Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

ACORN in *CCP*4 and its applications

Jia-xing Yao

York Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, York YO10 5DD, England

Correspondence e-mail: yao@ysbl.york.ac.uk

Received 29 January 2002
Accepted 12 September 2002

ACORN is a comprehensive and efficient phasing procedure for the determination of protein structures when atomic resolution data are available. Reliable phases can be developed from a fragment composed of a small percentage (less than 5%) of the scattering matter of the unit cell. For example, *ACORN* has been used to solve a structure of 1093 atoms from only one S atom. The map from *ACORN* typically reveals the 90% of whole structure. The initial model can be automatically built using *ARP*/*wARP* or *QUANTA*. *ACORN* can also be used to determine substructures of heavy atoms or anomalously scattering atoms from SAD or MAD data at much lower resolutions such as 3.5 Å. Some test results of using *ACORN* on known structures are presented here. *ACORN* has solved four new protein structures by users in the UK and France, and some results kindly provided by them are described in this paper.

1. Introduction

All proteins are built up from linear polymers of amino acids and have some common features in their configurations. For example, α-helices and β-sheets appear in many protein structures and can be taken as starting fragments for structure solution. The configurations of α-helices in protein structures are all close to a standard α-helix such as that found in the fragment library in *CCP*4 (Cowtan, 2001). The β-sheet is more flexible and no single standard exists, but a few commonly occurring β-sheets can be used (Oldfield, 2001). Since the required size of fragment is very small and the Protein Data Bank (PDB) is steadily becoming larger, it is normally possible to find a small matching motif from the structures in the PDB by sequence-searching approaches. S atoms commonly occur in proteins and Dauter *et al.* (1999, 2000) have shown that it is possible to determine the positions of S atoms from anomalous differences using direct-methods programs such as *RANTAN* (Yao, 1981), *SHELX* (Sheldrick & Gould, 1995) or *SnB* (Weeks *et al.*, 1994). Atoms heavier than sulfur can also be found in proteins such as metalloproteins or SeMet proteins, in which S atoms are replaced by Se atoms. It will be shown that using a single random-atom search it is possible to locate one of the heavy atoms from which the native data can be phased. Another common approach is to determine a substructure, *e.g.* one containing S or Se atoms, with anomalous scattering data. *ACORN* (Foadi *et al.*, 2000) can use all such information to start the phasing process. In order to be used for whole protein structures *ACORN* needs data at atomic resolution, higher than 1.3 Å, but for the determination of substructures the resolution can be as low as 3.5 Å.

© 2002 International Union of Crystallography
Printed in Denmark – all rights reserved

Mean phase errors given in this paper are for the 'strong' reflections only (see below) and the CPU times are on a Silicon Graphics O2. All reflection data input to *ACORN* are in *CCP*4 MTZ format and fragment data are in standard PDB format. The outputs of *ACORN* are phased reflection data in *CCP*4 MTZ format.

2. The general structure of *ACORN*

The reflections are divided into three groups (strong, medium and weak) according their normalized structure-factor (E) values. The strong reflections ($E > 1.2$) are used in the phase refinement by the dynamic density modification (DDM) and Patterson superposition (SUPP) procedures and both strong and weak reflections ($E < 0.1$) are used in Sayre-equation refinement (SER). The medium reflections ($0.1 < E < 1.2$) are used to calculate a correlation coefficient (CC) for each potential solution of DDM. These groups are default settings and can be changed by the user.

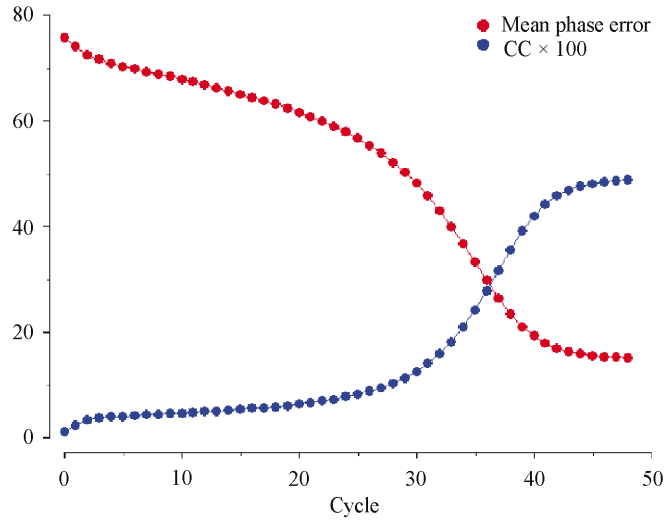

Figure 1
CC and mean phase error for penicillopepsin (PDB code 1bxo) against the cycle number of DDM.

