Acta Cryst. (1996). D52, 833–841

Towards the Automatic Interpretation of Macromolecular Electron-Density Maps: Qualitative and Quantitative Matching of Protein Sequence to Map

JIN-YU ZOU AND T. ALWYN JONES*

Department of Molecular Biology, Uppsala University, Biomedical Center, Box 590, S-751 24 Uppsala, Sweden. E-mail: alwyn@xray.bmc.uu.se

(Received 24 July 1995; accepted 5 December 1995)

Abstract

The matching of the known polypeptide sequence to the electron density is a critical step in solving protein structures by the crystallographic method. Tools have been developed to help in defining the placement of the sequence, both qualitatively and quantitatively. They have been tested with good results on two proteins whose structures were solved by the MIR method.

1. Introduction

The interpretation of an experimental electron-density map is one of the most difficult steps in solving the structure of a protein by X-ray crystallography. In spite of the enormous technical improvements in the last few decades the basic problem in protein crystallography, the phase problem, remains unsolved. In collecting diffraction data only the amplitudes can be measured but not the phases. The phases, therefore, have to be deduced with indirect methods, *e.g.* by multiple isomorphous replacement (MIR) methods. The experimental map usually contains numerous errors and its interpretation requires considerable effort. One of the critical steps during this stage is the placement of the known polypeptide sequence into the density. We report here tools we have developed to objectively judge which amino acid should be associated with a given electron density and to help crystallographers match their estimation of amino-acid identity with the known sequence.

2. Methodology

When trying to build a polypeptide chain into the electron density, one is forced to associate the known sequence with the map. This process is usually based on the shape of the electron density and other factors that are harder to quantify but are derived from the experience of the crystallographer. Since an initial map rarely shows continuous density from amino terminus to carboxyl terminus, a number of discontinuous segments are routinely produced into which attempts are made to place the sequence. The *slider* options in *O* (Jones, Zou, Cowan & Kjeldgaard, 1991) are tools that have been developed to help in defining the placement of the sequence. These tools can be used for both qualitative and quantitative estimation of how well a sequence matches the local electron density.

2.1. *Qualitative estimation*

The *slider_guess* option provides a qualitative estimate of which amino acid is to be associated with some electron-density feature. A guess of the sequence is entered based on the shape and the location of the density. This guess is then compared to the actual sequence *via* a lookup table and a sorted list is calculated for up to 20 stretches in the sequence that give the highest scores. Each score is an evaluation of how well this guess fits a continuous stretch in the sequence. The lookup table is a two-dimensional matrix, containing values between 1 and 10, that relates each amino-acid type with the one-letter symbol of the guess. The scoring matrix is stored in the *O* database as a text entry with the name *.slider_matrix* (the '.' at the start of the name signifies that this is an *O* system data block). Users are free to create their own, personalized matrices. To help describe the principle and problems involved, consider a simple scheme where the user has a choice of only three characters to input for each guess: *S, M, B* to identify small, medium and big residues. According to this scheme, glycine and alanine residues, for example, would score high when guessed as *S*, and score very low when guessed as *M* or *B*. Larger residues such as Val, Leu, would score high for *M*, but because lack of density is not proof for the absence of a large amino-acid type (its absence could be because of disorder or envelope errors in solvent flattening or averaging), a classification of *S* would give a reasonable score. Such residues are not usually associated with a large density so a classification *B* would result in a lower score. Large aromatic residues such as Trp, score highest with *B* assignments and lower with *M* and *S*.

In Fig. 1 the number of assignment symbols has been increased to look more like the one-letter amino-acid code. This table contains more information so special care should be taken with two groups of symbols, the acidic residues D/E and the medium-sized aliphatic L, I, M. Symbols D/E score L/I/M residues low, and symbols L/I/M score D/E residues also low. A medium-length residue of unknown polarity is best described by the symbol *X*. The slider matrix, therefore, can be used

to store more general knowledge about proteins than merely the size of each amino acid.

An example listing of the use of this option is shown in Fig. 2. It is important to note that knowledge of the chain directionality is implicit in making the guess. Often, especially at the beginning of a trace, the directionality is not known so that both directions may have to be checked. The results of the guess may be stored in the user's database for use with other options (described later).

Before deciding the guess of amino-acid type based on the electron density, it may be helpful to use the *O* option *slider_lego* which can show each of the 20 amino acids on the screen in each of their preferred rotamer conformations. This option requires that the crystallographer has built at least a polyalanine model through the density. The necessary methods have been discussed by Jones, Zou, Cowan & Kjeldgaard (1991) and in more detail by Jones & Kjeldgaard (1996). Alternatively, the dipeptide associated with the *baton* option can be used.

