Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Performance of phased rotation, conformation and translation function: accurate protein model building with tripeptidic and tetrapeptidic fragments

František Pavelčík[a]* and Jiří Václavík[b]

[a]Department of Chemical Drugs, University of Veterinary and Pharmaceutical Sciences Brno, 612 42 Brno, Czech Republic, and [b]Department of Natural Drugs, Faculty of Pharmacy, University of Veterinary and Pharmaceutical Sciences Brno, 612 42 Brno, Czech Republic

Correspondence e-mail: pavelcikf@vfu.cz

Received 23 April 2010
Accepted 29 July 2010

The automatic building of protein structures with tripeptidic and tetrapeptidic fragments was investigated. The oligopeptidic conformers were positioned in the electron-density map by a phased rotation, conformation and translation function and refined by a real-space refinement. The number of successfully located fragments lay within the interval 75–95% depending on the resolution and phase quality. The overlaps of partially located fragments were analyzed. The correctly positioned fragments were connected into chains. Chains formed in this way were extended directly into the electron density and a sequence was assigned. In the initial stage of the model building the number of located fragments was between 60% and 95%, but this number could be increased by several cycles of reciprocal-space refinement and automatic model rebuilding. A nearly complete structure can be obtained on the condition that the resolution is reasonable. Computer graphics will only be needed for a final check and small corrections.

1. Introduction

Automatic structure determination presents a challenge in X-ray crystallography. Several methods have been developed and implemented in computer programs [*e.g. ESSENS* (Kleywegt & Jones, 1997), *ARP/wARP* (Perrakis *et al.*, 1999), *RESOLVE* (Terwilliger, 2003*a*) and *Buccaneer* (Cowtan, 2006)]. Recently, fragment-based methods have been reviewed by Cowtan (2008). None of these programs were able to reliably build complete structures (Joosten *et al.*, 2008) and computer graphics had to be used. This is in sharp contrast to small-molecule crystallography, in which the direct inspection of an electron-density map was abandoned many years ago in routine structure determination. New methods for the automatic interpretation of electron-density maps need to be developed.

The phased rotation, conformation and translation function (PRCTF) is a novel computational tool for positioning flexible molecular fragments into electron-density maps (Pavelcik, 2006*a*). The position, orientation and internal torsion angles of molecular fragments can be found in a spherical harmonics Bessel representation of the electron density. The PRCTF represents an improvement over the phased rotational and translation function (PROTF; Friedman, 1999; Pavelcik *et al.*, 2002) because rigid fragments are not well suited to macromolecular structure building. A small disadvantage of both functions is that the results are slightly dependent on the grid selection. For this reason, the generalized coordinates were refined by phased flexible refinement (PFR; Pavelcik, 2003).

A description of a crystal structure in terms of small flexible molecular fragments has been developed by Pavelcik (2003). This concept was applied (Pavelcik, 2003, 2004) to the building of protein structures at 1.2–2.4 Å resolution. A small set of carefully designed flexible structure units of about ten atoms containing one or two peptide groups was used. Modelling with monopeptidic and dipeptidic fragments requires an eight-dimensional search. Three parameters are used for position, three for orientation and two for conformation. Monopeptidic and dipeptidic fragments proved to be quite effective for the accurate and automatic building of high-resolution macromolecular structures. At resolutions of about 2.3 Å or lower the PROTF (or PRCTF) based on these fragments was less effective and about half of the structure had to be built by its extension (Pavelcik, 2004). Recently, the PRCTF has also been used to build RNA structures, including 30S and 50S ribosome subunits (Pavelcik, 2008; Pavelcik & Schneider, 2008). For this purpose, flexible mononucleotide and rigid double-helical fragments were used.

In general, larger fragments are needed at lower resolutions because of the reduced information content of the low-resolution electron-density map. Main-chain modelling with tripeptidic fragments requires four-dimensional conformation searches and six-dimensional searches are required for tetrapeptidic fragments. The fragment size cannot be increased beyond a certain limit because of the exponential increase in the combinations of conformations (Levinthal, 1968). Using larger fragments increases the total number of parameters that have to be determined for the entire protein structure. The number of parameters is $8N$ for dipeptides, $10N$ for tripeptides and $12N$ for tetrapeptides, where N represents the number of protein residues. Nonetheless, the positioning of the larger fragments is only formally independent. The positioning of neighbouring fragments is correlated. In other words, not every tripeptide needs to be located to build the complete chain. The limiting number of parameters is $6 + 2N$ for positioning the complete polyalanine molecule as one flexible fragment. From this number, and from the resolution dependence of the measured reflections, the upper resolution limit can be estimated for the flexible-fragment concept. The determinacy point for a crystal structure with 50% solvent content is 5.4 Å (Brunger *et al.*, 2009). The real resolution limit, assuming several observables per parameter, seems to be in the region of 3.5 Å. At lower resolutions the PROTF can be used for positioning rigid helices or small domains determined at higher resolution.

Full systematic conformational searches with oligopeptides are impractical (they currently take weeks or even months of CPU time and this situation will probably remain the same in the near future). Searches have to be limited to preferred conformations, known as conformation families. Conformation families of polypeptides at higher dimensions have been established (Pavelcik & Vanco, 2006; Pavelcik & Pavelcikova, 2007) by the direct mapping of multidimensional periodic torsion-angle space. This opened up a way to build macromolecular structures at lower resolutions using oligopeptides by the PRCTF. The number of conformation families is large, particularly for tetrapeptides. We were faced with the principal question ‘how many search conformers are needed in fragment-based model-building methods and how can we reduce this number?’ For these reasons, we also analyzed the overlap of partially correct fits to find conformers which were not used as search conformers.

In addition to the resolution and the number of search conformers, the performance of the PRCTF is also affected by the quality of the spherical harmonics Bessel expansion, the quality of the phases of the structure factors and the quality of the subsequent refinement. In principle, the number of search conformers could be further reduced if we were able to refine partially correct fits. In this paper, the PRCTF has been improved by a real-space refinement (RSR; Diamond, 1971; Chapman, 1995).

Tripeptidic and tetrapeptidic fragments were studied and tested at various resolutions in an effort to find the best possible strategy for automatic model building. One of the aims of this paper was to establish optimal parameters for the PRCTF. The result of the PRCTF depends on many parameters. For example, a finer translation grid improves the performance of both the rotation and translation functions, but cubically increases the CPU time. We are seeking cost-effective modelling, optimal conditions and the best possible performance.

The overall performance of the PRCTF can be evaluated indirectly and independently by its ability to build a complete structure model.

2. Methods

Building with oligomeric fragments is more complicated than building with monomeric fragments. The oligomeric fragments can be defined in several ways, *e.g.* by amino-acid residues or by peptide groups. We prefer the peptide-group concept because five atoms of a peptide group form a planar rigid body and these fragments are the largest fragments with fixed pairs of (φ, ψ) torsion angles. The number of possible conformations increases exponentially with the number of (φ, ψ) pairs. This fact discriminates larger fragments. In this study, we selected fragments with three or four peptide groups as the basic model-building blocks and these search fragments were named AlphaT and AlphaQ. Oligomeric fragments located in the electron density have a more complicated fragment overlap than monomeric fragments. For instance, a fragment with three peptide groups can overlap with another located fragment at two peptide groups, one peptide group or even one CA atom only. In addition to this, there are many overlaps involving side chains at the position of peptide groups and *vice versa*.

The process of model building with oligopeptides can be divided into ten main steps.

(i) Flexible conformers of search fragments are localized in the electron-density maps by the PRCTF and refined by a real-space refinement. The number of search conformers is limited to the most probable ones.

(ii) The possible solutions (peaks) of the PRCTF are tested for partial overlap; new building blocks are generated and refined.

(iii) Localized and generated fragments are analyzed for mutual overlaps and arranged into chains of positioned fragments. The algorithm aims to dock the oligomeric fragments centred on each peptide group/residue; the fragment assembly then searches for fragments that overlap ideally by $(n-1)$ of n peptide groups.

(iv) The polypeptidic chains are extended at both ends by direct building into electron density with extensive sets of tripeptidic or tetrapeptidic conformers.

(v) Sequence docking.

(vi) Side chains are built based on the results of sequence docking.

(vii) Post-sequence corrections such as the elimination of overfitted residues, the building of small loops or a search for *cis*-peptide bonds.

(viii) The fragment model is converted into an atomic model and the PDB file is generated.

(ix) The model is checked and refined with the *REFMAC*5 program.

