Acta Cryst. (1997). D**53**, 665–672

Polyalanine Reconstruction from Cα Positions Using the Program *CALPHA* Can Aid Initial Phasing of Data by Molecular Replacement Procedures

ROBERT M. ESNOUF†

The Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford OX1 3QU, England, and The Oxford Centre for Molecular Sciences, New Chemistry Building, South Parks Road, Oxford OX1 3QT, England. E-mail: robest@biop.ox.ac.uk

(*Received 23 August 1996; accepted 16 April 1997*)

Abstract

The Cα positions for a protein can provide a scaffold for the reconstruction of more complete models. Reconstructions can be by manual rebuilding, from geometric solutions to the constraints on main-chain torsion angles or from databases of known protein structures. The last method is usually the most convenient and reliable. This paper describes a database reconstruction program, *CALPHA*, and assesses its accuracy and reliability by test reconstructions of well refined structures. Typically, backbone atoms are repositioned to within 0.3 Å of their original positions. This corresponds to regenerating main-chain torsion angles to within 15°. Uses of *CALPHA* for automating refinement procedures are discussed. In particular, the uses of Cα-only and reconstructed polyalanine models of HIV-1 reverse transcriptase for cross-rotation and translation function searches are compared. The *CALPHA* polyalanine model is found to provide more selectivity for approximately correct orientations. The effect on the translation function is dependent on the resolution shell employed. It is expected that these observations will be applicable in other cases.

1. Introduction

Structures are sometimes made available in a reduced form, usually consisting of the Cα-atom positions only; in the first stages of electron-density map interpretation it is often convenient to construct a Cα trace. In both cases, polyalanine protein models can be reconstructed by automated procedures with some degree of accuracy and reliability (Jones & Thirup, 1986; Jones, Zou, Cowan & Kjeldgaard, 1991). The major restraints controlling such reconstructions arise from the stereochemistry of the protein backbone: bond lengths and bond angles are (virtually) fixed and only certain combinations of the ω, φ and ψ torsion angles are possible. Normally, $\omega \simeq 180°$ (*trans* conformation) or $\omega \simeq 0°$ (*cis* conformation, occasionally prior to a prolyl residue) as a result of restricted rotation around the peptide linkage. The φ and ψ angles are more variable, but most combinations cannot occur since they result in conformations with bad steric clashes between atoms (Ramachandran, Ramakrishnan & Sasisekharan, 1963). These restrictions are sufficient to allow accurate positioning of the N and C atoms. The planarity of carbonyl groups allows positioning of the O atoms and the tetrahedral Cα-atom geometry dictates the position of the Cβ atoms. As the database of well refined high-resolution structures has grown, it has become evident that the side-chain torsion angles ($\chi1$, $\chi2$, ...) for each residue also have preferred values, corresponding to staggered conformations (Laskowski, MacArthur, Moss & Thornton, 1993). Selecting side-chain conformations which fit observed preferences and avoid bad steric clashes provides a method for building all-atom models with moderate accuracy.

Manual rebuilding is the most time-consuming method for model reconstruction, an example being provided by the study on flavodoxin by Reid & Thornton (1989). The main-chain atoms were repositioned with a root-mean-square (r.m.s.) error of 0.57 Å. Models can also be rebuilt geometrically (Purisima & Scheraga, 1984; Luo *et al.*, 1992; Payne, 1993) by considering each peptide plane to have only one degree of freedom: the rotation of the plane around the axis connecting consecutive Cα atoms. The well defined bond angles around Cα atoms then define relative plane orientations and usually only one complete set of orientations is consistent with good φ and ψ angles. Although elegant, such methods rely on accurate Cα positions and small deviations from idealized geometry or model errors can be amplified into significant distortions of the final structure.

The most popular method for reconstruction is the use of a database, where a set of well refined protein structures is considered as a collection of small oligopeptide units. These units are fitted, in turn, to fragments of the Cα framework and the best fitting ones are used as building blocks from which the final model is constructed. Apart from graphical database methods (Jones & Thirup, 1986; Jones *et al.*, 1991), two programs have been described (Claessens, van Cutsem, Lasters & Wodak, 1989; Holm & Sander, 1991). The programs fit

† Present address: Rega Institute for Medical Research, Katholieke Universiteit Leuven, Minderbroedersstraat 10, B-3000 Leuven, Belgium.

fragments in a similar way but the choice of fragment length, the selection of the best-fitting database units and the construction of the final model vary. For example, reconstructions of citrate synthase (2CTS; Remington, Wiegand & Huber, 1982) by these programs have been reported: the r.m.s. errors in the N, Cα, C and O positions (compared to the original model) were 0.62 (Claessens *et al.*, 1989) and 0.45 Å (Holm & Sander, 1991; however, this model was slightly incomplete). For the same reconstruction, the database program described herein produced a complete model with an equivalent r.m.s. error of 0.43 Å. Database programs have the important practical advantage that small errors in the Cα positions do not cause large distortions in the final model. Indeed, it has been shown that small random errors introduced into Cα coordinates can be partially corrected by database rebuilding (Holm & Sander, 1991).