2.2. *Quantitative estimation (the automatic slider procedure)*

We have introduced *O* options *rsr_seq* and *slider_calc* to quantitatively score how well the density fits each of the 20 amino-acid types. The *rsr_seq* option requires an initial, well fitting polyalanine chain. For each position in the chain, this option mutates the residue to each

.slider_matrix T 21 72

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
A	10	4	2	2	3	8	3	4	2	4	3	4	7	3	2	8	7	7	2	4	2	10
C	4	10	7	5	7	5	7	7	3	7	8	7	7	5	3	6	6	6	3	7	7	5
D	4	7	10	9	3	2	8	3	4	3	3	10	6	9	3	7	7	3	1	10	3	5
E	3	5	9	10	4	2	8	3	5	2	3	9	6	10	5	6	6	2	1	9	4	5
F	4	7	6	7	10	2	9	7	5	7	8	6	6	7	5	3	4	4	8	6	10	5
G	9	1	1	1	1	10	1	1	1	1	1	1	6	1	1	6	4	4	1	1	1	10
H	5	7	8	8	9	2	10	7	3	9	7	8	6	8	3	5	5	5	6	8	9	5
I	5	7	2	2	5	3	7	10	2	10	7	8	5	8	2	5	6	8	2	10	5	5
K	5	4	6	7	2	2	5	2	10	2	2	6	6	6	8	4	5	2	2	6	2	5
L	3	7	2	2	7	2	8	9	2	10	8	9	6	8	2	7	7	7	2	10	7	5
M	3	8	2	2	8	1	8	6	7	7	10	6	6	7	7	4	6	6	6	7	8	5
N	5	7	10	8	6	3	7	7	5	9	8	10	6	9	5	7	7	7	5	10	6	5
P	7	7	5	3	3	7	3	7	3	3	3	3	10	3	3	8	8	8	3	3	3	8
Q	3	7	8	10	8	3	8	8	6	8	8	8	6	10	6	6	6	6	3	10	8	5
R	2	5	6	7	2	2	6	2	8	2	2	6	5	7	10	2	4	4	2	7	2	5
S	8	7	7	5	2	6	2	6	2	5	4	7	7	5	2	10	8	8	2	7	2	8
T	8	8	7	5	2	6	2	8	2	6	4	7	7	5	2	8	10	10	2	6	2	7
V	8	8	2	2	5	6	5	8	2	7	5	7	7	5	2	8	10	10	2	6	2	7
W	2	6	5	5	8	2	7	5	5	5	6	5	4	5	5	3	4	4	10	5	8	5
Y	4	7	6	7	10	2	9	7	5	7	8	6	6	7	5	3	4	4	8	6	10	5

Fig. 1. The default *slider_guess* scoring matrix. Each row corresponds to one of the 20 amino-acid types. Each column corresponds to a class of residue as assigned by the crystallographer. The scores are integer values, normally in the range 1–10. For example, the column W shows how likely each of the 20 amino acids are, given the fact that the density resembles that of a Trp. Note that the number of classes can be smaller than, equal to or greater than 20. For example, class X comprises medium-sized residues of unknown polarity, and class Z comprises small residues.

of the 20 amino acids in turn. The rotamers of each amino acid are then evaluated to see which one fits the experimental density best. The *O* database includes a library of rotamers found in at least 10% of the representative side-chain conformations analyzed by Ponder & Richards (1987). For each of the rotamers, the real-space refinement feature in *O* is used to optimize the fit of the side-chain atoms to the experimental electron density. The experimental density map ρ_{obs} (*e.g.* an MIR or an averaged map) is pre-calculated on a suitable grid. Then the model density ρ_{calc} is calculated on the same grid using the relationships described by Jones & Liljas (1984; following Diamond, 1971). For each atom the electron density ρ at position r is modelled by a Gaussian distribution of the form,

$$\rho(r) = (Z/A^3)\exp(-\pi r^2 A^2),$$

where Z is the atomic number of the atom and A the effective atomic radius; the density for the model is the sum of the individual contributing atoms. The effective atomic radius is in turn generated from the atomic temperature factor using the relationship,

$$B = 4\pi(cA^2 - A_0^2),$$

where B is an overall temperature factor, c is a constant chosen to remove systematic differences between F_{obs} and F_{calc} and A_0 is the zero-temperature radius (Deisenhofer & Steigemann, 1975). The program then optimizes the fit of the side-chain atoms to the density by pivoting the side chain around the Cα atom to either maximize the convolution product,

$$\sum \rho_{obs}\rho_{calc},$$

or minimize the grid sum,

$$\sum |\rho_{obs} - \rho_{calc}|,$$

where the sums are taken over all the grid points within the box covering all the atoms being processed plus some extra space on all sides.

The real-space refinement algorithms in *O* have a large radius of convergence when a rough initial model is used. Usually at this stage, the model is not complete and there are phase errors so care must be taken to prevent the atoms from moving into density belonging to neighbouring atoms. Two measures are taken to prevent this. First, a residual experimental map is used in the refinement. It is obtained as follows: the model density is calculated only within a cuboid box which covers all the atoms of the residue being worked on, plus a wall of user-defined thickness, *e.g.* 3.5 Å, on all sides. All the atoms within this box, except the Cα and the side-chain atoms of the residue, are used in the calculation of the model density. This model density is then scaled to the experimental density ρ_{obs}. For each grid point within the chosen box, if the calculated model density is non-

zero, that grid point in the experimental density is set to zero. Second, the main-chain atoms are fixed during the refinement, and only the side-chain atoms are allowed to move, by pivoting around the Cα atom. Since the experimental density covering the main-chain N, C and O atoms is set to zero as described above, the side-chain atoms are unable to move into main-chain density.

For the best fitting rotamer of each residue type, a real-space *R* factor is calculated (Jones, Zou, Cowan & Kjeldgaard, 1991) which is used as an index of how well that amino-acid type fits the density. The score is saved in an *O* data block. When the calculation over the chosen zone is completed, each residue will have an entry in the data block (*_residue_rsrseq*) consisting of 20 scores, one for each of the 20 amino acids.