(x) The refined partial model is entered into the model-rebuilding process. This step can be coupled with phase improvement based on information from the partial model.

2.1. Molecular fragment and conformer tables

2.1.1. Tripeptidic and tetrapeptidic fragments. The line formula of a tripeptidic AlphaT fragment is $\mathrm{C}^{\alpha}-\mathrm{C}(=\mathrm{O})-\mathrm{Ala}-\mathrm{Ala}-\mathrm{N}-\mathrm{C}^{\alpha}$. This fragment contains three peptide groups, two pairs of (φ, ψ) torsion angles, four CA atoms, two CB atoms and is approximately of tripeptide size. Based on previous statistics (Pavelcik & Pavelcikova, 2007), the number of conformation families seems to be reasonable and the errors introduced by the assumption of fixed bond lengths and fixed bond angles do not appear to be large. The fragment was designed to be a compromise between size and number of conformers. The line formula of a tetrapeptidic AlphaQ fragment is $\mathrm{C}^{\alpha}-\mathrm{C}(=\mathrm{O})-\mathrm{Ala}-\mathrm{Ala}-\mathrm{Ala}-\mathrm{N}-\mathrm{C}^{\alpha}$. This fragment contains four peptide groups, three pairs of (φ, ψ) torsion angles, five CA atoms and three CB atoms. The number of conformation families exceeds 100. Only the most frequent conformations can be used in calculating the PRCTF. This fragment was designed for lower resolutions or maps of lower quality to locate the dominant part of the structure. We use the terms 'tripeptidic' or 'tetrapeptidic' fragment to distinguish these fragments from tripeptide or tetrapeptide fragments. The molecular fragments used for model building are shown in Fig. 1.

The flexibility of the fragments is defined in the PRCTF by conformer tables. The conformer table represents an irregular grid in the multidimensional fragment search. Each flexible torsion represents one search dimension. The recent data used for calculations and stored in the *NUT* program are presented in Tables 2 and 3 of Pavelcik & Pavelcikova (2007); the number of conformers in the program is limited to 16.

Table 1
Idealized and experimental radii for the AlphaT and AlphaQ fragments.

R_{mean}, R_{min} and R_{max} are related to experimental conformations. N is the number of experimental fragments. α_{R}, $\beta1$ and $\beta2$ are idealized conformations. Radii are given in Å.

	R_{mean}	R_{min}	R_{max}	N	α_{R}	$\beta1$	$\beta2$
AlphaT	4.4 (6)	3.31	6.11	375204	3.69	5.21	4.58
AlphaQ	5.5 (1.0)	3.89	7.81	335820	4.22	6.61	5.30

2.1.2. Fragment radii and virtual distances. The shape and size of the flexible fragment depends on the conformation. One of the simplest descriptors of a fragment is the fragment radius. The fragment radius is needed to determine the radius (R_{ex}) of the expansion sphere in the PRCTF. A smaller radius means a better description of the electron density with the same number of expansion coefficients and more accurate positioning. For this reason, the fragment radius represents the critical parameter for the PRCTF. Experimental fragment radii were analyzed for crystal structures from a protein database, using the same set as were used for analysis of the conformation families (Pavelcik & Pavelcikova, 2007). The theoretical α_{R} (φ, ψ) = (−64°, −41°), $\beta1$ (φ, ψ) = (−121°, 128°) and $\beta2$ (φ, ψ) = (−66°, 136°) represent typical secondary structures. These structures were modelled with torsion angles

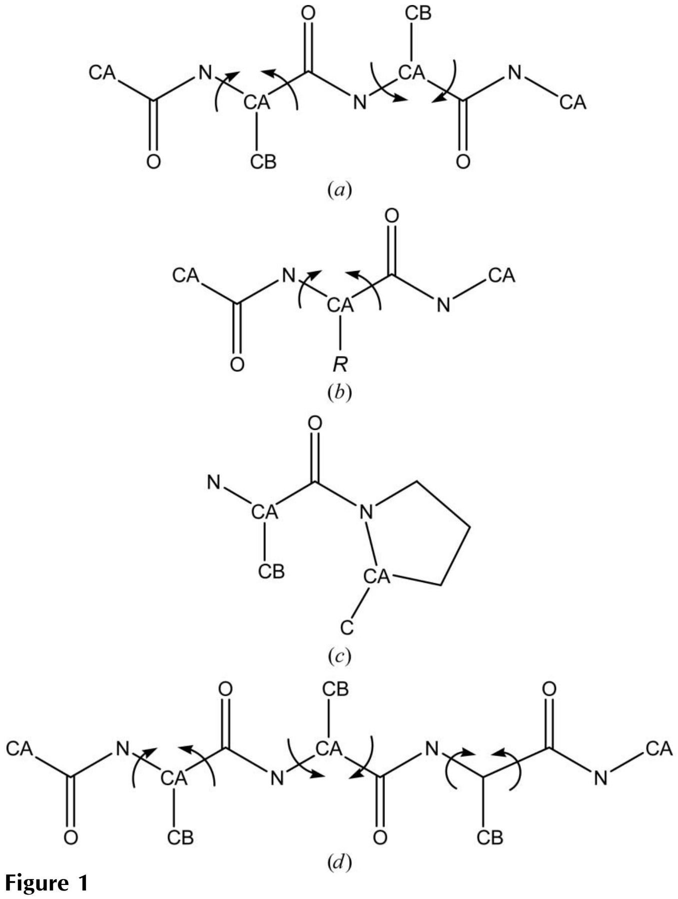

Figure 1
The fragments used for protein model building. (*a*) Main-chain AlphaT fragment. (*b*) Side-chain SideD fragment. (*c*) CisPro fragment. (*d*) AlphaQ fragment.

Table 2
Idealized and experimental virtual bonds (VB) and virtual 1–3 distances (VT) for the AlphaT and AlphaQ fragments.

Mean, min and max are related to experimental conformations. α_R, $\beta1$ and $\beta2$ are idealized conformations. The number of virtual bonds and VT distances is 300 825 and 269 580 for AlphaT and AlphaQ, respectively. Distances are in Å.

	VB			VT			VB			VT		
	Mean	Min	Max	Mean	Min	Max	α_R	$\beta1$	$\beta2$	α_R	$\beta1$	$\beta2$
AlphaT	2.4 (7)	1.33	3.82	4.7 (1.4)	2.59	7.62	1.68	3.24	2.48	3.20	6.44	4.66
AlphaQ	2.2 (7)	0.99	3.73	4.4 (1.4)	2.12	7.45	1.56	3.20	2.14	3.1	6.4	4.2

from Hovmöller *et al.* (2002). The radii are presented in Table 1.

Virtual parameters (virtual bonds and virtual angles) are useful for model building because they impose limits on the possible inter-fragment distances. Within the flexible-fragment concept, simple geometrical centres of fragments defined around each residue generate a set of points for the calculation of virtual parameters. A virtual bond (VB) is the distance between the geometrical centres of two fragments shifted by one residue on a polypeptide chain. A virtual angle between three successive fragments is represented by the distance (VT) between residue 1 and residue 3. Virtual parameters were harvested from the same set of structures as the fragment radii. The aim of the analysis was to determine critical values for fragment connection. The calculated values for AlphaT and AlphaQ fragments are given in Table 2. The conformer tables and virtual bond parameters represent the knowledge base for model building.

2.1.3. Other fragments. The fragment CisPro, described by Pavelcik (2003), was designed for the detection of *cis*-peptide bonds. The PRCTF was used to determine the position of the fragment in the electron density.

20 dipeptidic fragments were modelled for the final model representation, sequence alignment and creation of a PDB file for the protein model. These are of variable size and can be represented by the line formula $\mathrm{C}^{\alpha}{-}\mathrm{C}({=}\mathrm{O}){-}\mathrm{N}{-}\mathrm{C}^{\alpha}(R){-}\mathrm{C}({=}\mathrm{O}){-}\mathrm{N}{-}\mathrm{C}^{\alpha}$. On average, the size of these fragments is comparable to the size of the tripeptidic fragment. These fragments contain two planar peptidic groups. A pair of main-chain torsion angles (φ, ψ) and side-chain torsion angles (χ) are the variables in searches. The name of each individual fragment consists of the three-letter code for the amino acid and the letter D, representing two peptidic groups (*e.g.* HisD). The collective name for these fragments is SideD. A fragment AlphaD (AlphaA0; Pavelcik, 2003) is used in certain steps of sequence assignment and model extension.

