a fragment atom at the calculated atomic positions. SRO is a scaled mean electron density and n is the number of atomic positions. The rotamer with the best FOM is accepted and the value of the FOM is used directly in the scoring table. Lovell *et al.* (2000) rotamers were used in calculations.

Rotamer or marker scoring tables can be used separately for sequence alignment or can be combined together. Tables are normalized. Mean and standard derivations are calculated and values in tables are scaled by

$$S_{\mathrm{scaled}} = \frac{S - \langle S\rangle}{\sigma_S}. \tag{4}$$

The sum of scaled values is used as a new score. See Zou & Jones (1996) for a related approach.

2.4.2. Multiple-chain sequence alignment. The protein can be composed of N_{seq} sequence chains. The FSM consists of N_{fsm} chains of A0 structure units. Two or more smaller chains may be related to a single sequence chain. On the other hand, a false structure unit can artificially connect some A0 chains and this means that two sequences have to be assigned to one FSM chain. In the process of sequence alignment, each FSM chain is aligned with all sequences. An alignment is a one-dimensional search. The relative position of the chain and sequence is given by an offset parameter j. An FOM is calculated for each offset as a simple sum of scores,

$$\mathrm{FOM}(j) = \sum_i^k S_i(j). \tag{5}$$

S_i is the score (from the scoring table) for an amino acid (given by the sequence) and a particular SU. k is the number of residues common to both chains. The three top FOMs and related offsets are stored for each chain/sequence pair. Two tables $N_{\mathrm{seq}} \times N_{\mathrm{fsm}} \times 3$ are formed (one for FOMs and the second for related offsets). The pair with the highest FOM is aligned first. If the sequence is longer than the chain, then the assigned part of the sequence is removed from consideration and the second highest FOM and offset are moved to the first position in both tables. This process is repeated until all pairs are aligned.

2.5. Post-sequence modelling

When the sequence is aligned, it becomes clear where the boundaries of the main chain are: some residues may still be missing, while others may be extended too far by false extension (to chemically bonded groups at the chain ends or to side chains). Some smaller chain of A0 structure units may not be connected into a longer chain because of corrupted end fragments. There may also be larger gaps in the correlation table of the sequence *versus* structure units in the loop or in the random-coil regions.

The first step of post-sequence modelling is completion of the FSM with AlphaA0 structure units at the beginning and end of each chain. The procedure is essentially the same as the extension procedure (Pavelcik, 2003). This extension is well founded for the right end of the chain (only the CA atom of the A0 has no counterpart in the sequence-related atoms), but at the left end of the chain there are three excess atoms (CA, C, O) unless the chain is substituted (*N*-acetyl, *N*-formyl *etc.*). Because of this forced extension, some residues may not have the correct conformation.

If one sequence chain contains two or more FSM chains and if the number of missing structure units is 0–1 (0 corresponds to unconnected chains because of corrupted residues on connection), these missing SUs can be constructed or corrected and refined. At chain connections the interatomic distances between atoms of neighbouring structure units are analysed (the CA—CA distance should be either 0 or 3.8 Å), false atoms are identified and new coordinates for AlphaA0 are created. Symmetry codes for one chain have to be recalculated in order to obtain one continuous chain. Building a larger part of the missing chain (or disordered conformations) is a more complex problem that will be addressed in a future paper.

The conformations at the end positions are generally uncertain. The only way of finding a chain-end conformation is to build the side chain first. For this reason, a pseudo-C atom is modelled at a hydrogen position on the CA atom of the end structure unit. The CA atom has three singly bonded atoms and each of them can be CB. Three side chains are modelled (see next section) and refined by PFR. The best of them is accepted. When the side-chain position is fixed then the positions of CB and N (or CB and C for the right end) become available. The A0 SU is again constructed and refined. Theoretically, only terminal Ala and Gly conformations are not fixed by the rest of the structure. In an ideal case, the structure model is completed and the model contains all A0 structure units. However, the unmodified N-terminus of the chain is better represented by the S0 structure unit.