In the original formulation of the real-space *R* factor (Jones, Zou, Cowan & Kjeldgaard, 1991) the envelope used to calculate the grid sum $\sum|\rho_{obs}-\rho_{calc}|/\sum|\rho_{obs}+\rho_{calc}|$ was the set of grid points for which ρ_{calc} was non-zero. Such an envelope gives poor discrimination when the fit of a small side chain is evaluated in the density corresponding to a bigger amino acid, Fig. 3. Therefore, we use envelopes that depend on the amino acid being tested. For the small side-chain residues, Ser, Thr and Val, we also include as part of the envelope, a sphere with its center in the Cα–Cβ direction and 2.0 Å away from Cβ, and with a radius of 3.0 Å (Fig. 3). If a small side chain is placed in a bigger density many of the grid points within this expanded envelope will have non-zero ρ_{obs} values while ρ_{calc} is zero. This will result in a higher *R* factor and better discrimination. If the density is actually for a small side chain, the inclusion of this extra sphere should not affect the *R* factor too much since, in the worst case, ρ_{obs} will be affected by more distant side chains.

Alanine and proline residues are also treated specially. Since the starting model is a polyalanine model, alanine residues should already fit the electron density as well as possible. Also, the only side-chain atom in alanine is the Cβ and since fitting single atoms with the RSR features in *O* suffers severely from grid biasing, the Cβ position is not refined. The calculation of the real-space *R* factor for alanines also includes an extra sphere as described above for Ser, Thr and Val. The side-chain atoms of a proline residue are fixed once the main-chain atoms are placed, so pivoting around the Cα is not carried out.

The calculation of how well glycines fit to density presents a third special case. Since there is no side chain for Gly, the envelope is defined as a sphere with a radius

```
 O > slider_guess
Slid> There are    131 residues in molecule.
Slid> Estimated sequence: DNSQ    ⇒ A guess entered by the user.
Slid>  Average= 0.61,rms =  0.14 ⇒ O scans the whole sequence and find the 20 best fitting segments.
Slid>                       DNSQ
Slid>  Fit   1 0.900 A92    NQVQ ⇒ The best fitting segment starts at residue A92 with a score of 0.900
Slid>  Fit   2 0.850 A58    KNTE ⇒ The second best fitting starts at residue A58.
Slid>  Fit   3 0.850 A61    EISF
Slid>  Fit   4 0.825 A34    NLAK
Slid>  Fit   5 0.825 A15    NFDE
Slid>  Fit   6 0.825 A90    SLNQ
Slid>  Fit   7 0.800 A100   NETT
Slid>  Fit   8 0.800 A101   ETTI
Slid>  Fit   9 0.800 A107   KLVD
Slid>  Fit  10 0.775 A73    TTAD
Slid>  Fit  11 0.775 A71    EETT
Slid>  Fit  12 0.775 A65    KLGQ
Slid>  Fit  13 0.775 A54    ESPF
Slid>  Fit  14 0.775 A69    EFEE
Slid>  Fit  15 0.775 A85    TLAR
Slid>  Fit  16 0.775 A14    ENFD
Slid>  Fit  17 0.775 A59    NTEI
Slid>  Fit  18 0.750 A68    QEFE
Slid>  Fit  19 0.750 A97    WNGN
Slid>  Fit  20 0.750 A121   DVVC
Slid> Do you want to associate with a residue ([Y],n)? yes ⇒ The guess can be stored in
Slid> Molecule and residue name : P2 A92  ⇒ the database, associated with a particular residue that
Slid> Forwards or backwards ([F],b)? F    ⇒ has been built and given a name.
 O > yes
Slid> What name to associate with guess? G1
```

Fig. 2. An example of the result of a *slider_guess* calculation.

of 3.5 Å centered on the Cα atom. Upon completion of the zone calculation, the scores for Gly residues are scaled so that the average score for Gly is equal to the average score of all other amino acids.

Ponder & Richards (1987) found that only χ_1 and χ_2 angles of arginine and lysine residues assumed the prefered rotamers. Therefore, only the Cβ, Cγ and Cδ atoms of these residues are used to evaluate the real-space *R* factor.

The results obtained for the density of a large side chain (a tryptophan) are shown in Fig. 4. The modified envelopes of the small residues result in high scores. The medium-sized amino acids show moderate scores, while the large aromatics have very good scores. Because the density is well defined, in this case the correct residue (Trp) has the best score.

Once the *_residue_rsrseq* data block is calculated for some segments of the model, one can estimate how the scores match the expected sequence with the *slider_calc* option. For a segment of *i* residues, one calculates how well it fits the sequence of residues 1 to *i*, 2 to *i* + 1 and so on. (The calculation is performed from amino terminus to carboxyl terminus which implies that the directionality of the segment is correct.) Since we know the score for each of the 20 amino acids at each position in the segment, we can calculate an average *R* factor for fitting the portion of sequence (1 to *i*, 2 to *i* + 1 and so on) to this stretch of chain. The list is sorted so that the estimate with the lowest average *R* factor appears at the top. To be compatible with other slider options, the value of (1 – *R* factor) is actually stored. This estimate is the best fit of the sequence to the polyalanine segment based on a quantitative fit to the electron density. Depending on various factors (quality of the map, correctness of the initial polyalanine model, *etc.*) the correct result can appear anywhere in the sorted list.