An important component of *ACORN* is a CC that describes the extent to which the magnitudes of the calculated normalized structure factors (E_c) resemble the observed normalized structure-factor amplitudes (E_o). A fragment in a particular position and orientation in the unit cell will have an associated set of structure factors and the CC will be expressed by

$$\mathrm{CC} = \frac{\langle |E_c E_o| \rangle - \langle |E_c| \rangle \langle |E_o| \rangle}{\sigma(E_c)\sigma(E_o)},$$

where

$$\sigma = (\langle |E^2| \rangle - \langle |E| \rangle^2)^{1/2}.$$

E_c and CC values are calculated from the starting fragment for all reflections to find the correct orientation and position in molecular replacement (MR) or random MR or for single random-atom searching. In phase refinement E_c and CC values are calculated from the modified map for medium reflections, which are not used for computing the map, to indicate solutions of DDM. Fig. 1 shows the CC for the medium reflections is strongly correlated to the mean phase error and does indicate a correct solution clearly.

The *ACORN* procedure, as implemented in *CCP*4, is divided into two parts, *ACORN-MR* and *ACORN-PHASE*, as illustrated in the flow diagram in Fig. 2 and described below. The first part, *ACORN-MR*, deals with finding the position of a fragment of the structure, even a single atom, that provides an initial set of estimated phases. This set is passed into *ACORN-PHASE*, where phase refinement by a number of real-space processes is performed.

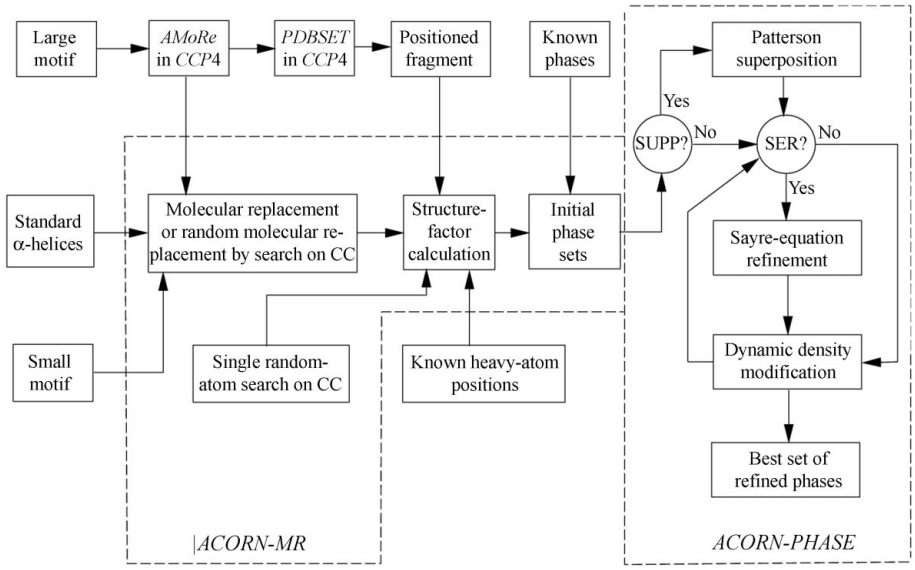

Figure 2
Flow diagram for *ACORN* in *CCP*4.

3. *ACORN-MR*

3.1. Single random-atom searching

In order to locate a single atom, this approach randomly generates thousands of positions in the asymmetric unit. E_c values and corresponding CCs are calculated for all reflections for each random position. The program saves the 1000 sets with highest CC as starting points. *ACORN-PHASE* will refine them in order until the solution is found. Normally, the solution is found in the top 100 sets. This approach can be used to determine a native protein structure from atomic resolution data if the structure contains at least one heavy atom (sulfur or heavier). Provided the heavy atom accounts for at least 5% of the total scattering power

of the structure, then it will usually provide good enough initial phases for *ACORN-PHASE* to successfully refine the single-atom fragment phases. However, experience has shown that a solution may sometimes be found even if the scattering power of the single atom (or fragment) is as low as 1%. The *ACORN* approach can also used to determine substructures with anomalous scattering data or isomorphous replacement data at much lower resolutions. The smaller number of atoms in the substructure means that one random atom has proportionately more scattering power in a substructure than in a protein structure but, since the quality of anomalous scattering data is worse than that of native protein data, the CC will be smaller for the correct position.