2.6. Building side chains

2.6.1. Marker method. Information from *PROTF* on special markers can be used to build side chains. The positioned Phe0 marker is in fact an SU of the side chain. The Cartesian coordinates of the S atoms of the SSlink are combined with the Cartesian coordinates of N, CA, C and CB of the AlphaA0s and two PDB Cys0 units are created. The PDB Cys0 SU is transformed into the Cys0 SU and refined. A similar procedure is used to create a Trp0 structure unit from the TrpM marker.

2.6.2. Full conformational search. Side-chain fragments contain N, CA and C atoms in addition to the standard side-chain atoms. These atoms fix the side chain to the main chain. At first, torsional groups of the side chain are rotated along chemical bonds (χ_1, χ_2) in order to obtain the required conformation. The side chain is then moved and rotated to fit N, CA, CB and C of the AlphaA0 SU. A one-dimensional (χ_1) search is carried out for Cys, Ser, Thr and Val. A two-dimensional search (χ_1, χ_2) is carried out for Aaa0, Leu0, Ile0, Asp0, Asn0, His0 and Phe0. The step for the torsion angle is 10°. Geometrical tests are used to avoid an overlap of the CD atom of the side chain and atoms of AlphaA0. (χ_1, χ_5) torsion angles are optimized for Pro0. Conformations having large CG—CD or CG—N distances are discarded. Distance tests are also carried out for the CD—C distance so that the conformation is consistent with the main-chain torsion angle φ. The electron density is evaluated at calculated atomic positions. The conformation with the best FOM based on the electron density is accepted. The side-chain parameters are refined by PFR (all above-mentioned fragments are within the 3.7 Å radius sphere).

Larger flexible side chains (Glu, Gln, Arg and Lys) are represented in this search by a special Aaa0 fragment in order to avoid extensive multidimensional searches. In the second step, (χ_1, χ_2) torsion angles are fixed at the refined position of Aaa0 and another (χ_3) one-dimensional or (χ_3, χ_4) two-dimensional search is carried out. Glu and Gln fragments can be refined by the standard refinement procedure. Arg, Lys and Trp are outside the 3.7 Å limit and are not refined using pre-calculated expansion coefficients. Tyrosines are treated as phenylalanines. The position of the phenolic O atom is calculated assuming sp^2 hybridization and the known length of the C—O bond.

A methionine has three variable torsion angles. It does not fit the Aaa0 building scheme as it has a long C—S bond instead of a single C—C bond. The methionine is treated within the (χ_1, χ_2) scheme. χ_3 is only considered in two conformations (110 and 290°). The best conformation is selected from $36 \times 36 \times 2$ points in the electron-density space. All three torsion angles are refined.

The tryptophan position is usually built by the marker method. Alternatively, it can be built by modelling in the electron density and refined by the single-point refinement method (see below).

2.6.3. Rotamer search. The method is the same as described in §2.4. There is no need to divide the search into two steps. The best electron-density fit is refined by PFR.

2.6.4. Single-point refinement. In the standard refinement procedure, the expansion coefficients are pre-calculated by FFT for each grid point of the asymmetric unit and stored. The radius of expansion is 3.7 Å. This radius is small for the large side chains Arg, Lys and Trp. The position of the SU is known from the modelling in the electron density. The expansion coefficients can be calculated for this point by a direct summation of structure factors. The radius of expansion was selected as 4.5 Å. Side chains are refined by PFR. The single-point refinement is a computationally costly process.