2.3. *Combining different sequence placements*

In both qualitative and quantitative estimations, the length of the stretch under consideration is an important factor. A long stretch is more likely to ensure that the correct result comes to the top of the list. In an initial experimental density map, however, a long stretch of assigned residues may already have an error where an insertion or deletion has been missed. It is, therefore, advantageous to only work with segments having well defined density. This usually limits the procedure to secondary-structure elements. Frequently during map interpretation, the situation arises where two segments are clearly defined but are connected by a less well defined region (often a loop) for which the number of residues may be difficult to determine. The *slider_combine* command evaluates matches of multiple

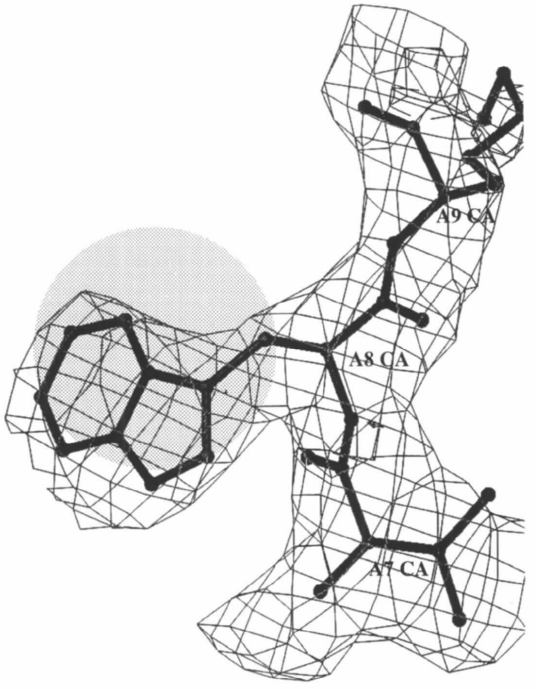

Fig. 3. A representative portion of the averaged electron-density map of P2 myelin. This is in the region around residue Trp8. The scores calculated with *rsr_seq* are shown in Fig. 4. The shaded area represents the sphere within which all grid points are used in calculating the real-space *R* factor of a threonine residue.

Residue A8

```
residue type GLY  R-factor  =  0.462
residue type ALA  R-factor  =  0.490
residue type SER  R-factor  =  0.543
residue type THR  R-factor  =  0.650
residue type VAL  R-factor  =  0.637
residue type LEU  R-factor  =  0.454
residue type ILE  R-factor  =  0.491
residue type CYS  R-factor  =  0.427
residue type ASP  R-factor  =  0.425
residue type ASN  R-factor  =  0.401
residue type GLU  R-factor  =  0.385
residue type GLN  R-factor  =  0.389
residue type PRO  R-factor  =  0.351
residue type MET  R-factor  =  0.377
residue type HIS  R-factor  =  0.287
residue type TYR  R-factor  =  0.291
residue type PHE  R-factor  =  0.260
residue type TRP  R-factor  =  0.248
residue type ARG  R-factor  =  0.516
residue type LYS  R-factor  =  0.348
```

Fig. 4. An example of the result of an *rsr_seq* calculation. The example shown here is for residue 8 in P2 myelin protein, a tryptophan residue. The scores show how well each of the 20 amino acids fits at that position.

segments that are connected by gaps. The user specifies the maximum gap size between each segment. Choosing a gap size of 0 has the same affect as combining the two segments into one long segment. Defining a very large gap, on the other hand, merely forces the first segment to occur before the second segment in the sequence.

3. Results of the automatic slider procedure

The qualitative approach has proven its usefulness in a number of *ab initio* map interpretations in our laboratory: transketolase (Lindqvist, Schneider, Ermler & Sundström, 1992), cellobiohydrolase I (Divne *et al.*, 1994), *Candida antarctica* lipase B (Uppenberg, Hansen, Patkar, & Jones, 1994) and elsewhere (Perrakis *et al.*, 1994). How well does the quantitative method perform? As test cases we have tried the procedure on two proteins whose structures were solved in our laboratory: P2 myelin protein (Jones, Bergfors, Unge & Sedzik, 1988) and lipase B from *Candida antarctica.*

3.1. *P2 myelin protein*

P2 myelin protein is a member of a family of cellular lipophilic transport proteins. This protein is made up of 131 amino-acid residues and it crystallized in space group $P2_12_12_1$ with three molecules in the asymmetric unit (Jones, Bergfors, Unge & Sedzik, 1988). The structure of P2 myelin was solved by the method of multiple isomorphous replacement and refined to a resolution of 2.7 Å (Cowan, Newcomer & Jones, 1993). The initial model was built from an electron-density map calculated with two heavy-atom derivatives (which occupied identical sites in the protein) and made use of anomalous data. This map was of surprisingly good quality, but had the sort of errors typically found in MIR maps. It has been used to teach map interpretation in numerous schools and workshops and is freely available. After the initial interpretation, the phases were further improved by five cycles of threefold averaging. The final averaged map is of excellent quality (Fig. 3) and is in fact too good for teaching purposes.