Reconstruction from Cα positions can help structure refinement. Apart from being used in the initial stages of model building, *CALPHA* was used in the largely automatic refinement of black-swan egg-white lysozyme (Rao, Esnouf, Isaacs & Stuart, 1995) from the Cα trace of an homologous lysozyme. One feature of this refinement was that a simple protocol for including side-chain atoms produced a model with few errors. This paper examines a third use: it describes how a polyalanine reconstruction can help identify an initial orientation and position for a protein model in a unit cell. This was used for the structure determination of a new crystal form of HIV-1 reverse transcriptase (RT) for which cross-rotation function searches using a partial Cα trace had failed to identify the correct orientation, but for which a model reconstructed using *CALPHA* was successfully used for the same search (Ren, Esnouf, Garman *et al.*, 1995; Esnouf *et al.*, in preparation).

2. Description of the *CALPHA* program

The program *CALPHA* is written in Fortran and provides all the functionality described below. The time taken to model a protein depends on the size of both the protein and the database, but is typically only a few minutes on current workstations. A central part of the program is the superpositioning of one structure fragment on another and obtaining the r.m.s. error in the superposition. *CALPHA* achieves this using the Fortran subroutine *MATFIT* (Remington, unpublished work) based on the algorithms of McLachlan (1972) and Kabsch (1976, 1978).

2.1. *Making a database*

CALPHA can work directly from files in PDB format (Bernstein *et al.*, 1977), creating a database before modelling each protein. However, it is more convenient to create and save a database. Database construction simply requires the names of proteins and the names of the relevant PDB-format files. Essentially, a database consists of main-chain and Cβ-atom coordinates for a set of well refined protein models representing diverse structures. Provided databases are sufficiently large, the resulting models are not substantially affected by the actual choice of proteins, and even very small databases (of a few hundred residues) can still produce adequate results (Esnouf, 1992). For this study, a database of 16 553 residues from 72 structures (generally refined against data to at least 2.0 Å resolution) was employed.

2.2. *Modelling individual fragments from a database*

The first stage of rebuilding matches the Cα trace against a database by considering the trace as a series of short overlapping fragments (Cα fragments). The default fragment and overlap lengths were selected on the basis of a series of modelling tests (Esnouf, 1992): with short fragments the Cα coordinates defined the structure less precisely, while with long fragments it became difficult to find good database matches. Six residue fragments represented the best compromise for the size of the database employed. A three-residue overlap then ensures that the ends of one Cα fragment (which are presumably

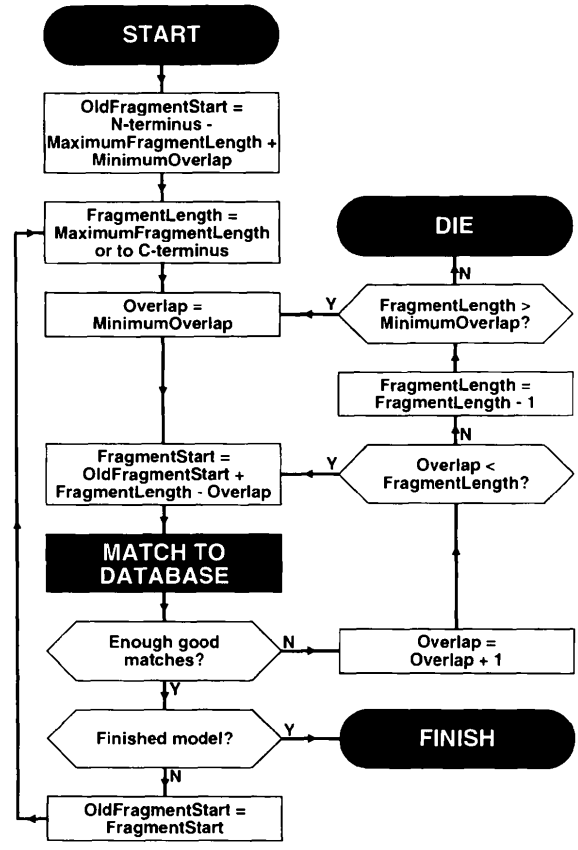

Fig. 1. A schematic diagram showing the algorithm used in *CALPHA* to divide a Cα trace into fragments and to match those fragments to protein structure fragments in a database.

less reliably modelled than the centres of fragments) are at the centres of adjacent fragments.