2.7. Least-squares refinement of the group model

Structure factors can be expressed as

$$F_c = \sum_j \sum_k f_k \exp[2\pi i \mathbf{H}(\mathbf{r}_j + \mathbf{O}^{-1}\mathbf{r}_k)] \exp(-B_j s^2),$$

$$F_c = \sum_j [\sum_k f_k \exp(2\pi i \mathbf{h}\mathbf{r}_k t)] \exp(2\pi i \mathbf{H}\mathbf{r}_j)] \exp(-B_j s^2). \quad (6)$$

In this formula, $\mathbf{r}_j$ is the positional vector of the SU in fractional coordinates and $\mathbf{r}_k$ is the positional vector of a group atom in Cartesian coordinates within the SU (local coordinate system). $\mathbf{O}^{-1}$ is the inverted orthogonalization matrix. $\mathbf{H}$ is the reciprocal vector in fractional coordinates; $\mathbf{h}$ is the same vector in Cartesian coordinates. B is an isotropic group temperature factor. The vector $\mathbf{r}_k$ depends on the group orientation and conformation but not on the group position. These coordinates are taken from PFR. See Navaza (2001) for a similar formulation.

The classical least-squares refinement is based on this scheme. Only the scale, group positions (not orientation) and group temperature factors were refined. The least-squares refinement benefits from a full transformation of the P0–A0 structure model to the A0–S0 structure model. In general, there are two structure units (A0, S0) per residue and the

Table 2
Test protein structures.

Code is the PDB code or the structure code used in structure determination. $n_{\rm SEQ}$ is the number of residues in the protein. $n_{\rm PDB}$ is the number of residues in the PDB file of the refined structure. $n_{\rm CHAIN}$ is the number of protein chains. Resolution is given in Å.

Code	Resolution	Space group	$n_{\rm SEQ}$	$n_{\rm PDB}$	$n_{\rm CHAIN}$	Reference†
1a32	2.1	$P2_12_12_1$	88	85	1	PDB
1a75	1.9	$P2_1$	216	214	2	PDB
1ab1	0.9	$P2_1$	46	46	1	PDB
1bfe	2.3	$P4_132$	119	110	1	PDB
1g7a	1.2	$R3$	204	201	8	PDB
1pen	1.1	$P2_1$	16	16	1	PDB
1rb9	0.9	$P2_1$	52	52	1	PDB
2fdn	0.9	$P4_32_12$	55	55	1	PDB
9rnt	1.5	$P2_12_12_1$	104	104	1	PDB
GIBR	1.3	$I222$	387	386	1	b
TP47	2.3	$P3_221$	830	815	2	t

† PDB, Protein Data Bank; b, Borek (2002); t, Tomchick (2001).

number of parameters to be refined is $8N$, where N is the number of amino acids in the protein.

Some atoms are included twice in the structure-factor calculation because of group overlap. This overlap is corrected by setting occupation factors of smaller than one. The group contribution is the same in each refinement cycle and can be pre-calculated for each reflection. Alternatively, only group temperature factors and the overall scale can be refined.

3. Results and discussion

The model-building method was tested on several protein crystal structures. The basic crystallographic data of these structures are given in Table 2. Structure factors and Cartesian coordinates were either obtained from the PDB or directly from the authors. Phases were calculated from the coordinates or were taken from the final stages of refinement. The phases for 1pen are the result of automatic structure determination (see Pavelcik *et al.*, 2002). Three phase sets were used for 9rnt and 1a75 in order to study the influence of phase error on model building. The 9rntA and 1a75A phase sets were calculated from polyalanine atoms (and metal atoms); the 1a75G and 9rntG phases were calculated from polyglycine atoms (and metal atoms). Two phase sets are used for TP47. Experimental density-modified phases (based on SAD, phased with Xe) at a resolution of 2.3 Å represent the first set. The second phase set was calculated from refined PDB coordinates of 1o75 at the same resolution. Observed structure-factor amplitudes are used in all calculations.

Details of the improved main-chain connection and extension are summarized in Table 3. Results in Table 4 are based on F_c fragments [see Pavelcik (2003) for definitions of Dirac and F_c fragments]. The number of residues traced was the main criterion for the evaluation of the method. The accuracy of the model building was the second criterion. The accuracy is reflected in the parameters CON and NABC. These parameters are not refined by any of the procedures and can be regarded as independent estimates (*i.e.* analogous to an $R_{\rm free}$ factor). CON is a mean overlap distance of the peptide-group atoms of two neighbouring A0 structure units (N, C and O atoms are used, not CA). NABC is an analogous criterion for A0 and S0 (overlap of N, CA, CB and C). The 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 to calculate the root-mean-square error CMP. All three criteria are calculated using (1). CMP is calculated from all common atoms in the model and in the PDB file.