The results of the calculation are ultimately dependent on the quality of the map and the initial polyalanine model. To test the procedure we used a polyalanine model derived from the final refined model together with the averaged map. The *rsr_seq* option was first used to calculate the scores for all the residues in the protein. These scores were then used in the *slider_calc* option to evaluate how well each segment fitted the expected sequence. Segments of five, eight, ten and 15 residues were used in the calculations. For a segment length of i, therefore, we evaluated the scores of residue 1 to i, 2 to $i+1$ *etc.* for a total of $(131-i+1)$ segments. From the sorted list of scores, we can determine the position of the correct sequence and evaluate the relative scores of correct and incorrect guesses. Fig. 5 shows the percentage of correct guesses coming at the top in the list as a function of segment length. For a segment length of eight residues, 66.9% of the top scores are correct guesses, compared with 35.4% for a segment length of five residues. For a 15-residue segment, 86.3% of the guesses are correct. P2 myelin is a relatively small protein and as the size of the protein increases we may expect an increase in noise and a decrease in the percentage of correct estimations. This was simulated by adding parts of the sequence of a completely unrelated protein, the periplasmic glucose/galactose binding protein from *Salmonella typhimurium*, to the end of the P2 sequence. As shown in Fig. 5, the percentage of correct estimations decreases with increasing protein size in particular with short segment lengths. This figure clearly shows the importance of using long stretches if possible. Fig. 6(*a*) shows the correct guess ranking in the list as a function of residue number for a segment of eight residues in the P2 myelin molecule. For the majority of residues, the correct answer appears at the top of the list. Most of the wrong guesses are due to the high real-space R factors for the correct amino acids in the region between residue 73 and residue 79 in P2 myelin. This is due in part to interactions between different groups of side chains. For residue 74, which is a Thr, the density of the side-chain atoms is close to the density for the side chain of Asp76, Fig. 7. The extra envelope sphere in the Cα–Cβ direction for Thr (see *Methodology*) also includes part of the density from Asp76 and results in a high R factor (0.544) for Thr74. Asn77 has very weak density for its side chain and thus a very high R factor for Asn (0.727), but relatively low R factors for Ala and Gly (both 0.55) at this position. Arg78 has reasonably good density, and its neighbour Trp97 has even stronger density for its side chain (Fig. 8). The real-space refinement moved the side chain of

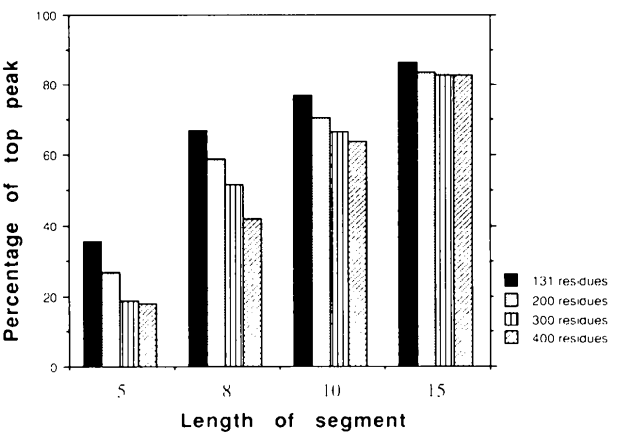

Fig. 5. The percentage of the highest scoring correct guesses as a function of segment length for P2 myelin. Also shown are the results after adding an unrelated protein sequence to the end of the P2 sequence to simulate the results for proteins 200, 300 and 400 residues long.

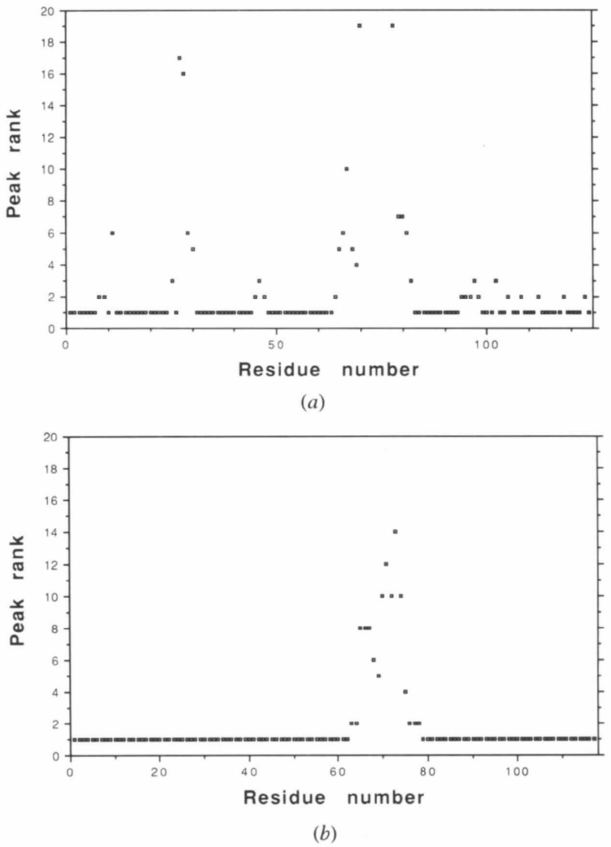

Fig. 6. The ranking of correct guesses for P2 myelin when segments of (*a*) eight and (*b*) 15 residues are used in *slider_calc*.

Arg78 into the density of Trp97 despite the fact that the refinement was carried out against the residual density. Since a polyalanine model was used in the calculation, the side-chain density for Trp97 was not subtracted and its side-chain density was still present. The residues in this region, therefore, have their correct answers quite low in the list, even when a segment length of 15 residues is used (Fig. 6*b*). In practice, attempts would be made to gradually build up the structure from the best determined indications. The residual density would then be less affected by intruding side chains as the model is developed. Therefore, once parts of the sequence have been assigned and the corresponding model has been constructed, the other parts of the sequence become easier to position.