Starting from the N terminus of the trace, a Cα fragment is selected (Fig. 1) and all database fragments of the same length are superposed on it. A list of the best-fitting database fragments is created (typically the 50 fragments with the smallest r.m.s. superpositioning errors). Provided that enough (usually ten) of these database fragments have r.m.s. errors less than a (user-defined) cut-off value (typically 0.5–1.0 Å), the program selects a new Cα fragment of default length and which overlaps the previous one by the default number of residues. This process is repeated until the C terminus is reached, the final fragment length being adjusted if necessary (Fig. 1). If too few database fragments match a Cα fragment well enough, then the program automatically tries to readjust: firstly, by increasing the overlap with the previous fragment by one residue at a time and then, if necessary, by decreasing the fragment length (Fig. 1). Overlap readjustment can be useful when secondary structural units are separated by only a couple of residues since it allows *CALPHA* to model each unit separately instead of having to find a single database fragment which ends one unit and starts the second. However, the more drastic measure of decreasing the fragment length is sometimes necessary and in the extreme case of the fragment length having to be reduced to the default overlap length the program terminates. In practice this only occurs when one tries (inadvertently!) to model over a chain break or when the Cα trace is extremely approximate.

2.3. *Selecting fragments for the final model*

Having divided the Cα trace into overlapping fragments for which there are enough good database matches, the program selects the individual database fragments from which to construct the final model. The algorithm used selects fragments to minimize the weighted sum of two r.m.s. errors. The first error is simply the r.m.s. superposition error between each database fragment and the appropriate Cα fragment. Once superposed, database fragments for neighbouring Cα fragments have an overlap region. The degree of continuity between each pair of database fragments can be measured as the r.m.s. error between atom positions in this region. This error should also be minimized since the program is modelling a continuous structure.

The selection algorithm is similar in concept to algorithms used in sequence alignment routines (Needleman & Wunsch, 1970; Smith & Waterman, 1981) and in other structure comparison programs, such as *SHP* (Stuart, Levine, Muirhead & Stammers, 1979). The r.m.s. error for the superposition of each database model on the first Cα fragment is stored in a set of running totals. Moving on to the next Cα fragment, the running total for each database model is its own r.m.s. superposition error plus the smallest sum of the weighted r.m.s. overlap error between this database model and each of the database models for the previous Cα fragment and the appropriate previous running total. A list of fragments leading to each running total is also stored and the algorithm moves to the next Cα fragment. When all Cα fragments have been covered, the list of database fragments leading to the lowest running total gives the optimal choice for that weighting scheme. Weighting in favour of small r.m.s. errors between database models and Cα fragments gives a model which fits the original trace closely at the expense of increased stereochemical strain; weighting in favour of small r.m.s. overlap errors allows the program more freedom to deviate from the Cα trace in order to use database fragments giving good stereochemistry. In practise, it has been found that giving equal weighting to the two terms gives a good compromise.

2.4. *Constructing the final model*

The database fragments selected in the previous section are used to construct the final model. Each database fragment is superposed on the Cα fragment it represents and each atomic coordinate in it is stored, being weighted in proportion to the number of backbone atoms separating the atom from the nearest fragment end. Cβ and O atoms are given the same weight as the backbone atoms to which they are attached. Thus, for a six-residue fragment the first N atom has a weight of 1/9, the first Cα and Cβ atoms have a weight of 2/9, and so on. The coordinates for each atom in the final polyalanine model are the weighted average of the coordinates for all equivalent atoms in the contributing database fragments. This smoothes the transitions between fragments, spreading the remaining distortion evenly over the whole overlap region without introducing major stereochemical errors. In a test reconstruction of phosphorylase *b*, the final model had r.m.s. deviations in backbone bond lengths and bond angles of 0.050 Å and 5.0°, respectively, when compared with the Engh & Huber (1991) dictionary values. The source of these errors is primarily the stereochemistry of the proteins making up the database used for the reconstruction, many of which were refined against (older) alternative dictionary values and whose equivalent r.m.s. errors are very similar. However, since a typical use for reconstructed models is in molecular replacement procedures, it is the accuracy of the position of the atoms that is of paramount importance.

2.5. *Adding side chains to the model*

To allow the construction of complete protein models, *CALPHA* has an option for adding appropriate side-chain atoms to the reconstructed polyalanine backbone. Each residue is added in its 'most likely' conformation [based on an analysis of well refined high-resolution structures

in the PDB (Esnouf, 1992)], ignoring any steric clashes with either main-chain atoms or other side-chain atoms. The resulting models have very few bad contacts (Rao *et al.*, 1995) and no attempt is made to refine the structures further (*e.g.* by energy minimization procedures) since the current methods of protein simulation cannot be expected to provide a substantial improvement (Rao *et al.*, 1995). All-atom reconstructions with *CALPHA* for test cases have typical r.m.s. errors in atom positions of 1.5–1.8 Å.

Disulfide bridges will not be reproduced well by this side-chain procedure. When using reconstructed models for molecular replacement procedures this can be crucial since the scattering of X-rays by the S atoms can account for a significant fraction of the total scattering. A simple solution could be to reconstruct each cystine pair based on the Cα coordinates of nearby residues in both polypeptide chains. For each disulfide linkage in the proteins making up a database, the Cα positions for both fragments of (say) five residues centred on the two cystine residues are stored in a separate database. To reconstruct a disulfide linkage one selects the coordinates for the cystine side chains for the best-fitting disulfide as judged from the fit of the nearby Cα atoms. Routines which match pairs of strands simultaneously have been implemented in related database programs for modelling β-strands (Esnouf, 1992) and double-stranded DNA (unpublished program) and have been found to improve the accuracy of reconstruction.