Table 3
Details of the automatic main-chain building based on F_c fragments.

$N_{\rm peak}$ is the number of refined pool peaks. Conn3 is the result of building chain seeds. Conn5 is the result of connection. Exten is the result of extension. The numbers given are number of built chains/the total number of all connected P0 structure units. ⟨DCA⟩ is the mean CA···CA distance at structure-unit connections in Å.

Code	$N_{\rm peak}$	Conn3	Conn5	Exten	⟨DCA⟩
1a32	230	7/71	3/74	2/88	0.28
1a75	561	20/175	9/191	2/216	0.29
1ab1	104	4/43	1/45	1/47	0.11
1bfe	121	9/44	9/48	2/100	0.45
1g7a	450	19/170	16/176	8/216	0.21
1pen	24	2/15	2/15	1/17	0.12
1rb9	115	3/49	2/50	1/53	0.14
2fdn	165	5/49	4/49	1/55	0.16
9rnt	236	12/96	1/102	1/104	0.25
GIBR	880	26/362	7/370	1/387	0.17
TP47	1880	92/511	48/593	17/789	0.53
dmTP47	1937	77/417	55/489	20/681	0.53
1a75A	371	17/192	6/204	2/216	0.29
1a75G	451	29/160	9/186	2/216	0.39
9rntA	183	11/99	7/100	1/105	0.24
9rntG	201	14/97	3/102	1/105	0.29

Table 4
Sequence alignment and the accuracy of the automatic model building of protein structures.

$N_{\rm res}$ is the number of residues found after post-sequence modelling. S/N is the smallest of the signal-to-noise ratios in the sequence alignment. CMP, CON and NABC are defined by (1) and are given in Å. NABC1 and CMP1 are the results of the marker and full search method. NABC2 and CMP2 are the results of the rotamer method. RG is the R factor of the group refinement.

Code	$N_{\rm res}$	S/N	CON	NABC1	NABC2	CMP1	CMP2	RG
1a32	85	1.5	0.76	0.23	0.17	0.30	0.34	0.42
1a75	215	4.4	0.21	0.21	0.18	0.22	0.25	0.31
1ab1	46	6.4	0.08	0.09	0.09	0.08	0.11	0.25
1bfe	103	1.9	0.37	0.22	0.19	0.36	0.39	0.47
1g7a	204	2.3	0.15	0.17	0.16	0.13	0.15	0.31
1pen	16	4.5	0.12	0.13	0.15	0.09	0.11	0.21
1rb9	52	5.4	0.08	0.11	0.11	0.08	0.12	0.34
2fdn	55	3.0	0.10	0.12	0.12	0.11	0.14	0.34†
9rnt	104	5.6	0.19	0.18	0.18	0.16	0.19	0.31
GIBR	387	16.8	0.11	0.14	0.13	0.12	0.16	0.33
1a75A	216	1.8	0.20	0.21	—	0.35	—	—
1a75G	215	1.1	0.28	0.22	—	—	—	—
9rntA	104	3.1	0.19	0.18	—	0.21	—	—
9rntG	104	2.9	0.20	0.20	—	0.25	—	—

† Fe_4S_4 clusters were located by *PROTF* and included in the least-squares refinement.

Another criterion for the evaluation of the method is an R factor from the group refinement (RG). The group temperature factors were compared with individual temperature factors in the PDB file.

Electron-density maps were not inspected. A direct analysis of the map by a graphics program should be avoided as far as possible in automatic structure determination.

Calculations were carried out on a 2 GHz Intel Celeron CPU with 512 Mb of RAM under Windows XP with a Fortran90 program compiled by a Compaq (Digital) compiler. The total computer time (not the CPU time) was approximately 1 min per residue. Parallel tests were performed under Linux Red Hat 9 with a g77 compiler.