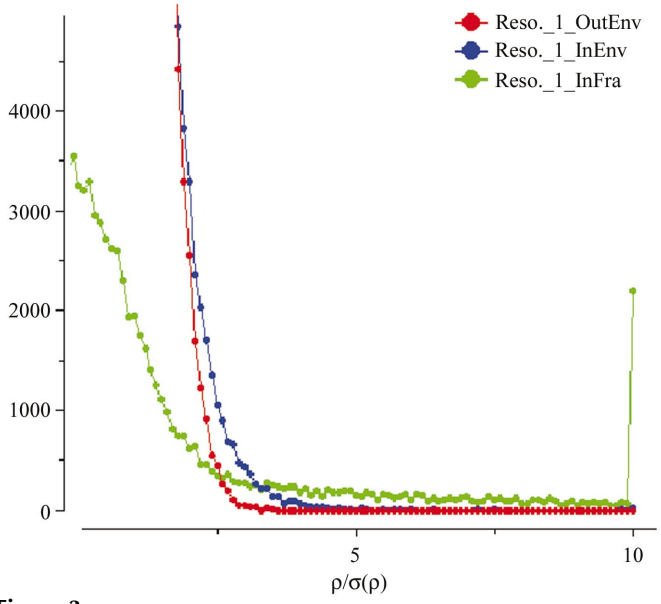

Figure 3
Histogram of an initial map calculated from a fragment.

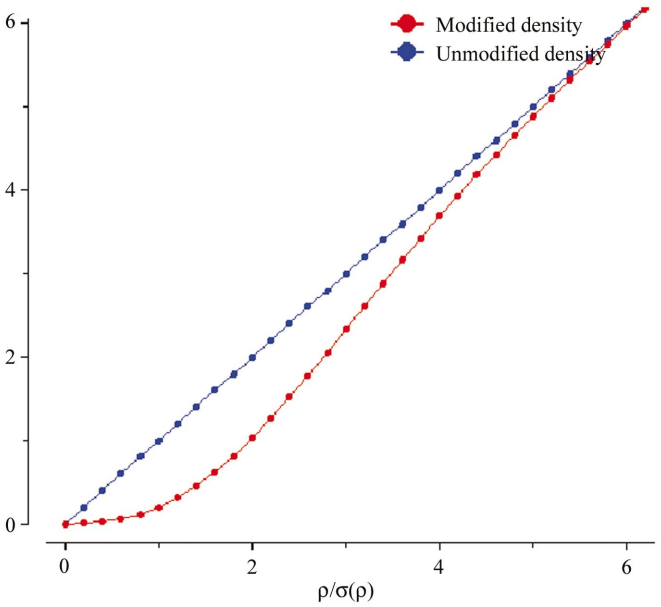

Figure 4
Density-modification curve in DDM.

3.2. Molecular replacement (MR) and random MR searching

For a many-atom fragment it is necessary to find both the position of a representative point and the orientation. *ACORN* splits the required six-dimensional search into two consecutive three-dimensional searches for reasons of speed (Rossmann & Blow, 1962). Therefore, MR (or random MR) first explores rotational space and then, for the highest values of CC from rotation alone, finds positions for different translations. MR is a step-by-step search process on a grid based on the rotation angles and random MR just searches a predetermined number of rotation angles generated randomly. To save CPU time, the process is divided into two parts. In the first part, the CC is calculated for only 10% of reflections for each rotation angle or each translation shift and

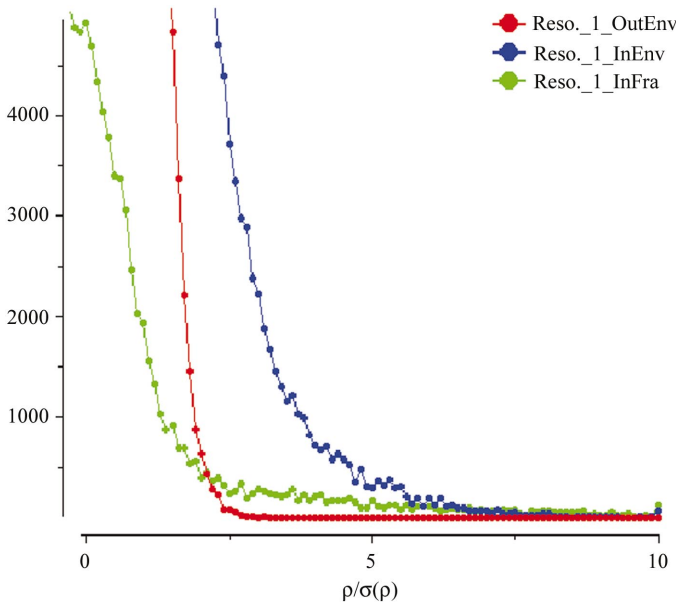

Figure 5
Histogram of a map modified by one cycle of DDM.