2.6. *Modelling part of a structure*

CALPHA can be used to model part of a structure, for example when a loop has been rebuilt as a Cα trace and a full model is required. In this case, a region longer than the loop by three residues at either end is modelled and the program effectively treats the structures at either end of the loop as the first and last fragments. The fragment selection routine then considers the true structures as the only models for the terminal fragments and thus an intervening loop is constructed which overlaps smoothly the structure to either side of it. This feature has been used in the early stages of several structure refinements, for example with phosphorylase kinase (Owen, Noble, Garman, Papageorgiou & Johnson, 1995) and bovine enterovirus (Smyth *et al.*, 1995).

3. Accuracy of the *CALPHA* program

The accuracy of reconstruction by *CALPHA* can be assessed easily from test cases, in each of which a well refined structure is stripped to its Cα trace and then reconstructed from a database that does not include the test structure. Testing with a series of proteins belonging to different structural classes (Table 1) showed that typically the backbone atoms are repositioned to within 0.3 Å (a figure comparable to the intrinsic error of the coordinate sets themselves) and that Cα atoms are repositioned in the rebuilding process almost as much as other backbone atoms. *CALPHA* models α-helical structures more closely than β structures (compare the reconstructions for myoglobin and plastocyanin summarized in Table 1). It is not clear whether this reflects a greater intrinsic error in the Cα positions in β structures, either for the test traces or in the database, results from some feature of the fragment fitting algorithm of *CALPHA* or results from greater variability among β structures in general. However, the last explanation seems most plausible.

Changes in atom positions on rebuilding also produce changes in the φ and ψ torsion angles, the change typically being less than 15°. However, this figure excludes occasional 180° 'flips' of the peptide plane, usually in loop regions and usually producing conformations in more favoured regions of the Ramachandran plot. These flips are indicative of the essentially conservative nature of the algorithm: *CALPHA* rebuilding frequently results in structures with torsion angles in 'more allowed' positions of the Ramachandran plot.

CALPHA not only models good structures well, but models implausible ones poorly (assuming that the structures in the database are reliable). The use of databases for detecting (possible) errors in structures has been highlighted previously (Brändén & Jones, 1990) and is demonstrated here (Table 2) for two models of ferredoxin: 2FD1 (Ghosh, O'Donnell, Furey, Robbins & Stout, 1982) and 4FD1 (Stout, 1989). As well as the *CALPHA* rebuild being far less faithful for 2FD1 than for 4FD1, especially with respect to the backbone torsion angles, the mean fragment length used by *CALPHA* was substantially less for 2FD1 (5.13 compared with 5.95 residues), revealing the difficulty experienced in the search for suitable database fits.

4. An example of the use of *CALPHA*: HIV-1 RT

Many different crystal forms of HIV-1 RT have been obtained, although few show diffraction to high resolution (*e.g.* Unge *et al.*, 1990; Lloyd, Brick, Mei-Zhen, Chayen & Blow, 1991; Kohlstaedt, Wang, Friedman, Rice & Steitz, 1992; Jones *et al.*, 1993; Jacobo-Molina *et al.*, 1993; Stammers *et al.*, 1994; Rodgers *et al.*, 1995). Recently, a series of related crystal forms have been described (space group $P2_12_12_1$, but with great variation in unit-cell dimensions), and some of these crystals show diffraction to a high resolution limit of 2.2 Å (Stammers *et al.*, 1994; Ren, Esnouf, Garman *et al.*, 1995; Esnouf *et al.*, 1995; Ren, Esnouf, Hopkins *et al.*, 1995). The initial crystal form had unit-cell dimensions of $a \simeq 147$, $b \simeq 112$ and $c \simeq 79$ Å, and diffracted to minimum Bragg spacings of 3.7 Å (Stammers *et al.*, 1994). A data set to 5.0 Å resolution for this crystal form was used for initial structure determination attempts (Table 3). Isomorphous

Table 1. *Sample CALPHA reconstructions for some well refined structures*

Representative well refined protein structures belonging to different structural classes were selected. They were reduced to Cα traces and rebuilt using the program *CALPHA*. The differences between the original and reconstructed models are described in terms of r.m.s. positional and mean angular errors.

Protein	PDB code	Positional r.m.s. error (Å)		Mean angular error (°)		Excluded residues (see text)
		Cα	N, Cα, C	⟨δφ⟩	⟨δψ⟩	
Crambin	1CRN	0.24	0.28	13.0	13.4	46
Plastocyanin	1PCY	0.28	0.33	13.2	13.4	6, 67
Myoglobin	1MBO	0.17	0.23	7.8	9.2	153
Flavodoxin	3FXN	0.30	0.32	14.8	14.8	43, 90, 93, 112, 119–20
HEW lysozyme	†	0.23	0.26	10.1	10.1	67, 74, 102–3
Citrate synthase	2CTS	0.25	0.29	13.0	12.5	2, 67–8, 81–2, 201, 239, 295, 317–9, 437
Carboxypeptidase	5CPA	0.29	0.34	13.3	14.4	2, 56, 58, 70, 134–5, 140, 152, 155, 199, 205, 272, 275, 277–9

† D. C. Philips, unpublished data.