The results of the polyalanine building (Table 3) show a better performance of the new connecting algorithm compared with the previous method (Pavelcik, 2003). The flipped refinement increased the number of AlphaP0 structure units with the correct (φ, ψ) conformation. The refinement also increased the peak height and moved the true peaks higher on the sorted peak list. The overlap-based sorting of the *PROTF* peaks is approximately equivalent to looking for an optimum path through a non-perfect and complicated connectivity graph.

Despite this improvement, the whole structure cannot be constructed solely by the connecting approach, because some SUs belonging to disordered residues are simply not present in the peak list (*PROTF* is shape-sensitive and the presence of higher electron density is not a guarantee of finding a *PROTF* peak). Extension is of prime importance for building disordered chains. The efficiency of the extension algorithm was increased by reusing *PROTF* peaks. Very small chains that were omitted in the connection procedure can contribute to chain extension.

In the new algorithm, the P0 chain is discarded after transformation to the A0 chain. The transformation from P0 to A0 represents model refinement. The main-chain (φ, ψ) torsion angles are known with sufficient accuracy if the DCA is not large. AlphaA0 structure units are constructed and refined in a space of spherical harmonics Bessel expansions. The disadvantage of AlphaA0 in *PROTF* (a flexible central part) compared with AlphaP0 (a rigid central part) is useful in the refinement. The torsion angles (φ, ψ) and the position of the CB atom are stereochemically fixed by two rigid peptide groups. For this reason, AlphaA0 SU is better suited for building of side chains than AlphaP0 SU. Checking the overlap of two peptide groups of neighbouring AlphaA0 SUs can immediately detect a suspicious part of the chain. The corrupted residue or missing residue can be corrected in post-sequence modelling. A more sophisticated routine is needed (see the case of 1a32) for building larger loops (Pavelcik, work in progress).

The signal-to-noise ratio in the sequence alignment is high (the noise is represented by the next highest maximum) because the chains are of sufficient length. The marker method is a relatively safe method of sequence alignment if disulfide bridges and tryptophan residues are present in the protein. The rotamer method gave the same results. The data in the table are given for combined results.

During the post-sequence modelling, structure units at chain ends are built. This brings one more CA atom to each chain end (unless a 'false' residue has already extended the chain). Main chains in most of the tests were completed and new inter-chain connections were formed.

After the sequence has been assigned, the building of side chains is a rather simple and well established process. The results, given in Table 4, are encouraging. All methods gave good results. The rotamer method is the fastest method. The full conformation search is a method for building unusual conformations, *e.g.* conformations of hydrogen-bonded residues or residues involved in short van der Waals contacts. This is a good method of establishing the conformation in a group with a low energy barrier (*e.g.* an end carboxyl group). A disadvantage of the full conformation search is that the end of a long side chain may fall onto the main-chain atoms if the electron density of the side chain is not well defined. Many such cases are excluded by distance testing. The distance tests are limited to distances within one AlphaA0 SU only. The PFR can correct for slightly displaced CB atoms and is able to improve torsion angles, but is not able to correct wrong conformations.

A new method in side-chain building is the marker method. The marker method may build some electron-rich side chains even if the sequence is not known. The method was used to build Cys, Phe, Tyr and Trp. The marker and full search method gave slightly better results in overall accuracy than the rotamer method, as can be seen from values of CMP1 and CMP2 in Table 4. Detailed comparison of individual side chains showed that the rotamer method gave more chains with wrong conformations.

The optimum method for side-chain building will probably be a combination of all methods. The conformations with a threefold torsion barrier (rotation about a single carbon–carbon chemical bond) are predictable (*gauche* and *trans*; torsion angles close to +60, 180 and −60) and these can be built safely by the rotamer method. Carboxyl, amidic and guanidine groups can be built by the full conformation search. In a more sophisticated procedure, hydrogen-bond contact information can be utilized. This may resolve the problem of amidic conformations in Asn and Gln. The marker method is probably best for tryptophan, phenylalanine, tyrosine and cystine.