2.2. Phased rotation conformation and translation function

The PRCTF was used as described and implemented by Pavelcik (2006*a*). The result of the PRCTF is a set of positioned and oriented fragments. Positioned fragments are called peaks [the peak of the translation map of the PRCTF]. The peaks accepted into a protein model and arranged into chains are called localized fragments or virtual atoms.

2.3. Real-space refinement (RSR)

The PRCTF peaks can be refined using a method developed by Chapman (Chapman, 1995; Korostelev *et al.*, 2002; see also Diamond, 1971). The only difference from the Chapman method is that individual atoms are not refined, but the position, orientation and conformation of a whole localized fragment are. Instead of restrained refinement, we have programmed constrained refinement. Otherwise, the calculation of derivatives and the conditions for numerical integration remain the same. Both resolution-dependent atomic shapes (Chapman shapes) and spherical Gaussian shapes can be used. The former give us more realistic temperature factors that are comparable to a reciprocal-space refinement. The greatest advantage of RSR is that the scale factor between the calculated and experimental electron densities can be determined and a real-space *R* factor can be calculated. This *R* factor seems to be more sensitive than the correlation coefficient that we have used to date (Pavelcik, 2004) to discriminate between correct and false peaks. The *R* factor can also be used for sequence assignment.

2.4. Model building

2.4.1. Partial overlap. In an ideal situation for model building, two located tripeptidic fragments overlap at two peptide groups (three CA atoms; Fig. 2*a*). Fragments shifted by more peptide groups are regarded as partial overlaps (Figs. 2*b* and 2*c*). The information collected from analysis of these overlaps can be used to construct new 'peaks'. Two partially overlapped original peaks of the PRCTF are trans-

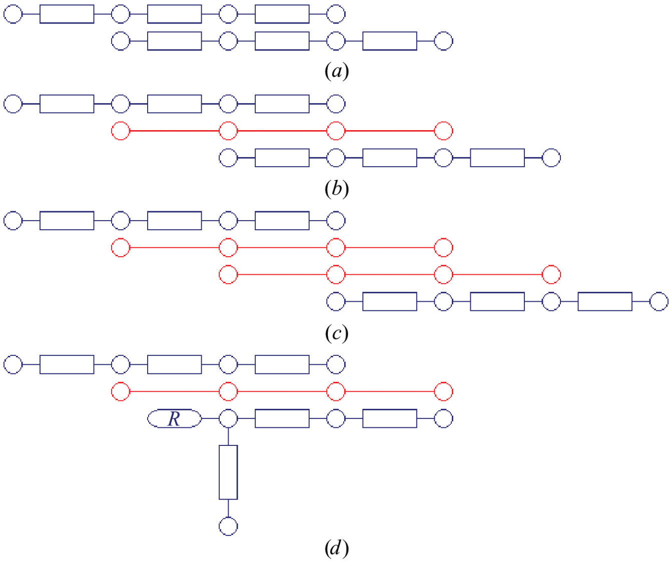

Figure 2
Partial overlaps. Rectangles represent planar peptide groups and circles represent CA atoms. (*a*) Normal three-point overlap of two AlphaT fragments. This type of overlap is used for fragment connecting. (*b*) Partial two-point overlap. One new fragment is constructed (in red). (*c*) Partial one-point overlap. Two new fragments are constructed. (*d*) Partial overlap involving side chain. One new fragment can be constructed.

formed into Cartesian atomic coordinates. The atoms of these two peaks are combined together to create a new peak. The bond lengths and bond angles are fixed. The torsion angles of this new peak are refined against the electron density and if the correlation coefficient is good then the peak is accepted. In this way, partially overlapped peaks can create new building blocks with unique conformations. The entire set of peaks enters the connecting process.

The peptide group of the search fragment can be fitted either to the electron density of the main-chain peptide group or to a side chain of similar stereochemistry. Also, the side chain (Asn, Asp) can be fitted to main-chain atoms (Fig. 2*d*). These partially correct fits can have a good correlation with the electron density but are difficult to identify and have a tendency to disrupt protein model building.

An analogous analysis of fragment overlaps has also been programmed for tetrapeptidic fragments.

2.4.2. Connecting fragments. The successful connection of the positioned fragments into the chain is a practical criterion for evaluating the performance of the PRCTF. The connection of fragments has been described in previous papers (Pavelcik, 2003, 2004) and the connecting algorithm is implemented in the *HEL* program (Pavelcik, 2006*b*). Only a marginally modified algorithm is used in the most recent versions of our computer programs.

From the perspective of recent developments (DiMaio *et al.*, 2006, 2007), the connecting algorithm can be regarded as a very simplified version of a graph-based probabilistic approach to protein backbone tracing. The correlation coefficient of the fragment and the crystal electron density represent a vertex potential. The edge potential is given by a restriction on the virtual distance: edge probability 0 for distances longer than the cutoff distance and a fragment overlap. The AlphaT or AlphaQ fragment is larger than a monomer and two AlphaT or AlphaQ fragments are overlapped at all atoms of two or three peptide groups. This overlap is quite discriminative and leads to edge probabilities close to either 1 or 0. A few branches are resolved using a connectivity index (Morgan, 1965) and the VT distances (Table 2). A single pass through the connectivity graph connects the positioned fragments, making the algorithm practically deterministic.

The refined peaks of the PRCTF are pre-sorted on the basis of peak height and peak connectivity (equation 2 in Pavelcik, 2004). Connections are processed in two steps. In the first step, only chains with peaks that have all peptide groups overlapped are accepted. Many smaller chains are created (seeds). These seeds usually correspond to rigid parts of the macromolecular structure and are the only seeds that are used for further chain building. In the second step, the requirement for perfect overlap is dropped and all peaks with correct CA—CA distances are connected. Chains are represented by series of oligomeric virtual atoms. The protein model generated in this way is called a virtual model.

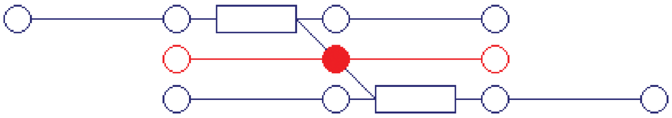

Figure 3
Transformation of the AlphaT model to the AlphaD model. The central parts of two positioned AlphaT fragments are used to create AlphaD fragments (in red). The coordinates of two CA atoms are averaged to create the central CA atom of the AlphaD fragment (filled red circle).

2.4.3. Extension of chains. The chains of the virtual model are extended at both ends with tripeptidic or tetrapeptidic fragments. Reference tables of preferred tripeptidic and tetrapeptidic conformations were generated for this purpose. The values in the tables are those for the conformation families (Pavelcik & Pavelcikova, 2007). Some large conformation families, such as those belonging to the β-region, are further split into several conformations. Each conformation is built and refined and the best conformer is accepted. The extension fragment is anchored either on the last peptide group or on the central peptide group of the neighbouring virtual atom. Several new building blocks are constructed in this way and the combined set of originally connected virtual atoms and new localized fragments are entered into the connecting procedure. In this way, larger chains are built. These new chains are then checked for partial overlaps and new virtual atoms are constructed and mixed with unused PRCTF peaks and again entered into the connection. The procedure is repeated several times. The process is finished when the number of connected fragments is comparable to the number of protein residues or if no new virtual atom with a satisfactory correlation with the electron density is found. Usually, several of the smaller chains formed end with a virtual atom having an incorrect conformation, at a *cis*-peptide, at a disordered chain or at a residue that exhibits a very large displacement parameter in a fully refined structure.

2.4.4. Conversion of AlphaT fragments. Each AlphaT virtual atom can be transformed into two dipeptidic fragments, but for a chain of connected virtual atoms we use the scheme depicted in Fig. 3. Centrally, the most reliable peptide groups of two AlphaT fragments are combined together.

The structure is regularized in this way. Two additional but less reliable fragments are constructed at each end of the chain. The AlphaD virtual atoms are connected into new chains; some tripeptidic building errors can be removed at this stage. These AlphaD virtual atoms are converted into 20 SideD localized fragments for sequence assignment and then refined.

If the resolution and phase quality are good, the AlphaD fragments can be extended by a systematic search (Pavelcik, 2004). Instead of using the table of preferred conformations, the two main-chain torsion angles are systematically varied with a step of 10°. The two conformers with the best fits to electron density are stored and entered into the connecting process, making the extension a multi-solution process. Very rare conformations can also be modelled using this method. The number of separated chains is then usually further reduced. In favourable cases, the number of chains is equal to the number of protein chains in the asymmetric unit and the number of AlphaD virtual atoms is equal to the number of residues. Occasionally, such an extension has a tendency to overfit and the chains are extended beyond the protein borders. CisPro markers can also be entered into the connection process and the resulting AlphaD virtual atom containing information on the *cis*-proline peptide bonds can be constructed.