Table 2. *CALPHA reconstructions for two models of the same protein*

The Cα traces for two models of ferredoxin were used for reconstruction. The model with code 2FD1 is inaccurate and has been removed from the PDB. The data show that *CALPHA* is unable to regenerate accurately the main chain for this model, whereas the statistics for rebuilding the model with code 4FD1 are typical of a well refined structure.

Protein	PDB code	Positional r.m.s. error (Å)		Mean angular error (°)		Excluded residues (see text)
		Cα	N, Cα, C	⟨δφ⟩	⟨δψ⟩	
Ferredoxin	2FD1	0.45	0.60	67.4	61.2	43 out of 106 residues have φ and ψ deviations more than 90°
Ferredoxin	4FD1	0.30	0.32	11.8	11.1	14, 29, 90

Table 3. *Data set used for cross-rotation and translation function searches*

The data set used in this study was the same as the one used in the initial structure determination for this crystal form of HIV-1 RT. Fuller details of the data collection, processing and structure refinement will be presented in the appropriate place (Esnouf *et al.*, in preparation).

Data collection site	In-house, 1992
Wavelength (Å)	1.54
Number of crystals	3
Unit-cell dimensions (Å)	$a = 147$, $b = 112$, $c = 79$
Maximum resolution (Å)	5.0
No. of unique reflections	4792
Completeness (%)	80
R_{merge} (%)†	9.2

† $R_{merge} = \sum |I - \langle I \rangle| / \sum \langle I \rangle$.

replacement methods did not yield an interpretable map and the structure was solved (Esnouf *et al.*, in preparation) using a molecular replacement procedure (Rossmann & Blow, 1962; Rossmann, 1972) based on a *CALPHA* reconstruction of an unrefined partial Cα trace for HIV-1 RT (1HVT; Kohlstaedt *et al.*, 1992). To examine why this procedure was successful while attempts using only the Cα trace were inconclusive, the initial stages of the structure determination were repeated using both starting models; the results are analysed below. In this discussion, the model orientations and positions evaluated by the search procedures which are closest to the orientation and position of the refined RT model are referred to as the correct orientations and positions, respectively.

4.1. *Construction of the model*

The partial coordinate set 1HVT (Kohlstaedt *et al.*, 1992) contained 767 Cα positions (out of 1000), corresponding mainly to α and β secondary structural units, but with a complete trace for the RNase H domain (129 residues) based on a structure determination of the isolated domain (Davies, Hostomska, Hostomsky, Jordan & Matthews, 1991). It comprised 55 separate sections of polypeptide chain which were individually modelled using *CALPHA*, except for two sections (residues 186–187 and 666–667, using 1HVT numbering) which were too short. Three sections (residues 161–174, 175–180 and 369–372) could not be modelled satisfactorily by *CALPHA* and these sections were excluded from the model on the assumption that these difficulties indicated unreliable traces. The reconstructed model contained 3696 atoms corresponding to 739 alanine residues and the r.m.s. change in Cα positions was 0.32 Å. Thus, the use of *CALPHA* had increased the completeness of the model from 9 to 45% of non-H atoms and led to the

elimination of 28 residues for which the trace was questionable.

4.2. *Cross-rotation function searches*

The Cα trace and polyalanine model were used for separate cross-rotation function (RF) searches using *X-PLOR* (Brünger, 1992) with identical protocols (Table 4): self-vectors from 5–40 Å were included which resulted in 3757 peaks being used, and a search defined in terms of Lattman angles with $\delta = 2.8^\circ$ resulted in 177 053 orientations being evaluated. For each search the 500 orientations with the highest values for the RF were divided into clusters using $\varepsilon = 0.25$ (see *X-PLOR* version 3.1 manual, page 235; Brünger, 1992). For both searches (Fig. 2*a*) the correct orientation had one of the top 100 RF values (73rd for the Cα trace, fifth for the polyalanine model), although no orientation had a particularly large RF value. With the polyalanine model, not only did the RF peak for the correct orientation have a much higher rank, but orientations in the same cluster as this orientation were much more prevalent among the top 100 peaks (seven occurrences with the Cα trace; 27 with the polyalanine model, these peaks generally having a much higher rank). Fig. 2(*b*) compares the highest RF peaks for the Cα trace directly with the RF values for the equivalent orientations of the polyalanine model. In general, the noise peaks in the RF are less significant relative to the weak peaks for roughly correct orientations with the polyalanine model. Thus, for the resolution range 15–6 Å at least, there is evidence for substantially greater selectivity for orientations near to the correct one with the rebuilt model.