The least-squares refinement of the group model presently serves only as a source of group temperature factors for the PDB file and for calculating the final *R* factor. The 'halfway' approach (only refining scale, position and temperature factors, but not orientation and conformation) was not successful. The independent parameters such as CON and NABC were not improved by the refinement. There is an open question whether to write a full 'flexible-rigid body' refinement program or to leave this problem to the current refinement programs. Only temperature factors and overall scale were finally refined. Nevertheless, the group temperature factors follow the temperature factors published in coordinate files. RG factors seem to be reasonable for the group model with only heavy peptide atoms and no solvent correction.

Details of the modelling are discussed for each structure separately.

3.1. 1a32

The number of reflections is 6149. There are (on average) 16 parameters needed to describe the geometry of one residue (the temperature factor is not included). The observation-to-parameter ratio is 4.4. In small-molecule crystallography this ratio is usually about 10. Although the number of found residues is 85, the chain is interrupted at His17 and Glu18. The histidine side chain of the model is situated at the position of the main chain of the protein and Glu18 is incorrect. In addition, Asn19 and Asp20 are not modelled accurately. These residues have the highest temperature factors in the group refinement. The loop-correcting procedure is not designed to correct two or more corrupted residues.

3.2. 1a75

This structure is close to the limit where the observation-to-parameter ratio of the group model is the same as the ratio for an atomic model at resolution 1.2 Å. The connection and extension of fragments led to the formation of two chains. In chain A one residue at the N-terminus was not found (two residues are missing in the PDB file). There are only a few errors in the side-chain building.

3.3. 1ab1

This structure was built practically without errors. There is a different conformation of Arg17 in comparison with the original PDB file.

3.4. 1bfe

The resolution of this cubic structure is 2.3 Å (about three reflections per parameter). A limited number of structure units was found by *PROTF*. This is reflected by the fact that only 48 residues were connected. Half of the structure was constructed by extension. *PROTF* is losing its power to reveal the structure with search fragments of about ten atoms. Two chains were constructed by connection and extension instead of one. The sequence was assigned correctly. These two chains were connected in post-sequence modelling. The first residue and six residues at the C-terminus were not found in comparison with the PDB file. The high temperature factor region 319–321 was modelled differently to that in the PDB file.

3.5. 1g7a

The total number of residues in this insulin structure is 204. The connection and extension procedures gave 216 residues distributed over eight chains. All residues and all side chains were built by the present method. In the PDB file, residues Thr30D, Phe1H and Thr30H are not present. Several parts of side chains are also missing. Otherwise, there is very good agreement between both models. The coordinate error of atoms in the PDB file is 0.12–0.13 Å. CMP1 and CMP2 are of comparable values. This supports the view that the model-building method presented in this paper can be regarded as a refinement method in electron-density space.

3.6. 1pen

Depending on the criteria for the chain extension, the number of connected P0 structure units was 16–18. This represents 17–19 residues (the first P0 represents two residues) and reflects the ability of the extension procedure to build the chain artificially into the lower electron density of (hydrogen-bonded) solvent molecules. This is a good property for building loops or disordered parts of the chain. On the other hand, false extensions and incorrect chain connections may hamper the process of sequence assignment. The conformation of Gly1 is incorrect. There is no side chain to fix the conformation in the post-sequence refinement. Both the CB and N atoms of the P0 SU have the same probability of hitting the N atom of the protein. Side chains have only minor conformation differences from the refined structure. The largest difference is 0.52 Å for Asn11 ND2.

3.7. 1rb9

Rubredoxin is another well behaved structure. Only Met1 CE is displaced by 0.9 Å from the closest refined position. Met1 is a disordered residue in the PDB file. Otherwise, the overall fit is almost perfect.