2.4.5. Treatment of AlphaQ fragments. Localized AlphaQ fragments are not converted into SideD virtual atoms like AlphaT fragments because they were designed to interpret electron-density maps of lower quality. SideD fragments are not stable when refined into low-quality electron-density maps. Instead, the partially overlapped fragments are regularized. The external peptide groups of one fragment are replaced by the central peptide groups of neighbouring fragments in a similar process to that shown in Fig. 3. The rebuilt AlphaQ virtual atoms are then refined into the electron density. The result of the regularization is that all of the connected virtual atoms are overlapped at three peptide groups. Finally, the central part of the localized fragment is converted into 20 SideD virtual atoms for sequence assignment. These SideD blocks are not refined.

2.4.6. Sequence docking. We mutate each SideD virtual atom into each of the 20 different amino acids. For each amino acid the fit of each rotamer is optimized to the electron density. Rotamer tables (Lovell *et al.*, 2000) are used for each fragment type. Small side chains (χ_1 or χ_1 and χ_2) are optimized by systematic searches with a torsion-angle step of 10°. Systematic searches are useful for locating S atoms in high-resolution structures. Distance tests are calculated and the conformations that interfere with the main-chain atoms of the same fragment are discarded. The real-space fit of the best-fitting rotamer is then used as a SCORE to determine how well that particular amino-acid type fits the density. The quality of the fit is evaluated by a correlation coefficient (CC) and a real-space R factor (RF). The empirical SCORE is defined as

$$\mathrm{SCORE} = \mathrm{CC} - 0.5\mathrm{RF}. \qquad (1)$$

To improve the sensitivity of small fragments such as GlyD, AlaD or SerD that fit well into any larger side chain, a set of dummy atoms is added to the side-chain atoms. The dummy atoms are given positions where other amino acids have atoms. A large electron density for these dummy atoms represents a penalty function that is added to the SCORE. Dummy atoms are also added to intermediate large chains to distinguish between similar residues such as Asn or His. A SCORE matrix with dimensions $N \times 20$ is generated and normalized (as a Z score), with N being the number of located SideD virtual atoms. A second matrix is calculated from positioned markers (Pavelcik, 2003; Hattne & Lamzin, 2008). The restricted conformation (φ torsion angle) of proline can be utilized in the same way as the markers.

Chains longer than 60 residues need to be split into smaller chains at the weakest connection point in order to avoid 'out-of-register' errors. The technique for sequence assignment has been described in Pavelcik (2004), but the algorithm is quite similar to the sequence docking described, for example, by Zou & Jones (1996), Terwilliger (2003*b*), Cowtan (2008) and others. A chain of residues is shifted along the sequence and the Z score from the matrix for each amino acid given by the sequence is added into a final combined Z score for each particular chain position. The chain position with the maximum combined Z score is recorded. All chains are analyzed in this way. Chains with the highest signal-to-noise ratio (usually the longest chains) are docked first. The sequence assignment for smaller chains is then repeated only in the sequence gaps left by positioned larger chains. A large overlap with an already assigned chain is treated as a penalty; small overlaps are tolerated.

If the sequence is not reliably assigned owing to a low signal-to-noise ratio then the result of the model building is a model with side chains that fit best into the electron density.

2.4.7. Post-sequence modelling. When the sequence is aligned and the boundaries of the chains are known, overfitted virtual atoms can be removed. Chains that are overlapped by sequence but unconnected by overlap criteria can be corrected and connected. Finally, the side chains are rebuilt according to the sequence. The overlap of CisPro fragments located by the PRCTF with the residues of the model is analyzed, particularly at disconnected sites, and *cis*-peptide bonds are built. Development of the post-sequence stage of model building is not yet finished. Procedures for disordered chains, very flexible loops, very long or modified side chains, heterogroups and solvent need to be further elaborated and will be reported later.

2.5. *REFMAC* recycling

2.5.1. *REFMAC* refinement. A model transformed from generalized coordinates into atomic coordinates is exported as a PDB file and this file can be refined using the *REFMAC*5 program (Murshudov *et al.*, 1997). The *REFMAC*-refined model is analyzed and dubious residues can be deleted such as those with very high displacement parameters or incorrect connections. Several corrected models can be refined and the final selection can be based on $R_{\rm free}$. The refined model is transformed back into AlphaQ, AlphaT or SideD virtual atoms and returned to the model-building stage. The actual strategy depends on the results of the sequence alignment, refinement, quality of phases and the completeness of the model.

In favourable cases, one can try to locate only a few missing residues in a new electron-density map (based on the FP/PHIC or FWT/PHWT coefficients produced by *REFMAC*). In many intermediate steps, the existing model is combined with the original or newly calculated PRCTF peaks, extended, the sequence is reassigned and the rebuilt model is again refined by *REFMAC*. Several cycles of this process can be carried out with minimal user intervention. A few selected cases of *REFMAC* recycling will be discussed later.

2.5.2. Phase combination. In the initial stages of *REFMAC* recycling one can combine the original (usually *DM*) and *REFMAC*-calculated phases and start model building into a better map. We use a simple scheme for the phase mixing. All

Table 3
Test macromolecular structures.

Code is a PDB code or a structure code used in structure determination. RES is the resolution. N_{REFL} is the number of reflections. N_{SEQ} is the number of residues in the asymmetric unit. N_{PDB} is the number of residues in the PDB file of the refined structure. N_{CHAIN} is the number of protein chains. MPE is the r.m.s. phase differences for experimental and final refined phases.

Code	RES (Å)	Space group	N_{REFL}	N_{SEQ}	N_{PDB}	N_{CHAIN}	Reference†	MPE (°)
1ab1	0.9	$P2_1$	28737	46	46	1	1	—
1g7a	1.2	$R3$	53700	204	201	8	2	—
9rnt	1.5	$P2_12_12_1$	13115	104	104	1	3	—
1a70	1.7	$P2_12_12_1$	12461	97	97	1	4	—
1a32	2.0	$P2_12_12_1$	6149	88	85	1	5	—
12gs	2.1	$C2$	25527	420	418	2	6	—
1bfe	2.3	$P4_132$	6387	119	110	1	7	—
1zx3	2.5	$I4_122$	4685	122	86	1	8	—
2b17	2.7	$P4_3$	3288	121	121	1	9	—
2gpi_mad	1.6	$C222_1$	15540	91	91	1	JCSG	80.4
RNASE_dm	1.8	$P2_12_12_1$	17211	192	192	2	D	75.3
RNASE	1.8	$P2_12_12_1$	17211	192	192	2	10	—
1vku_mad	2.0	$P3_221$	6572	100	85	1	JCSG	80.7
TP47_dm	2.3	$P3_221$	66966	830	801	2	T	63.5
1vkn_dm	2.5	$P2_1$	52490	1360	1351	4	JCSG	51.7
2re7_dm	2.5	$I4_122$	8859	133	133	1	JCSG	59.5
3cuc_dm	2.7	$P3_121$	36748	582	530	2	JCSG	61.2
3cuc	2.7	$P3_121$	36748	582	530	2	JCSG	—
3dxo_dm	2.7	$P6_122$	11371	242	235	2	JCSG	63.8

† 1, Yamano *et al.* (1997); 2, Smith *et al.* (2001); 3, Martinez-Oyanedel *et al.* (1991); 4, Binda *et al.* (1998); 5, Clemons *et al.* (1998); 6, Oakley *et al.* (1999); 7, Doyle *et al.* (1996); 8, J. Osipiuk, M. Cuff, X. Xu, A. Savchenko, A. Edwards & A. Joachimiak, unpublished work; 9, Singh *et al.* (2006); 10, Sevcik *et al.* (1991); D, Dodson (2004); T, Tomchick (2001), *DM* data at 2.3 Å resolution, current PDB code is 1o75 (Deka *et al.*, 2002); JCSG, Joint Centre for Structure Genomics (unpublished work).

dubious residues or long side chains of the model are manually deleted. A mask is created, *i.e.* all grid points of the map that are within a radius RM of known atoms (*e.g.* RM = 2 Å) of the model are selected. Two maps are calculated: the first is based on the original (*DM*) phases and the second is based on the refined phases. A new map is created as a combination of the *DM* map and the *REFMAC* map. Masked grid points are given *REFMAC* electron density; the rest of the pixels are from averaged *DM* and *REFMAC* density. Special care may be given to negative densities or very high densities belonging to heavy atoms of heterogroups. Back FFT of the combined map produces new phases. Many different variants of this simple scheme can be developed for phase calculation.