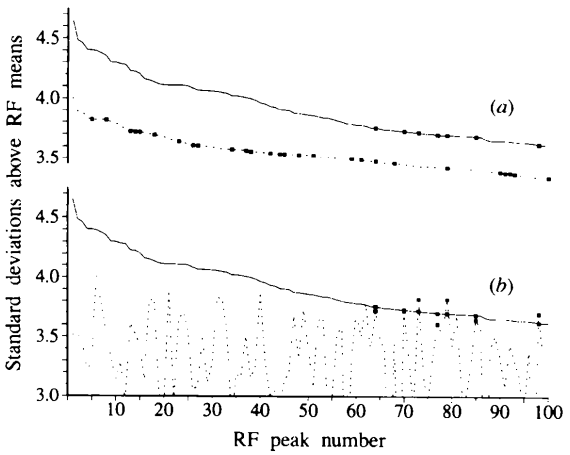

Fig. 2. (*a*) The highest peaks for RF searches using the 1HVT Cα-only model (solid line) and the *CALPHA* polyalanine model (dashed line). Orientations belonging to the cluster around the correct orientation are marked with squares. (*b*) The same plot for the 1HVT model (solid line), but this time the dashed line indicates the RF values for equivalent orientations in the search using the *CALPHA* polyalanine model. Again, orientations belonging to the cluster around the correct orientation are marked with squares.

Table 4. *Cross-rotation-function searches and Patterson correlation refinements*

Results of RF searches and PC refinements for the 1HVT Cα trace and *CALPHA* polyalanine model. Data are given both for the orientation with the maximum RF value and for the correct orientation.

	1HVT Cα trace	*CALPHA* polyalanine model
Resolution range (Å)	15–6	15–6
No. of reflections	2926	2926
No. of atoms in model	767	3696
Mean value of RF	0.518	0.560
Standard deviation of RF	0.080	0.083
Maximum value of RF	0.891	0.893
Value of RF for correct orientation	0.815	0.877
Rank of RF for correct orientation	73rd	5th
Relative height for correct orientation†	3.7	3.8
PC coefficient of original orientation	0.049	0.053
PC coefficient of refined orientation	0.079	0.082

† Standard deviations above RF mean.

Patterson correlation (PC) refinement (Brünger, 1990) of these orientations proved to be a good discriminator, both with the Cα trace and the polyalanine model. In each case an approximately correct orientation gave the best PC coefficient after refinement (Table 4), although the coefficients were small compared to their theoretical maximum (0.250). The orientation of the polyalanine model could be refined to give a slightly higher PC coefficient than could the orientation of the Cα trace.

4.3. *Translation function searches*

For the original structure determination, the translation function (TF) search used the polyalanine model, data from 10 to 8 Å resolution and a search grid spacing of 0.018 in fractional coordinates. The highest TF value was 4.1 standard deviations (σ) above the mean TF. This position was later shown to be approximately correct and led to the first interpretable electron-density maps (Esnouf *et al.*, in preparation).

For this study, searches were performed for a range of resolution shells using both models (Table 5). To accommodate the resolution limits all searches were carried out on a finer search grid (0.013 in fractional coordinates). As a result, the correct position could be located more accurately and the TF value for this position in an otherwise identical search to that above rose to 5.5σ above the mean. A TF search using the Cα trace also gave the maximum TF value at the correct position, but the peak was not quite as prominent as that with the polyalanine model (4.9σ above the mean; Table 5). The correct position gave the highest TF value for both models irrespective of resolution range, but the relative selectivity was not always better with the polyalanine model. Thus, either model could have been used with similar success for the search, but identifying the same position using different models increases the confidence that can be placed in the solution.

Table 5. *Translation function searches using data with different resolution ranges*

The orientation for the polyalanine model after PC refinement was used for TF searches over different resolution ranges with both the Cα-only and polyalanine models. In each case the maximum value for the TF was obtained for the correct position.

Resolution range (Å)	12–10	12–8	12–6	10–8	10–6	8–6
Number of reflections	292	944	2766	652	2474	1822
1HVT Cα-only model						
Mean value of TF	0.057	0.076	0.078	0.062	0.070	0.069
Standard deviation of TF	0.066	0.043	0.021	0.050	0.021	0.023
Maximum value of TF	0.349	0.347	0.235	0.318	0.209	0.171
Relative height of TF maximum†	4.4	6.1	7.6	4.9	6.8	4.7
CALPHA polyalanine model						
Mean value of TF	0.056	0.086	0.075	0.074	0.065	0.063
Standard deviation of TF	0.066	0.039	0.021	0.045	0.022	0.024
Maximum value of TF	0.311	0.332	0.233	0.308	0.210	0.175
Relative height of TF maximum†	3.8	6.6	7.4	5.5	6.4	4.4

† Standard deviations above TF mean.