3.8. 2fdn

A principal problem in model building of this protein is disorder. A relatively small part of the structure was built by connection compared with other atomic resolution structures. The extension procedure is very important for the building of

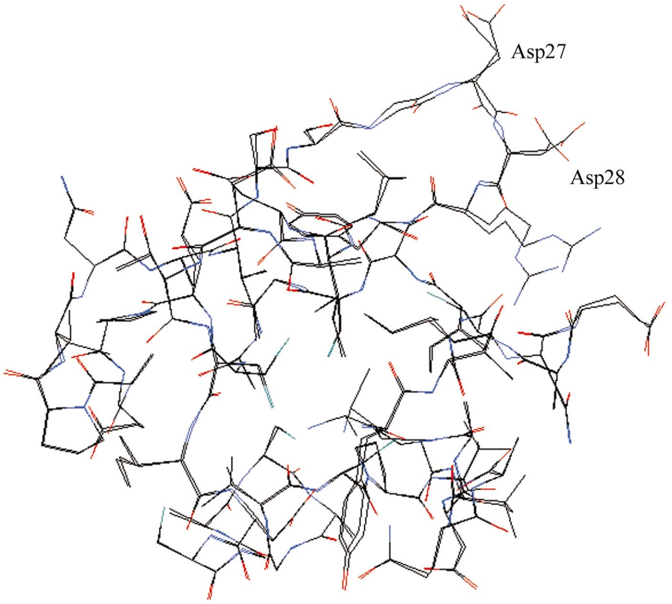

Figure 1
Comparison of the modelled structure and the structure of protein 2fdn. The difference in the main-chain building can be seen near the marked Asp28 and Asp29 residues.

this structure. No *cis*-peptides were found and the chain was constructed using the standard *trans*-peptide bond. The differences can be seen in Fig. 1. Depending on the fine details of the extension and connection procedures, various results were obtained. F_c fragments gave one polypeptide chain. It is interesting to note that the F_c-based model follows one disordered chain more closely, while the model from Dirac fragments follows the second disordered chain [not described in tables; see Pavelcik (2003) for definition of Dirac fragments]. There is a chance of building disordered parts of the protein automatically. Despite all this, the positions of the side chains are only slightly affected. There are disagreements with the published structure only in this loop (Asp28). Otherwise, there are no significant differences. Some conformations for rotamer side chains are slightly different, *e.g.* Tyr2 OH is displaced by 1.6 Å.

3.9. 9rnt

All 104 residues were built. Both *cis*-peptides were found with the help of CisPro0 *PROTF* peaks. There are few side chains (Lys25, Val33, Lys41, Ile61, Thr91, Asn98 and Val101) with different conformations for the marker and full search method. Rotamer side chains are of comparable quality.

3.10. GIBR

All 387 residues are built and connected. The *cis*-peptide bond was correctly located. There are very few disagreements in side-chain conformations. There is a difference in the position of the Asn1 side chain compared with *REFMAC* refinement (Borek, 2002).

3.11. TP47

This larger structure was used to demonstrate the overall applicability of the method. The structure was difficult to solve. At a resolution of 1.9 Å, one loop of the main chain was not found (Tomchick, 2001). Two protein molecules of 415 residues are present in the asymmetric unit. Calculations were carried out with phases calculated from the PDB file (PDB phases) and with experimental density-modified phases. The mean-square phase difference between these two sets is 63.5°. The results for connection and extension are given in Table 3. The structure is broken into many small chains; the sequence was not assigned correctly, so the data cannot be presented in Table 4.

3.11.1. PDB phases. 789 P0 structure units were found. This represents 806 protein residues. 17 chains resulted from connection instead of two. The largest connected chain consisted of 180 P0s and the smallest of six P0s. The second largest chain (140 residues) was correctly built and the sequence was aligned easily (signal-to-noise ratio 122/50). The largest chain has an insertion error. The left part of the chain (chain B, residues 40–141) was aligned correctly. Residue 142 was missing. The right end (143–219) is shifted by one residue and the wrong side chains were constructed. This sequence problem was reflected in the FOMs (65, 31, 37) for three successive offsets. Intermediate chains were also aligned correctly. Another problem was that some smaller isolated chains were interchanged among A and B chains and the single molecules were not constructed.

3.11.2. DM phases. 20 chains were constructed. The longest chain consists of 122 residues. There were two insertion errors in this chain, but the central part (sequence 59–115) of the chain was aligned correctly. Because of problems with the automatic sequence alignment, side chains were not built.