3. Calculations, results and discussion

3.1. Test structures

The PRCTF was tested on several randomly selected structures from the JCSG database and from the PDB database with deposited structure factors. We also included some older tests (Pavelcik, 2004) for comparison. Different space groups and a broad range of resolutions were selected for testing purposes. The basic crystallographic data are presented in Table 3. We used phases calculated from the deposited coordinates, phases archived in structure-factor files, experimental *DM* phases and, in two cases, MAD phases. Observed structure factors were used in all calculations. Experimental phases are specified by an extension added to the structure code.

Test structures can be divided into two main groups: structures with calculated phases and various resolutions and structures with various resolutions and experimental phases.

The first group is used to reveal the limitations of the method, represented by its ability to build the model into 'perfect' data. This group should provide answers to the following very basic questions. How large a fragment should be used for a particular resolution? How many search conformers are required in fragment-based model-building methods? To what extent can the analysis of overlapped PRCTF peaks reduce the number of search conformers? What are optimal parameters for calculating the PRCTF?

The second group simulates a real process of structure determination. The RNASE_dm data set was phased by density modification based on *MLPHARE* data of Hg and Pt derivatives (Dodson, 2004), while the RNASE data set has phases calculated from PDB file 2sar. The TP47_dm data set has phases from density modification of the SAD phase set phased with Xe (Tomchick, 2001). 2gpi, 1vku, 1vkn, 2re7, 3cuc and 3dxo are data sets obtained from Joint Centre for Structure Genomics (JCSG). These structures were solved using the MAD technique and data were taken from either the DM or PHASING directories and used without any changes. Related fully refined data originated directly from the PDB. The electron-density maps were calculated using the full resolution range as specified in the mtz files with the exception of 1vkn, where more complete data are available and density-modified data at a resolution of 2.45 Å were used.

3.2. Optimization of the PRCTF parameters

The performance of the PRCTF depends essentially on the following factors: the grid of the translation function (GrdZ), the number of conformers used for searching (ncon), the radius of the electron-density expansion (R_{ex}), the number of spherical harmonics and spherical Bessel functions ($l_{\max}$ and $n_{\max}$), the method for calculating fragment expansion (F_{c}, E_{c} or D_{c}; Pavelcik, 2003), the grid for Euler angles of the rotation function, the target function and, finally, the mask, which designates the search area in the unit cell.

We chose structure 1bfe for optimization of several of these parameters for AlphaT fragments. Based on previous experience, the starting values for the optimization were selected as follows: GrdZ = 0.5 Å, $n_{\max}$ = 5, $l_{\max}$ = 7, R_{ex} = 5.0 Å. R_{ex} does not need to be equal to $R_{\max}$. Even if the fragment is larger than the expansion radius the rotation function can still give satisfactory results because the section of the fragment outside the expansion sphere does not contribute to the target function.

The performance of the PRCTF is evaluated by the number of correctly localized peaks and the accuracy of positioning. Atoms belonging to the refined peak are compared with the atomic coordinates in the original PDB file. An AlphaT peak is considered to be positioned correctly if the r.m.s.d. (root-mean-square distance) is smaller than 1.0 Å and the sequence numbers of residues are in the correct order. The calculation of the r.m.s.d. is performed using the formula

$$\text{r.m.s.d} = \left(\frac{\sum d^2}{n}\right)^{1/2} \qquad (2)$$

for all fragment atoms (n = 15), where d is the distance between two related atoms. The criterion is rather strict because all partial fits, such as a fit into two peptide groups and one side chain, are not considered. The 1.0 Å limit may be too low for low-resolution structures with less accurate phases. The second condition, the correct sequence order, is particularly important in order to avoid fragment fits going in the opposite direction from the C-terminus to the N-terminus.

The aim of the optimization is not to find the maximum number of residues at any cost, but rather to find a reasonable compromise with the calculation time. The principal results of the parameter optimization of the PRCTF are shown in Figs. 4, 5 and 6.

The most important parameter is the number of conformers. The computer time is linearly proportional to ncon. The number of residues in the PDB file for 1bfe is 110 and the maximum number found by the PRCTF was 105. The performance of the PRCTF is proportional to the number of conformers used, but beyond ncon = 7 changes are at the level of few percent (Fig. 4). Unsurprisingly, this is consistent with the statistics of tripeptide conformation families. ncon seems to have an optimum at 7.

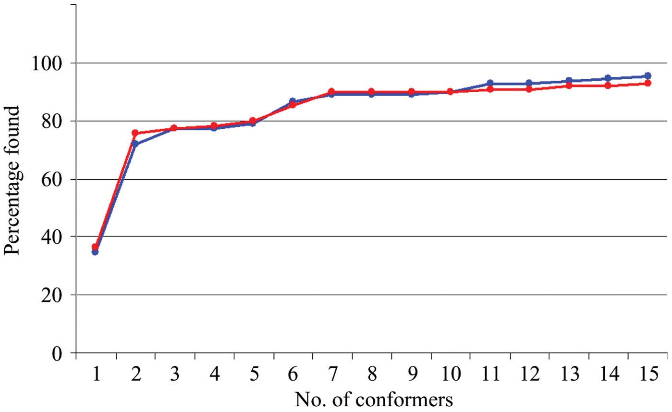

Figure 4
Optimization of the PRCTF parameters on structure 1bfe. Performance is shown as a function of the number of conformers. F_c fragments. R_{ex} = 5.0 Å, GrdZ = 0.5. Red line, program settings n_{max} = 5, l_{max} = 7. Blue line, program settings n_{max} = 6, l_{max} = 8.

As can be seen in Fig. 5, the most effective calculation in terms of CPU time appears to be the use of only two conformers. On the other hand, using the maximum number of conformers takes only several minutes of CPU time on a reasonably powerful personal computer. In general, all of the residues are not usually found by graphic programs at these resolutions. Thus, all of these residues can rarely be found by automatic programs because not all of the conformation families are used in the search. It is supposed that the few remaining residues would be modelled directly into the electron-density map either by human or other computer methods.

Another important factor for the CPU time is the grid. The computation time is inversely proportional to the third power of GrdZ. A finer grid increases the quality of the performance of both the rotation and translation parts of the PRCTF. The results are presented in Fig. 3. Suitable values seem to gather around 0.5–0.8 Å.

The number of expansion coefficients, controlled by n_{max} and l_{max} (Pavelcik *et al.*, 2002) directly influences the quality of the electron-density expression in the expansion sphere. This has an impact on the rotation. We found only a limited impact on the calculated results (Fig. 4). The current default in the computer program (n_{max}, l_{max}) = (5, 7) derived for monopeptides and dipeptides probably represents a sufficient number of coefficients [$n_{max}(l_{max} + 1)(l_{max} + 1)$] and also seems to be acceptable for tripeptidic fragments. An increase in these parameters has a limited impact on the ability of the PRCTF to find the correct peaks but does have some influence on the accuracy of the positioning. The CPU time is proportional to the number of expansion coefficients (Fig. 5). In this case, the optimum values seem to be roughly (n_{max}, l_{max}) = (6, 8). Another parameter is the radius for the electron-density expansion. The radius R_{ex} = 4.5 Å is close to the mean value of the fragment radius in the real crystal structures (see Table 1)

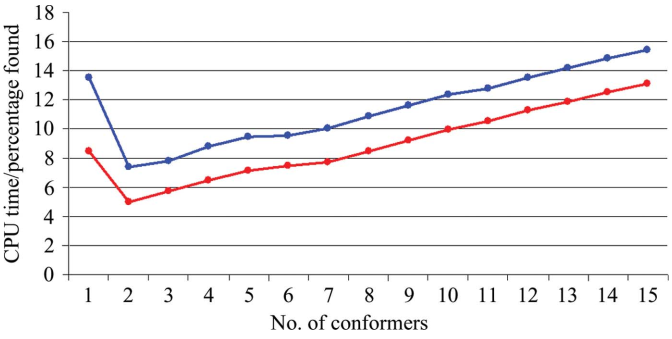

Figure 5
Optimization of the PRCTF parameters on structure 1bfe. CPU time/performance is shown as a function of the number of conformers. F_c fragments. R_{ex} = 5.0 Å, GrdZ = 0.5. The variable is the number of conformers. CPU time is on an arbitrary scale. Red line, program settings n_{max} = 5, l_{max} = 7. Blue line, program settings n_{max} = 6, l_{max} = 8.