5. Discussion

CALPHA is one of several programs that reconstruct main-chain coordinates from Cα positions. Although originally written some years ago, its reliability (assessed by reconstruction of well refined protein structures from Cα traces) is at least comparable to those of similar database procedures (Jones & Thirup, 1986; Claessens *et al.*, 1989; Jones *et al.*, 1991; Holm & Sander, 1991). Database rebuilding methods have been found to be more robust than geometrically derived solutions since they have some ability to correct small errors in Cα positions (Holm & Sander, 1991), whereas geometric methods tend to amplify them. Another feature of *CALPHA*, and database rebuilding programs in general, is the very 'conservative' nature of the algorithm: the rebuilt φ and ψ main-chain torsion angles are usually in good regions of the Ramachandran plot. This property can be useful in the early stages of model refinement; either building just a Cα trace or discarding all but the Cα atoms of a crude model and then rebuilding with *CALPHA* can save much manual effort (for an example see Grimes, Basak, Roy & Stuart, 1995). The inability of database programs to rebuild part of a model satisfactorily may indicate either a model error (Brändén & Jones, 1990) or an unusual, but genuine, backbone conformation. In either case these areas of the model should be checked carefully.

This study of the initial HIV-1 RT molecular replacement procedure has shown another use for *CALPHA* and similar programs, *i.e.* models for initial molecular replacement searches can be improved by reconstruction from a database. Three areas of model improvement were anticipated: increasing the number of non-H atoms in the model fivefold, improving areas of the structure and helping to decide which residues to eliminate. The effect on the cross-rotation function search was not to increase the RF value of correct orientations, but to decrease the RF value for incorrect ones. Thus, *CALPHA* effectively reduced the 'noise level' in the RF, allowing the correct, but weak, peaks to stand out more clearly. When *X-PLOR* was used to cluster the peaks of the cross-rotation function search for the Cα trace, only seven of the top 100 orientations corresponded to approximately correct orientations (ranking between 64th and 98th). With the polyalanine model the contrast is striking: 27 of the top 100 orientations are in the cluster with the best being ranked fifth. The value of rebuilding a polyalanine model from a Cα trace to help identify the initial orientation is clearly demonstrated for this example. By contrast, it is found that translation function searches produce similar results with either the Cα trace or polyalanine model.

HIV-1 RT has a two-chain nine-domain structure which results in a very flexible molecule; the orientations and positions of the individual domains can be substantially affected by crystal packing forces (Kohlstaedt *et al.*, 1992; Jacobo-Molina *et al.*, 1993; Rodgers *et al.*, 1995; Ren, Esnouf, Garman *et al.*, 1995; Ren, Esnouf, Hopkins *et al.*, 1995; Esnouf *et al.*, 1995; Ding, Das, Moereels *et al.*, 1995; Ding, Das, Tantillo *et al.*, 1995; Hopkins *et al.*, 1996). For the crystal form described above, two of the domains [the p66 and p51 thumb domains, using the nomenclature of Kohlstaedt *et al.* (1992)] were substantially repositioned from the 1HVT model used for phasing. The false contribution from 136 residues comprising these domains along with the approximate nature of the 1HVT model are the likely causes of the difficulties encountered in obtaining an initial orientation and position. Thus, the benefit obtained by using *CALPHA* to enhance a crude model at a very early stage was crucial in allowing a successful structure determination. This benefit is not expected to be unique for RT and such a procedure may be more widely applicable.

I am grateful to D. Stuart, Y. Jones and D. Stammers for allowing me to re-examine the RT data, D. Stuart and G. Taylor for discussions about *CALPHA*, the referee for pointing out the need for modelling disulfide bridges

specially, G. Taylor and R. Bryan for computing facilities and I. Geens for typing the original manuscript. The Oxford Centre for Molecular Sciences is supported by the BBSRC and MRC.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Brändén, C.-I. & Jones, T. A. (1990). *Nature (London)*, **343**, 687–689.

Brünger, A. T. (1990). *Acta Cryst.* A**46**, 46–57.

Brünger, A. T. (1992). *X-PLOR.* Version 3.1. *A System for X-ray Crystallography & NMR.* Yale University, Connecticut, USA.

Claessens, M., van Cutsem, E., Lasters, I. & Wodak, S. (1989). *Protein Eng.* **2**, 335–345.

Davies, J. F. II, Hostomska, Z., Hostomsky, Z., Jordan, S. R. & Matthews, D. A. (1991). *Science*, **252**, 88–95.

Ding, J., Das, K., Moereels, H., Koymans, L., Andries, K., Janssen, P. A. J., Hughes, S. H. & Arnold, E. (1995). *Nature Struct. Biol.* **2**, 407–415.

Ding, J., Das, K., Tantillo, C., Zhang, W., Clark, A. D. J., Jessen, S., Lu, X., Hsiou, Y., Jacobo-Molina, A., Andries, K., Pauwels, R., Moereels, H., Koymans, L., Janssen, P. A. J., Smith, R. H. J., Kroege Koepke, R., Michejda, C. J., Hughes, S. H. & Arnold, E. (1995). *Structure*, **3**, 365–379.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Esnouf, R. M. (1992). DPhil thesis, Oxford University, England.