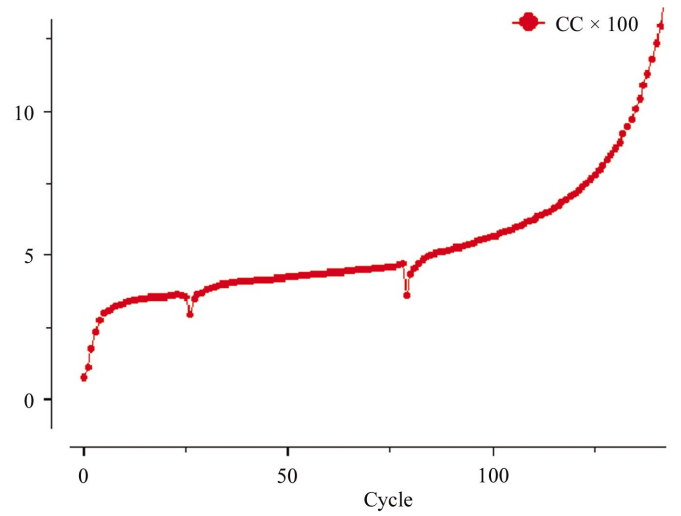

Figure 6
Combination of DDM with SER to re-solve the structure of catalase.

the top 1000 sets with highest CCs are saved. The CC is then recalculated for all reflections and the top 1000 sets are ranked on the basis of the new CC. Normally, the correct orientation and position is at the top of the list.

4. *ACORN-PHASE*

4.1. Dynamic density modification (DDM)

DDM is the main phase-refinement approach in *ACORN-PHASE* and can develop an initial phase set with a mean phase error approaching 80° to a final mean phase error around 15°. Each cycle of DDM starts from a set of phases with weights and calculates a weighted E map with the strong E values. The map is modified and back-transformed by FFT to obtain a new set of structure factors. At the end of one cycle a new set of phases with weights is calculated for the next cycle and the CC is calculated for medium reflections to check if a solution has been found.

The initial map from the starting fragment contains a lot of noise. Fig. 3 shows a histogram of an initial map from a fragment for a known structure.

Since the test is on a known structure, one can divide the map into three parts: red densities are outside the protein region (or envelope), green densities are inside the fragment region and blue densities are in the protein region but outside the fragment region. The last category of densities is the one we want to develop and will grow through the cycles of DDM as the phases are evolving towards the true solution. The protein region contains densities higher than the 'solvent' region densities and densities over $2\sigma(\rho)$ are the densities we want to enhance. The fragment region densities are what we put in and we do not want this part of the density to dominate the density-modification process, so the highest fragment-region densities are truncated to be comparable to protein-region densities. Therefore, DDM is designed to modify the densities in three steps:

$$\begin{cases} \rho' = 0 & \text{if } \rho<0 \\ \rho' = \rho\tanh\{0.2[\rho/\sigma(\rho)]^{3/2}\} & \text{if } \rho>0 \\ \rho' = kn\sigma(\rho) & \text{if } \rho'>kn\sigma(\rho), \end{cases}$$

(i) It sets all negative densities to zero.

(ii) It modifies the positive densities according the ratio $\rho/\sigma(\rho)$.

(iii) It truncates the modified densities to a value of $kn\sigma(\rho)$, where k is a constant given by the user (default value is 3); n is the cycle number of DDM, but after five cycles n always equals 5.

These properties can be seen clearly from the DDM curve in Fig. 4. It will depress the densities around and less than $1\sigma(\rho)$, which are mostly noise, and truncate high densities to $3\sigma(\rho)$ in first cycle to relatively enhance the protein region densities in the range 2–$3\sigma(\rho)$. In the second cycle it will truncate at $6\sigma(\rho)$ and after five cycles it will always truncate at $15\sigma(\rho)$ (default). The user can change this level by giving a different value of k.

Fig. 5 shows the histogram of the modified map after one cycle of DDM applied to the map corresponding to Fig. 3. The protein region (blue) densities are greatly increased and although the fragment region (green) densities are still there, they are much lower than in the initial map. The solvent region (red) densities have not changed very much. As the protein region densities become larger, the map will finally reveal the structure.