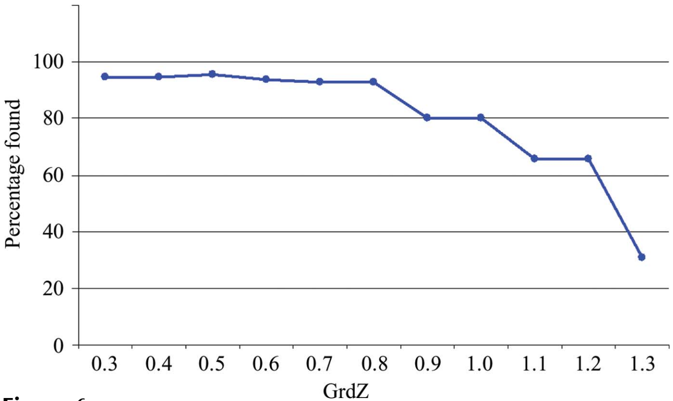

Figure 6
Optimization of the PRCTF parameters on structure 1bfe. Performance is shown as a function of grid (GrdZ). F_c fragments. n_{max} = 6, l_{max} = 8, R_{ex} = 5.0 Å.

Table 4
Details of the *NUT*, *HEL* and *BED* calculations based on AlphaT.

F_c fragments. 15 conformers were used. $N_{\rm NUT}$ is the number of residues correctly located by the PRCTF. % is the percentage (100% is $N_{\rm PDB}$). ⟨r.m.s.d.⟩ is the mean r.m.s.d. (see text). $N_{\rm CNH}$ is the number of connected AlphaT peaks divided by the number of chains of connected fragments in *HEL*. $N_{\rm HEL}$ is the number of amino-acid residues correctly located by *HEL*. $N_{\rm CNB}$ is the initial number of connected fragments in *BED* comparable to $N_{\rm HEL}$. $N_{\rm BED}$ is the final number of correctly located residues and $N_{\rm SEQ}$ is the number of located residues with the correct sequence.

Code	$N_{\rm NUT}$	%	⟨r.m.s.d.⟩ (Å)	$N_{\rm CNH}$	$N_{\rm HEL}$	$N_{\rm CNB}$	$N_{\rm BED}$	$N_{\rm SEQ}$
1ab1	46	100.0	0.14	44/2	46	45/1	46	46
1g7a	187	93.0	0.17	171/15	175	177/12	196	196
9rnt	99	95.2	0.18	88/9	88	94/7	98	98
1a70	89	91.8	0.19	82/8	84	84/7	97	97
1a32	81	95.3	0.15	83/3	79	82/2	82	82
12gs	399	95.0	0.17	376/18	374	380/11	415	415
1bfe	103	93.6	0.28	84/7	82	94/8	104	104
1zx3	82	95.3	0.35	62/7	60	76/4	81	77
2b17	94	77.7	0.43	45/11	54	68/11	101	7
2gpi_mad	73	80.2	0.32	64/7	64	67/6	78	78
RNASE_dm	173	90.1	0.43	159/15	146	170/12	176	14
RNASE	189	98.4	0.21	185/8	184	187/6	187	185
1vku_mad	74	87.1	0.38	65/6	63	72/4	76	72
TP47_dm	643	80.3	0.48	510/72	479	624/79	598	589
1vkn_dm	1116	82.6	0.43	884/137	836	1016/124	1051	998
2re7_dm	88	66.7	0.46	55/11	56	73/12	67	57
3cuc_dm	221	41.7	0.61	138/31	75	226/47	184	13
3cuc	426	80.4	0.41	287/54	289	397/44	391	357
3dxo_dm	146	62.1	0.50	79/18	67	113/23	118	11

Table 5
Details of the *NUT*, *HEL* and *BED* calculations based on AlphaQ.

F_c fragments. 15 conformers were used. $N_{\rm NUT}$ is the number of residues correctly located by the PRCTF. % is the percentage (100% is $N_{\rm PDB}$). ⟨r.m.s.d.⟩ is the mean r.m.s.d. (see text). $N_{\rm CNH}$ is the number of connected AlphaQ peaks divided by the number of chains of connected fragments in *HEL*. $N_{\rm HEL}$ is the number of amino-acid residues correctly located by *HEL*. $N_{\rm CNB}$ is the initial number of connected fragments in *BED* comparable to $N_{\rm HEL}$. $N_{\rm BED}$ is the final number of correctly located residues and $N_{\rm SEQ}$ is the number of located residues with the correct sequence.

Code	$N_{\rm NUT}$	%	⟨r.m.s.d.⟩ (Å)	$N_{\rm CNH}$	$N_{\rm HEL}$	$N_{\rm CNB}$	$N_{\rm BED}$	$N_{\rm SEQ}$
1ab1	46	100.0	0.17	43/2	41	42/1	44	44
1g7a	191	95.0	0.20	147/14	152	164/13	180	180
9rnt	100	96.2	0.27	80/10	84	84/8	89	89
1a70	90	92.8	0.24	73/9	75	78/7	80	80
1a32	81	95.3	0.15	84/3	80	84/3	78	78
12gs	398	94.8	0.18	362/23	379	368/15	381	381
1bfe	96	87.3	0.28	80/12	82	90/10	90	90
1zx3	84	97.7	0.32	71/9	81	81/5	83	79
2b17	101	83.5	0.40	65/11	73	75/8	87	85
2gpi_mad	82	90.1	0.34	60/6	58	67/5	73	72
RNASE_dm	184	95.8	0.46	148/15	150	158/13	160	158
RNASE	192	100.0	0.27	175/14	179	174/9	171	169
1vku_mad	78	91.7	0.37	74/7	75	73/4	79	77
TP47_dm	713	89.0	0.52	552/79	545	595/66	604	576
1vkn_dm	1195	88.5	0.44	913/115	975	996/93	1023	982
2re7_dm	121	91.7	0.49	63/10	72	86/12	105	103
3cuc_dm	412	77.7	0.59	259/44	270	354/41	320	290
3cuc	486	91.7	0.42	357/44	384	427/36	441	408
3dxo_dm	210	89.4	0.49	120/22	138	157/20	161	151

and $R_{\rm ex}$ = 5.0 Å also covers intermediate extended β_2. The optimum was found at $R_{\rm ex}$ = 5.0 Å, but again values that differ by 0.5 Å may be acceptable. $R_{\rm ex}$ can be estimated as an average of $R_{\rm max}$ and $R_{\rm mean}$ (Table 1) calculated from the known protein structures. The recommended values for tripeptidic fitting resulting from this analysis are $n_{\rm max}$ = 6, $l_{\rm max}$ = 8, GrdZ = 0.5 Å, $R_{\rm ex}$ = 5.0 Å. The analogous recommendation for AlphaQ fragments based on analogy with AlphaT and a limited number of tests is $n_{\rm max}$ = 6, $l_{\rm max}$ = 8, GrdZ = 0.5 Å, $R_{\rm ex}$ = 6.0 Å. These values were subsequently verified by numerous calculations.

3.3. The performance of the PRCTF

Calculation of the performance of the PRCTF with AlphaT and AlphaQ fragments was carried out with optimized parameters, but only ncon increased to 15. The number of searched peaks was 1.5$N_{\rm SEQ}$. The number of residues found was the main criterion for evaluating the PRCTF. We evaluated the performance on the basis of residues because the peaks were overlapped and some fits were only partial. In the case of AlphaT a single peak can locate up to two residues, while an AlphaQ peak can locate up to three residues. The accuracy of fragment positioning, reflected in the r.m.s.d., is the second criterion. ⟨r.m.s.d.⟩ is the mean r.m.s.d. for all correctly located fragments. A local program *CMP* (Pavelcik, unpublished work) was developed for comparing the results with the PDB file deposited in the PDB.

The fragments located by the PRCTF were connected using the *HEL* program (Pavelcik, 2006*b*) and a research version of the program *BED* (*Building into Electron Density*). The output of the connecting is a PDB file of either a polyalanine or a polypeptide model. The number of connected peaks as well as the number of formed chains represents the third criterion for the evaluation of the performance of the PRCTF. The details of the PRCTF and the main-chain connecting are summarized in Tables 4 and 5.