Esnouf, R. M., Ren, J., Ross, C., Jones, Y., Stammers, D. & Stuart, D. (1995). *Nature Struct. Biol.* **2**, 303–308.

Ghosh, D., O'Donnell, S., Furey, W., Robbins, A. H. & Stout, C. D. (1982). *J. Mol. Biol.* **158**, 73–109.

Grimes, J., Basak, A. K., Roy, P. & Stuart, D. (1995). *Nature (London)*, **373**, 167–170.

Holm, L. & Sander, C. (1991). *J. Mol. Biol.* **218**, 183–194.

Hopkins, A. L., Ren, J., Esnouf, R. M., Willcox, B. E., Jones, E. Y., Ross, C., Miyasaka, T., Walker, R. T., Tanaka, H., Stammers, D. K. & Stuart, D. I. (1996). *J. Med. Chem.* **39**, 1589–1600.

Jacobo-Molina, A., Ding, J., Nanni, R. G., Clark, A. D., Lu, X., Tantillo, C., Williams, R. L., Kamer, G., Ferris, A. L., Clark, P., Hizi, A., Hughes, S. H. & Arnold, E. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 6320–6324.

Jones, E. Y., Stuart, D. I., Garman, E. F., Griest, R., Phillips, D. C., Taylor, G. L., Matsumoto, O., Darby, G., Larder, B., Lowe, D., Powell, K., Purifoy, D., Ross, C. K., Somers, D., Tisdale, M. & Stammers, D. K. (1993). *J. Cryst. Growth*, **126**, 261–269.

Jones, T. A. & Thirup, S. (1986). *EMBO J.* **5**, 819–822.

Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.

Kabsch, W. (1978). *Acta Cryst.* A**34**, 827–828.

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A. & Steitz, T. A. (1992). *Science*, **256**, 1783–1790.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Lloyd, L. F., Brick, P., Mei-Zhen, L., Chayen, N. E. & Blow, D. M. (1991). *J. Mol. Biol.* **217**, 19–22.

Luo, Y., Jiang, X., Lai, L., Qu, C., Xu, X. & Tang, Y. (1992). *Protein Eng.* **5**, 147–150.

McLachlan, A. D. (1972). *Acta Cryst.* A**28**, 656–657.

Needleman, S. B. & Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443–453.

Owen, D. J., Noble, M. E. M., Garman, E. F., Papageorgiou, A. C. & Johnson, L. N. (1995). *Structure*, **3**, 467–482.

Payne, P. W. (1993). *Protein Sci.* **2**, 315–324.

Purisima, E. O. & Scheraga, H. A. (1984). *Biopolymers*, **23**, 1207–1224.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). *J. Mol. Biol.* **7**, 95–99.

Rao, Z., Esnouf, R., Isaacs, N. & Stuart, D. (1995). *Acta Cryst.* D**51**, 331–336.

Reid, L. S. & Thornton, J. M. (1989). *Proteins*, **5**, 170–182.

Remington, S. J., Wiegand, G. & Huber, R. (1982). *J. Mol. Biol.* **158**, 111–152.

Ren, J., Esnouf, R., Garman, E., Somers, D., Ross, C., Kirby, I., Keeling, J., Darby, G., Jones, Y., Stuart, D. & Stammers, D. (1995). *Nature Struct. Biol.* **2**, 293–302.

Ren, J., Esnouf, R., Hopkins, A., Ross, C., Jones, Y., Stammers, D. & Stuart, D. (1995). *Structure*, **3**, 915–926.

Rodgers, D. W., Gamblin, S. J., Harris, B. A., Ray, S., Culp, J. S., Hellmig, B., Woolf, D. J., Debouck, C. & Harrison, S. C. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 1222–1226.

Rossmann, M. G. (1972). *The Molecular Replacement Method.* New York: Gordon & Breach.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.

Smith, T. F. & Waterman, M. S. (1981). *J. Mol. Biol.* **147**, 195–197.

Smyth, M., Tate, J., Hoey, E., Lyons, C., Martin, S. & Stuart, D. (1995). *Nature Struct. Biol.* **2**, 224–231.

Stammers, D. K., Somers, D. O'N., Ross, C. K., Kirby, I., Ray, P. H., Wilson, J. E., Norman, M., Ren, J. S., Esnouf, R. M., Garman, E. F., Jones, E. Y. & Stuart, D. I. (1994*). J. Mol. Biol.* **242**, 586–588.

Stout, C. D. (1989). *J. Mol. Biol.* **205**, 545–555.

Stuart, D. I., Levine, M., Muirhead, H. & Stammers, D. K. (1979). *J. Mol. Biol.* **134**, 109–142.

Unge, T., Ahola, H., Bhikhabhai, R., Backbro, K., Lovgren, S., Fenyo, E. M., Honigman, A., Panet, A., Gronowitz, J. S. & Strandberg, B. (1990). *AIDS Res. Human Retroviruses*, **6**, 1297–1303.