Large aromatic residues are frequently used in map interpretation as markers for the placement of the sequence. This is due to their distinctive side-chain shape and because they frequently occur in the hydrophobic interior of proteins that usually show the clearest, best-defined electron density. The automatic slider procedure is capable of marking the possible positions of aromatic residues if these residues have clear density. Fig. 9 shows how well each residue in the chain fits a phenylalanine residue. The residues having low *R* factors in the plot are very likely to be aromatic residues, though not necessarily phenylalanine residues. Both tryptophan and tyrosine will score quite well in a density meant for

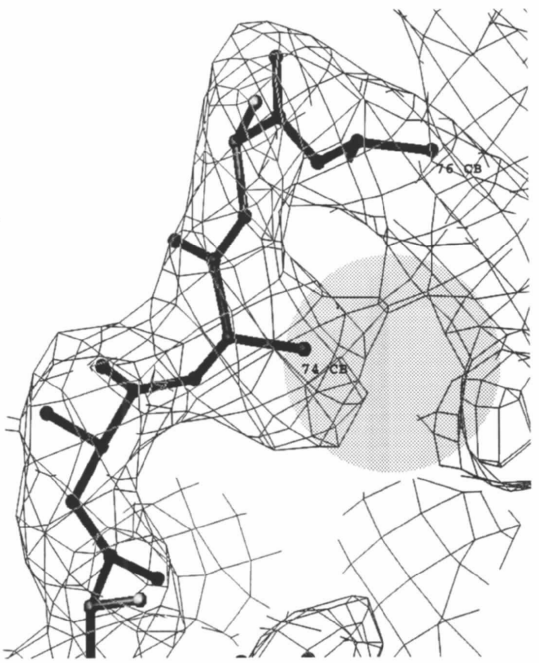

Fig. 7. The averaged electron density around residue Thr74 in P2 myelin.

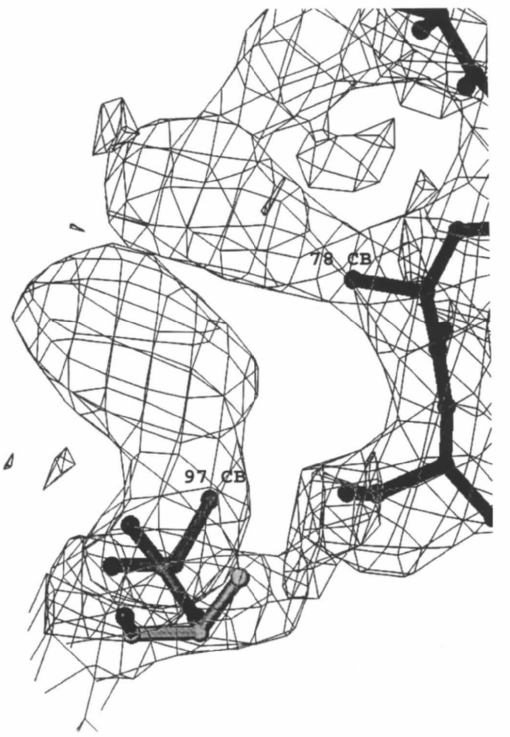

Fig. 8. The averaged electron density around residues Arg78 and Trp97 in P2 myelin.

a phenylalanine. In principle, a histidine residue should also have a low R factor, but this is not shown in the plot since P2 myelin does not contain any histidine residues. All aromatics show up as troughs as expected except for Phe57. This residue adopts two different conformations in the different NCS units which leads to poor density after averaging. The false positive at residue 113 is caused by the strong scattering of the Sδ atom of the methionine in that position. The density around Sδ has a rounded shape and a phenylalanine residue fits rather well. Plots similar to that in Fig. 9 can be made for each of the 20 amino acids and their validity in the density of this structure be evaluated. Consider tryptophan residues. The scores are evaluated placing tryptophans at each position in the chain, and the average and standard deviation are calculated. The scores are then re-evaluated in standard deviations above or below the average where the chain sequence really is a tryptophan. The averages of these scores have been plotted for the 20 amino acids in Fig. 10. Although care has to be taken because of the sampling of only 131 residues that make up this protein (there are only two tryptophans and two tyrosines in this structure, for example), the trend is quite clear and agrees well with our experience of map interpretation. The high signal for proline is a little unexpected. There are two prolines in the protein. Pro38 fits extremely well and has an R factor of 0.21, while Pro56 has a more 'ordinary' R factor of 0.33. This results in an average of score 2.3σ below the average value for proline throughout the chain.

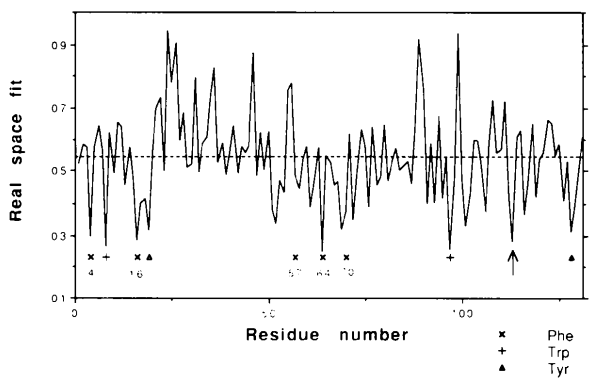

Fig. 9. Plot of how well each residue in chain A of P2 myelin fits a phenylalanine. The arrow indicates residue 113 (see text).