As previously mentioned, the performance of the PROTF employing five monopeptide conformers of the AlphaP0 fragment at 2.3 Å resolution (Pavelcik, 2004; Table 3) remained low and only about 50–75% of the structure could be located. As can be seen in Table 4, the performance increased considerably using AlphaT conformers. About 95% of residues could be located at 1.8–2.3 Å resolution and 90% could be located at 2.5–2.7 Å resolution, providing good-quality phasing. The situation is also quite satisfactory for *DM* phases, but beyond a resolution of 2.4 Å only about 50% of the structure could be located by the PRCTF (3cuc, 3dxo) with AlphaT conformers.

The PRCTF with AlphaQ fragments shows approximately the same performance as with AlphaT fragments despite the fact that 15 search conformers represent only a small fraction of all conformation families. Building with AlphaQ fragments demonstrates a considerable improvement over the AlphaT fragments at low resolution and in the poorer maps; about 80% of the structure could be directly located at a resolution of 2.5–2.7 Å. The data in Tables 4 and 5 show that the performance of the PRCTF is clearly a function of resolution, phase quality and fragment size.

The accuracy of the positioning is also a function of resolution and phase quality. The ⟨r.m.s.d.⟩ is 0.2 Å at a resolution of 1.9 Å and is 0.4 Å at a resolution of 2.7 Å with good-quality phasing. Building with the AlphaT fragments is more accurate than building with the AlphaQ fragments, probably owing to the effect of the fixed bond angles. The accuracy has increased by approximately a factor of two using real-space refinement compared with the previous method based on phased flexible refinement (Pavelcik, 2003).

Building into the experimental map is more difficult than building into maps with accurate phases, as can be seen from the results for RNASE and 3cuc; however, at high resolution the difference is not critical. The output is sufficiently high in both cases. The method of map calculation may influence the results at low resolution, but at the same time with automated procedures it is highly impractical to calculate several maps and then visually inspect the quality. The output of the PRCTF calculated with the oligopeptide fragments is a good starting point for connecting, extending and sequence docking.

3.4. Connecting, extending and sequence docking

Analysis of partial overlaps, connecting localized fragments into chains, extending the chains and sequence docking were carried out with the program *BED* and then compared with results of the program *HEL*. The results of connecting with a new algorithm that incorporates partial overlaps show an improved performance, as can be seen in the $N_{\rm CNH}$ and $N_{\rm CNB}$ columns of Tables 4 and 5, particularly at lower resolutions. The program *HEL*, which does not use analysis of overlapped peaks and is based on an accurate overlap of $(n - 1)$ peptide units of an n-peptidic fragment, is not able to connect all of the located peaks and thus creates many small chains. The connecting algorithm in the program *BED* was therefore relaxed to also include partially overlapped oligopeptides in order to produce longer chains. Analysis of partial overlaps improved the performance by more then 10% and in some AlphaQ cases by more than 30%.

Extension contributed positively to the completion of the model. The number of chains decreased and the number of correctly built residues increased, as can be seen from the $N_{\rm CNB}$ and $N_{\rm BED}$ columns. The overall performance of model building resembles that of the PRCTF. Structures up to a resolution of 2.3 Å are built almost completely with AlphaT fragments and good-quality phases. Performance is significantly reduced beyond this point, particularly with less accurate phases. In contrast, the ability of AlphaQ fragments to complete the structure under ideal conditions is reduced because of the relatively incomplete set of search conformers compared with the AlphaT fragment, but the performance is very good at low resolutions and with *DM* phases.

The development of methods for sequence docking is almost finished, but not completely. A critical point seems to be the correct building of side-chain conformations. The rotamer method frequently fails to build the longest side chains (Arg and Lys) because of fitting into distant main-chain or side-chain electron density (see also Jones & Kjeldgaard, 1997). A method based on the rotamer-averaged electron density (Cowtan, 2008) may be a useful alternative for the longest chains. Using 20 dipeptidic side-chain fragments instead of short side-chain fragments [N—C$^{\alpha}$(R)—C; Pavelcik, 2004] improved the positioning of the CB atom and the direction of the CA—CB vector. The insertion of dummy atoms further increased the sensitivity, particularly for short side chains. A 20 × N table can be used instead of the 18 × N table used in Pavelcik (2004). Relatively small SideD fragments are not very stable when refined into low-quality maps, which is reflected by the failure of the sequence docking in a few cases. This can be seen in the $N_{\rm SEQ}$ column in Table 4.

3.5. *REFMAC* recycling

*CCP*4 v.6.1.2 was used for this calculation. The model built with *BED* is exported as a PDB file that can be read directly by *REFMAC*5 and *vice versa*. The program *MTZ2VARIOUS* serves as a communication tool between *CCP*4 and *BED*. The coefficients h, k, l, FP, SIGP, FC, PHIC, X, PHIX (where X and PHIX are coefficients for the calculation of Fourier synthesis, *e.g.* FWT and PHWT) are exported for model rebuilding or a local phase-mixing program.

The first model is built into an experimental (*DM*) map and refined with *REFMAC*. The *REFMAC* log file and the PDB file of the refined model are then analyzed for inconsistencies; some incorrect residues can be deleted. The model corrections are controlled by $R_{\rm free}$ and zBOND. Several cycles of model building into a *DM* map and refinement may be carried out if the model improves. In the next step, the *DM* phases and *REFMAC* phases are combined together and model building progresses into a map of higher quality. If the model is large enough and the sequence is assigned, only phases calculated by *REFMAC* are used in the next model-building step. We present two representative cases below.

3.5.1. Case study of RNASE. In this study, we used reflection data with density-modified phases (Dodson, 2004) as a starting point for model building. The r.m.s. phase error of the input phases compared with the final refined phases was 75.3°. We applied the same parameters for calculating the PRCTF as were used for the calculations in Table 4. The number of PRCTF peaks entered into the model building was 288. The first run of model building with the *BED* program produced a model that formally consisted of 206 residues (192 virtual atoms in seven chains), 150 residues of which were correct when compared with the final refined model. The sequence was not assigned correctly. This model was refined with *REFMAC* to R and $R_{\rm free}$ values of 0.37 and 0.37, respectively, and subsequently entered into a second run of *BED*. The PDB file of the refined model was deconvoluted back into the AlphaT model. The deconvolution verified 186 virtual atoms. In this case, the sequence was assigned correctly and we located 160 correct residues out of 192 (but many residues were partly correct). The R factor of the *REFMAC* refinement fell to 0.2692 and $R_{\rm free}$ fell to 0.303. In the third cycle of model building the phases of the reflections produced by *REFMAC* were mixed with the original *DM* phases. In this case, we obtained 184 correct residues. However, six more cycles were needed to obtain all 192 residues with correct *cis*-peptides and disulfide bridges. The polypeptidic model was refined to an R factor of 0.2286 and an $R_{\rm free}$ of 0.2693. No water molecules were included. Several side chains had partly incorrect conformations. CPU times on a 2.8 GHz Intel Core2 Duo CPU were 424 s for calculating the PRCTF and 435 s for model building. The *REFMAC* time (30 cycles) was 163 s (for comparison).

3.5.2. Case study of 2re7. In this study, the 58118-FH7490A-1-dm_reflect_file.mtz file obtained from the DATA_PROCESSING/DM directory of JCSG was used in *CCP*4. Because of the lower resolution (2.5 Å), we used an AlphaQ fragment for the PRCTF calculation. The 201 acquired PRCTF peaks were entered into subsequent model building. Three cycles of model building and refinement were performed employing the original *DM* map. The first model building produced 120 residues, of which 105 were located correctly. The sequence was assigned and *REFMAC* refined the structure to R = 0.323 and $R_{\rm free}$ = 0.388. In the next two cycles $R_{\rm free}$ fell to 0.367 and in cycles 4–10 we used *DM* and *REFMAC* mixed phases, weighted $F_{\rm o}$ synthesis or a $2mF_{\rm o} - DF_{\rm c}$ map using modified FWT coefficients (F. Pavelcik, unpublished work) and PHWT phases. The model building stalled at 920–940 atoms and 111–114 correct residues. Two blocks of the chain were missing: residues 1–6 and 26–35. In cycle 11 the missing link 26–35 was finally built, but the subsequent refinement had problems refining these residues. The R factor and log-likelihood were oscillating, there were obviously contradictory requirements of the geometrical restraints and $F_{\rm o}/F_{\rm c}$ in refinement and the $R_{\rm free}$ increased to 0.391. Therefore, we deleted the residues with the highest B factors (His32, Ala33 and Trp34) and the $R_{\rm free}$ fell to 0.351. The model was rebuilt again in cycle 12 and we then deleted Pro35, the true source of the incorrect building. In cycles 13–14 we finally built the entire chain correctly with residues 6–131. The R factor converged to a more satisfactory 0.259 and the $R_{\rm free}$ fell to 0.315. It is worth mentioning that normal methionines were employed instead of selenomethionines because the model-building program only uses the 20 standard amino acids, with no option to modify the side chains. Unfortunately, we were not able to find residues 1–5 as the PRCTF was not able to locate any of these residues in the difference map.