Since DDM modifies a map solely according the ratio $\rho/\sigma(\rho)$, DDM can be used to modify not only an E map or an F map, but also a Patterson map or a sharpened Patterson map, for example in the Patterson superposition method in *ACORN-PHASE* (see below).

4.2. Real-space Sayre equation refinement (SER)

It has been found that a couple of cycles of Sayre equation refinement, SER, can help DDM to reach a global rather than local minimum (Foadi *et al.*, 2000). When the CC for the medium-strength reflections no longer increases, or the phases change little on use of DDM, then one or two cycles of SER can be employed. In fact, the application of SER sometimes makes the CC lower, but thereafter DDM can lead to a higher CC than previously. SER is a real-space procedure carried out using the fast Fourier transform (FFT) and inverse FFT. No phase relationships are involved, so there is no practical limit to the number of reflections that can be included. Fig. 6 shows how the combination of DDM with SER re-solved the known structure of catalase starting from nine S-atom positions. SER was used for two cycles at cycle 25 when the CC was not increasing and was used again for two cycles at cycle 77 when the average phase change was less than 0.5°. Thereafter, the CC after each cycle of DDM always increased until it reached 0.49 and the mean phase error was 13°.

4.3. Patterson superposition method (SUPP)

A sum function is used for Patterson superposition on each atom in the fragment. The summation of each origin-shift Patterson map can be carried out in reciprocal space by the summation of contributions from each atom in the fragment. Therefore, a weighted semi-sharpened Patterson superposition map is calculated using Fourier coefficients $|E_o||F_o||E_c|$ with phases from the fragment. The map shows high peaks corresponding to the fragment and some additional low densities that give a small but valuable phase improvement. The map is then modified by one cycle of DDM. The application of SUPP can improve the initial phases from the starting fragment by 1 or 2° when the fragment contains more than ten atoms.

5. Testing *ACORN* on known structures

A number of known structures have been tested using *ACORN* (Foadi *et al.*, 2000) and some details are given here for three examples. All three sets of data extend to resolutions better than 1.0 Å, but in the tests described here have been restricted to 1.0 Å resolution.

5.1. Lysozyme

This $P1$ crystal form of lysozyme (PDB code 3lzt) contains 1230 non-H atoms (Walsh *et al.*, 1998) and could be solved by *ACORN* from three kinds of starting fragments. There are ten S atoms in this structure, but only two of them were needed as a starting fragment, giving an initial mean phase error of 75°. A total of 19 min was required for *ACORN-MR* and *ACORN-PHASE* to solve this structure, giving a final mean phase error of 15°. A single random-atom search could also solve this structure together with another atom at the origin (because the space group $P1$ needs at least two atoms to fix the origin). In this case, the mean phase error went from 79 to 19° after 6.4 h CPU time. The extra CPU time is compensated by not needing to collect the anomalous scattering data and locate the positions of S atoms. Since lysozyme contains α-helices, a standard α-helix from the *CCP*4 fragment library could also be used as a starting fragment. Ten alanine residues (50 atoms) were taken from the library and random MR was used to find the correct orientation, no translation function being needed for space group $P1$. The mean phase error went from 76 to 15° using 15.3 h of CPU time. A second lysozyme study with 1093 non-H atoms and space group $P4_32_12$ (PDB code 1bwi) was solved by a single random-atom search approach with a 15° mean phase error after 14.5 h CPU time.

5.2. Penicillopepsin (PDB code 1bxo)

Penicillopepsin (Ding *et al.*, 1998) contains 2977 non-H atoms in the asymmetric unit with space group $C2$. This example illustrates the solution of a structure using *ACORN* where the only available information was a sequence. Sequence alignment against structures in the PDB was performed and 57 residues (398 atoms) were taken from a related structure with PDB code 1er8 (Pearl & Blundell, 1984). The first test assumed space group $P1$, reduced from $C2$, so that no translation function was needed. The correct orientation was found by *ACORN-MR* and provided an initial 76° mean phase error. *ACORN-PHASE* refined the initial phases to a mean phase error of 15°. The *ACORN* map was in $P1$ and the origin needed to be shifted to return to $C2$. The second test used the proper space group $C2$ and *ACORN-MR* found the correct orientation and position that resulted in a better initial mean phase error of 71° and a final mean phase error of 14°. Fig. 7 shows an E map from *ACORN* that was in very good agreement with the final model. *QUANTA* built the structure automatically from this map taking less than 20 s (Oldfield, 2002).