3.2. *Lipase B from Candida antarctica*

Lipases constitute a family of enzymes that hydrolyze triglycerides. The crystal structure of lipase B from the yeast *Candida antarctica* was solved recently (Uppenberg, Hansen, Patkar, & Jones, 1994). This lipase is an enzyme of 317 amino-acid residues. It crystallized in two crystal forms, with space groups $P2_1$ and $P2_12_12_1$. The monoclinic crystal form was initially used to solve the structure by MIR. There are two molecules in the asymmetric unit. The MIR map was calculated using several heavy-atom derivatives and it was further improved by twofold averaging at 3.5 Å and then used for the initial chain tracing. About two-thirds of the protein could be built into the averaged map, namely residues 9–184, 224–242 and 253–258. The structure of the orthorhombic crystal form was then solved by molecular replacement using this partial model, and heavy-atom sites in this crystal form could then be located and an MIR map calculated. The rest of the model was subsequently built using both MIR maps and with phase combination (Uppenberg, Hansen, Patkar, & Jones, 1994).

The averaged map, calculated at 3.5 Å resolution, in the monoclinic form has also been used as a test for the automatic slider procedure. As in the P2 myelin case, the final refined coordinates were used to produce the polyalanine chain. For clarity only residues 9–184 were used in the automatic slider test. Scores were first calculated with *rsr_seq* and then the sequence placement was made with *slider_calc*. All 317 amino acids of the

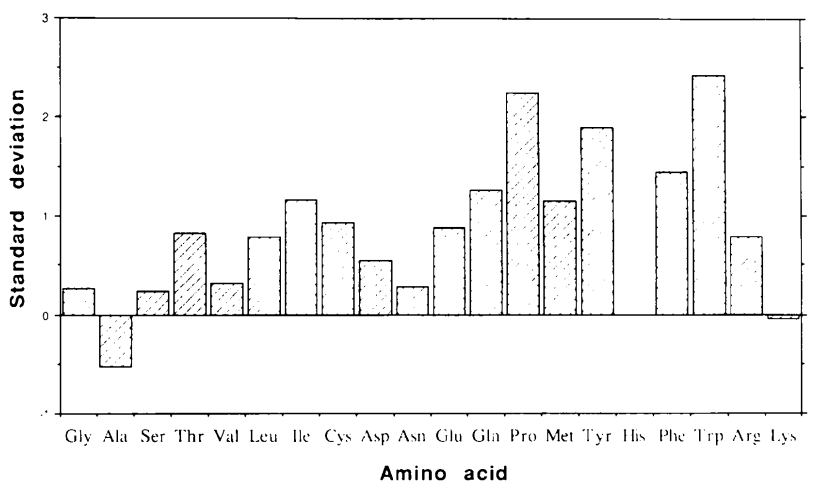

Fig. 10. Histogram of the average signal/σ values for P2 myelin. Note that there are no histidine residues in this protein.

Table 1. *The automatic slider procedure tested on some of the secondary-structure elements of lipase B*

Segments with fewer than ten residues were used in their entirety. Longer units were split into ten-residue segments.

Secondary structure		Range First and last residue No.	Segment length (residues)	Correct guess as peak No.
Helix	No. 1	13–18	6	20
Strand	No. 1	20–22	3	6
Strand	No. 2	33–37	5	5
Helix	No. 2	44–53	10	1
		45–54	10	1
		46–55	10	1
		47–56	10	1
Strand	No.3	62–66	5	6
Helix	No. 3	76–85	10	1
		77–86	10	1
		78–87	10	1
		79–88	10	1
		80–89	10	2
		81–90	10	1
		82–91	10	1
Strand	No. 4	99–104	6	2
Helix	No. 4	106–115	10	1
		107–116	10	2
		108–117	10	2
Strand	No. 5	125–131	7	65
Helix	No. 5	141–148	8	196
Helix	No. 6	151–158	9	1
Helix	No. 7	162–169	8	82
Strand	No. 6	179–183	5	56

lipase B sequence were used. For segments of length of eight, ten and 15 residues the percentages of correct highest scores were 26.6, 29.9 and 43.2%, respectively. Fig. 11 shows the correct guess ranking in the list as a function of residue number when the segment length is ten residues. These results are clearly worse than those obtained for P2 myelin. This was not unexpected since the averaged map for P2 myelin was judged to be significantly better than the averaged map of lipase B. The initial mask used in the lipase averaging contained errors and resulted in some parts of the density in the averaged map being chopped off. The scores for these parts are bad as a consequence. We therefore evaluated

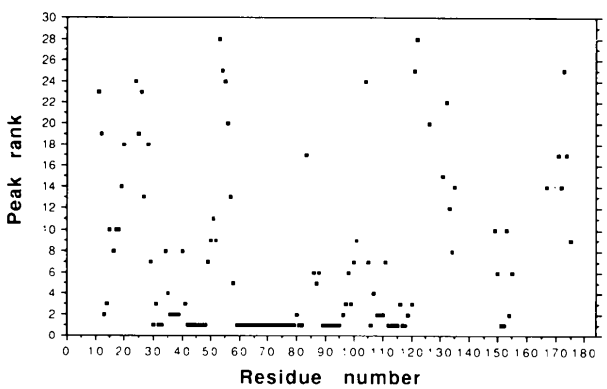

Fig. 11. The ranking of correct guesses for lipase B when a segment length of ten residues is used.