4. Conclusions

The availability of conformation families for tripeptidic and tetrapeptidic fragments (not only the clusters of CA atoms usually analyzed in bioinformatics) extended the applicability of the PRCTF to model building at lower resolution. It is possible to build almost an entire protein structure with only 15 conformation families. The overlap analysis of partially located fragments has enabled us to considerably reduce the number of search conformations in the PRCTF. This is quite a striking result, especially for the tetrapeptidic fragment (AlphaQ), because the number of conformation families has been estimated to be as high as 133 (Pavelcik & Pavelcikova, 2007). More then 80% of the structure could be directly located as a result of the PRCTF at resolutions up to 2.7 Å. Fragments consisting of 3–4 peptide units are large enough to be located reliably and refined into low-quality electron density, but at the same time they are small enough to suffer from the cumulative errors of fixed bond angles. Considering this and previous calculations, we can conclude that AlphaT/Q-based model building proves to be more effective in practice than the AlphaP0 strategy (Pavelcik, 2003) because it covers a broader resolution range (0.9–2.7 Å). The CPU times are conveniently shorter because the FFT grid can be almost doubled in the PRCTF, which leads to an eightfold reduction of CPU time, while the number of searched conformers only increases by a factor of three.

Each fragment is located and refined 'locally' and independently of other fragments. Implementation of the RSR further increased the reliability of fragment location and the accuracy and provided new tools such as the real-space R factor to detect false fragments. The correctness of fragment location and refinement can be monitored by the quality of the overlap of the fragment with its neighbouring fragments in the polypeptide chain. These overlap-based parameters that are independent of position may be used as a real-space analogue of the reciprocal $R_{\rm free}$ factor in the evaluation of the quality of the model in model-building procedures. The presented model-building techniques (not including *REFMAC* refinement) can build a model with an overall r.m.s.d. better than 0.5 Å.

A minor disadvantage of the present AlphaT concept is that the modelling of the *cis*-peptide bond is more difficult to implement and has to be treated differently than in the AlphaM/AlphaD concept (Pavelcik, 2004). A practical compromise is that *cis*-peptide bonds are constructed after the transformation of the AlphaT virtual model to a SideD virtual model. Until this point the chains are usually disconnected at the *cis*-peptide bond. *Cis*-peptide bonds cannot be built using the methods developed for the AlphaQ fragment. One disadvantage of the AlphaQ concept is that the chain ends of the model cannot be built completely unless the fragments are partly extended into the solvent.

The combination of model building, *REFMAC* refinement and model rebuilding is particularly fruitful. The restraint refinement can correct many small building errors and back-transformation into the model of independent building blocks verifies the *REFMAC* changes. In principle, the entire protein model can be constructed in this way because the starting models are sufficiently large. The partial model can be used to improve the phases of structure factors and the next building step takes advantage of a better electron-density map. In the case of RNASE the entire model was built completely without computer graphics.

Principles for automatic, accurate and complete model building have been developed. The user-oriented version of the computer program *BED* is in the final stages of development and will be published later. The presented combined AlphaT/SideD strategy seems to be a good choice for resolutions of around 2 Å and the AlphaQ strategy for resolutions of around 2.5 Å. At lower resolutions partial models can be obtained. We expect that analogous pentapeptidic and hexapeptidic strategies will shift reliable and accurate model building to even lower resolutions. This study is an important contribution to achieving our goal of developing methods that are capable of building complete biomacromolecular structures automatically (Pavelcik, 2003). Computer graphics will only be needed for a final check and small corrections.

The authors are grateful to Professor E. J. Dodson and Dr D. R. Tomchick for supplying experimental data. We would also like to thank the JCSG data archive for providing a source of test data and to an unknown referee for suggesting the JCSG as a source of data. The research was partly supported by a grant from the Ministry of Education, Youth and Sport of the Czech Republic (No. MSM6215712403).

References

Binda, C., Coda, A., Aliverti, A., Zanetti, G. & Mattevi, A. (1998). *Acta Cryst.* D**54**, 1353–1358.

Brunger, A. T., DeLaBarre, B., Davies, J. M. & Weis, W. I. (2009). *Acta Cryst.* D**65**, 128–133.

Chapman, M. S. (1995). *Acta Cryst.* A**51**, 69–80.

Clemons, W. M. Jr, Davies, C., White, S. W. & Ramakrishnan, V. (1998). *Structure*, **6**, 429–438.

Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.

Cowtan, K. (2008). *Acta Cryst.* D**64**, 83–89.

Deka, R. K., Machius, M., Norgard, M. V. & Tomchick, D. R. (2002). *J. Biol. Chem.* **277**, 41857–41864.

Diamond, R. (1971). *Acta Cryst.* A**27**, 436–452.

DiMaio, F., Kondrashov, D. A., Bitto, E., Soni, A., Bingman, C. A., Phillips, G. N. & Shavlik, J. W. (2007). *Bioinformatics*, **23**, 2851–2858.

DiMaio, F., Shavlik, J. W. & Phillips, G. N. (2006). *Bioinformatics*, **22**, e81–e89.

Dodson, E. J. (2004). Personal communication.

Doyle, D. A., Lee, A., Lewis, J., Kim, E., Sheng, M. & MacKinnon, R. (1996). *Cell*, **85**, 1067–1076.

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

Hattne, J. & Lamzin, V. S. (2008). *Acta Cryst.* D**64**, 834–842.

Hovmöller, S., Zhou, T. & Ohlson, T. (2002). *Acta Cryst.* D**58**, 768–776.

Jones, T. A. & Kjeldgaard, M. (1997). *Methods Enzymol.* **277**, 173–208.

Joosten, K., Cohen, S. X., Emsley, P., Mooij, W., Lamzin, V. S. & Perrakis, A. (2008). *Acta Cryst.* D**64**, 416–424.

Kleywegt, G. J. & Jones, T. A. (1997). *Acta Cryst.* D**53**, 179–185.

Korostelev, A., Bertram, R. & Chapman, M. S. (2002). *Acta Cryst.* D**58**, 761–767.

Levinthal, C. (1968). *J. Chim. Phys. Phys. Chim. Biol.* **65**, 44–45.

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

Martinez-Oyanedel, J., Choe, H. W., Heinemann, U. & Saenger, W. (1991). *J. Mol. Biol.* **222**, 335–352.

Morgan, H. L. (1965). *J. Chem. Doc.* **5**, 107–113.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Oakley, A. J., Lo Bello, M., Nuccetelli, M., Mazzetti, A. P. & Parker, M. W. (1999). *J. Mol. Biol.* **291**, 913–926.

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

Pavelcik, F. (2004). *Acta Cryst.* D**60**, 1535–1544.

Pavelcik, F. (2006*a*). *J. Appl. Cryst.* **39**, 483–486.

Pavelcik, F. (2006*b*). *J. Appl. Cryst.* **39**, 287.

Pavelcik, F. (2008). *J. Appl. Cryst.* **41**, 62–67.

Pavelcik, F. & Pavelcikova, P. (2007). *Acta Cryst.* D**63**, 1162–1168.

Pavelcik, F. & Schneider, B. (2008). *Acta Cryst.* D**64**, 620–626.

Pavelcik, F. & Vanco, J. (2006). *J. Appl. Cryst.* **39**, 315–319.

Pavelcik, F., Zelinka, J. & Otwinowski, Z. (2002). *Acta Cryst.* D**58**, 275–283.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Sevcik, J., Dodson, E. J. & Dodson, G. G. (1991). *Acta Cryst.* B**47**, 240–253.

Singh, N., Jabeen, T., Sharma, S., Somvanshi, R. K., Dey, S., Srinivasan, A. & Singh, T. P. (2006). *Acta Cryst.* D**62**, 410–416.

Smith, G. D., Pangborn, W. & Blessing, R. H. (2001). *Acta Cryst.* D**57**, 1091–1100.

Terwilliger, T. C. (2003*a*). *Acta Cryst.* D**59**, 38–44.

Terwilliger, T. C. (2003*b*). *Acta Cryst.* D**59**, 45–49.

Tomchick, D. R. (2001). Personal communication.

Yamano, A., Heo, N. H. & Teeter, M. M. (1997). *J. Biol. Chem.* **272**, 9597–9600.

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