High-resolution structures were built completely. More residues were found by this method than had been published in the PDB files. These extra residues seem to have good stereochemistry, but there is no verification of the correctness unless these structures are redetermined. This is outside of the scope of this paper. The present method was not able to automatically build the 'whole protein' at a resolution of 2.1 Å because the main chain was not connected. More than 90% of the structure can be obtained at a resolution of 2.3 Å, provided the phases are good. In general, larger fragments are needed at resolution 2.3 Å for statistically grounded modelling. Moving to lower resolution with the present fragments is equivalent to overfitting.

The results of this paper can be compared to those in Table 1 of Perrakis *et al.* (1999). The phases refined by the *ARP*/*wARP* method are of comparable quality to the data presented here. The criteria for evaluation are similar: the number of residues found and the number of chains traced.

The influence of phase quality on the fragment positioning was studied by Pavelcik *et al.* (2002). Further analysis is presented here. Two structures at intermediate resolutions were analyzed. The mean-square phase differences are 54.5 and 59.8° for 1a75A and 1a75G, respectively, compared with 1a75. Analogous phase errors for 9rnt are 54.3 and 59.1°. In general, there is a reduction in model-building accuracy, as can be seen from the data in Table 4. Main-chain building with polyalanine phases has been improved, as phases are biased towards this structure, and the influence of fitting fragments into side chains is reduced. The phase error has little influence on the side-chain building in the 1.5 Å 9rnt structure. More problems were encountered at 1.9 Å 1a75. The signal-to-noise ratio was reduced significantly; one chain of 1a75G was not assigned correctly to a sequence (the CMP1 parameter cannot be calculated).

Nevertheless, the phase error is not the principal obstacle in model building. A partial model can be used in phase improvement and new building can be started with better phases. The procedure for phase improvement can be similar to recycling procedures used in small-molecule crystallography, but 'group atoms' should be used instead of normal atoms in order to obtain a reasonable observation-to-parameter ratio (Pavelcik, work in progress). For example, the 82% polyalanine structure of dmTP47 plus some marker side chains may be a good starting point for this improvement.

The structure representation by the generalized atoms (flexible fragments) is a promising approach to low-resolution crystallography. This is valid not only for model building, but also for structure refinement and density modification

(Terwilliger, 2001; Pavelcik, 2003; Kosik & Pavelcik, 2004). In addition, the method shows how chemical information can be introduced into the process of structure determination [see also Bricogne (1997) for a rigid-group 'blueprint' in macromolecular direct methods]. This concept may be the tool used to overcome the 1.2 Å barrier in *ab initio* structure determination.

The author is grateful to Drs D. Tomchick and D. Borek for supplying experimental data and to Dr. B. Schneider for reading the manuscript.

References

Borek, D. (2002). Private communication.

Bricogne, G. (1997). *Methods Enzymol.* **227**, 14–18.

Cowtan, K. D. (1998). *Acta Cryst.* D**54**, 740–756.

Cowtan, K. D. (2001). *Acta Cryst.* D**57**, 1435–1444.

Friedman, J. M. (1999). *Comput. Chem.* **23**, 9–23.

Greer, J. (1974). *J. Mol. Biol.* **84**, 279–301.

Greer, J. (1985). *Methods Enzymol.* **115**, 206–224.

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.

Kosik, G. & Pavelcik, F. (2004). *Mater. Struct. Chem. Biol. Phys. Technol.* **11**, 46.

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.

Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). *Proteins Struct. Funct. Genet.* **40**, 389–408.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 968–975.

Navaza, J. (2001). *Acta Cryst.* D**57**, 1367–1372.

Oldfield, T. (2002). *Acta Cryst.* D**58**, 487–493.

Oldfield, T. (2003). *Acta Cryst.* D**59**, 483–491.

Pavelcik, F. (2003). *Acta Cryst.* A**59**, 487–494.

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.

Zou, J. Y. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 833–841.