the strategy of identifying those parts of the structure that have the clearest density and do not possibly have insertions or deletions within them. Table 1 shows the results of the automatic slider procedure tested on only the secondary-structure elements. This shows a significant improvement over the results where all residues are used. The procedure is able to identify most of the sequence placement when the secondary structure units have a reasonable length (eight to ten residues), especially for helices 2, 3, 4 and 6. For the shorter secondary-structure parts (between three and six residues in length) the procedure gives reasonable rankings for the correct guesses although there are exceptions. Helix 5 has its correct ranking down at number 196! This region is on the surface of the protein and the mask was too small around this region which made the averaged density very weak for the whole helix. The poor result for strand 5 is caused by density associated with Ala130. There is no density for the Cβ atom of the residue and results in an R factor of 0.925 for Ala, since the side chain contains only Cβ for Ala. If only residues 125–129 are used, the correct guess ranks number 12.

4. Discussion

Scanning a protein sequence and trying to find matches with the density is a difficult and laborious task even for an experienced crystallographer. Tools have therefore been developed to help identify the sequence placement during map interpretation. They can be used

both for qualitative and quantitative estimations. The *slider_guess* option compares the crystallographer's guess of the amino-acid type to the actual sequence *via* a scoring matrix. This enables the crystallographer to concentrate on portions of the map and to try different hypotheses which the computer scans against the sequence to find possible matches. The scoring matrix is used to assess the matching and it can be modified by users as they will. The automatic slider option gives an objective measure of how well the density of a certain residue fits each of the 20 amino acids. This is performed entirely without manual intervention. The calculation of *rsr_seq* for P2 myelin with 131 residues took about 7 h of CPU time on a Silicon Graphics Indigo workstation. With the introduction of ever more powerful computers, this calculation time will not be a problem.

Results of the automatic slider option on two test cases are presented. For P2 myelin, the map was of very good quality and the procedure was able to find the correct sequence for the majority of the residues when the segment length was about eight residues, a value routinely achievable in map interpretation. In the lipase B case, the test was performed on the map that was actually used for the structure determination. This map was of the more moderate quality routinely obtained by the MIR method and standard density-modification procedures. Even in this case, our procedure showed promising results. It was able to identify the correct sequences for most parts of the secondary structure provided that the density was reasonably good and the segments were of reasonable length. With the inclusion of *slider_comb* the procedure is potentially even more powerful. The method described here works for both moderate resolution (2.7 Å in the P2 myelin case) and at lower resolution (3.5 Å in the lipase B case).

In the calculations for both P2 myelin and lipase B, a polyalanine model derived from the final refined model were used. To evaluate how much an preliminary polyalanine model would affect the results, a ten-residue long segment corresponding to a β-strand in P2 myelin was manually built with the *baton* option in *O* (Jones & Kjeldgaard, 1996) and manually refitted to the averaged map. Calculations using this model produced very similar results to those obtained with the refined model.

A modified envelope has been used in calculating the real-space *R* factor to overcome the problem of small side chains scoring well in density meant for a bigger amino acid. Without this modification, all the bulky residues with good density will also show good scores for residues with small side chains and make the placement of the sequence somewhat more difficult. This technique, however, does result in some drawbacks because of interacting side chains. As shown in Fig. 7, for example, Thr74 in P2 myelin has good density, but *rsr_seq* gave an *R* factor of 0.544 for Thr as a consequence of the envelope including part of the density for Asp76. Different formulations of the envelope will have both pros and cons. We believe the envelope used here is most suitable for the automatic slider procedure.

The use of the automatic slider procedure presented here is not limited to sequence placement. It can also be used, for example, to find out-of-register errors in an initial model. This can be performed by mutating all residues of the part in question to alanine and then calculating scores with *rsr_seq*. The whole sequence can then be scanned to find the best matches and possible errors can be detected.

This work was supported by Uppsala University. We would like to thank Dr Gerard Kleywegt for critical reading of this manuscript. The improvements suggested by the referees are gratefully acknowledged.

References

Cowan, S. W., Newcomer, M. E. & Jones, T. A. (1993). *J. Mol. Biol.* **230**, 1225–1246.

Deisenhofer, J. & Steigemann, W. (1975). *Acta Cryst.* B**31**, 238–250.

Diamond, R. (1971). *Acta Cryst.* A**27**, 436–453.

Divne, C., Ståhlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G. Knowles, J. K. C., Teeri, T. T. & Jones, T. A. (1994). *Science*, **265**, 524–528.

Jones, T. A. & Liljas, L. (1984). *Acta Cryst.* A**40**, 50–57.

Jones, T. A., Bergfors, T., Unge, T. & Sedzik, J. (1988). *EMBO J.* **7**, 1597–1604

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

Jones, T. A. & Kjeldgaard, M. (1996). *Methods Enzymol.* In the press.

Lindqvist, Y., Schneider, G., Ermler, U. & Sundström, M. (1992). *EMBO J.* **11**, 2373–2379

Perrakis, A., Tews, I., Dauter, Z., Oppenheim, A. B., Chet, I., Wilson, K. S. & Vorgias, C. E. (1994). *Structure*, **2**, 1169–1180.

Ponder, J. W. & Richards, F. M. (1987). *J. Mol. Biol.* **193**, 775–791.

Uppenberg, J., Hansen, M. T., Patkar, S. & Jones, T. A. (1994). *Structure*, **2**, 293–308.