5.3. Catalase

This is the largest structure (Murshudov *et al.*, 1992) that has been tested on *ACORN*. There are 4762 non-H atoms (502 residues plus a haem group) in the asymmetric unit with space group $P4_22_12$. The iron and sulfur positions are easily determined by direct methods or Patterson searches using anomalous scattering data. The data provided by Murshudov (unpublished results) reaches a resolution of 0.88 Å with excellent quality. *ACORN* used only data extending to 1.0 Å resolution and solved this structure using DDM and SER starting from nine S atoms or using DDM alone starting from the iron-containing haem group (43 atoms in total). The final mean phase error was 13° for both starting fragments.

6. New structures solved by *ACORN*

6.1. Metalloproteinase deuterolysin (PDB code 1eb6)

The data of metalloproteinase deuterolysin (McAuley *et al.*, 2001) were collected to 1.0 Å resolution with synchrotron radiation. The structure contains 177 residues and one Zn atom in the asymmetric unit with space group $P2_1$. The position of the Zn atom was determined either from an anomalous difference Patterson or from a sharpened normal Patterson with native data. Starting from the zinc position, DDM solved the structure using 168.9 s CPU time. The single random-atom

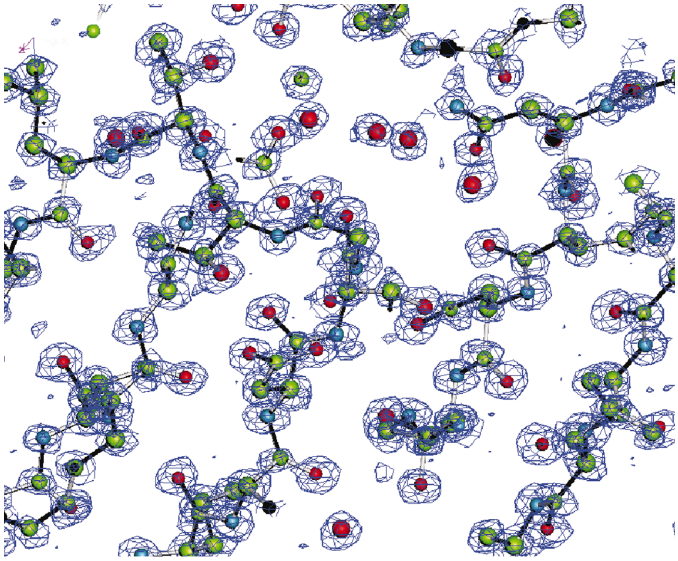

Figure 7
E map calculated from *ACORN* phases and weights for penicillopepsin.

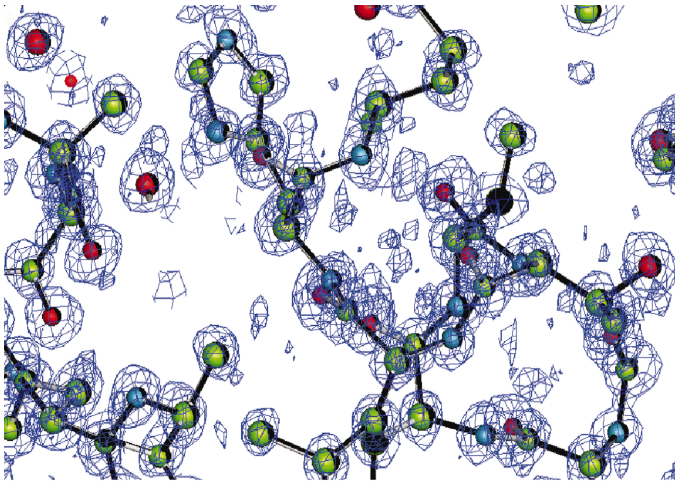

Figure 8
E map calculated from *ACORN* phases and weights for metalloproteinase deuterolysin.

search procedure in *ACORN* also solved this structure using 494.1 s of CPU time using only native protein data. The searching region is restricted to the plane $y = 0$ because the origin can be anywhere along the y axis for this space group. The structure contained α-helices and could also be solved by *ACORN* starting from ten residues of a standard α-helix with random MR giving the correct orientation and position. DDM provided the same quality phases as for the other starting points. The model was built up automatically using *ARP*/*wARP* (Perrakis *et al.*, 1999). Fig. 8 is an *E* map overlaid on the refined model. Most atoms appear in the map very clearly. The whole process, starting from the processed and merged data and ending with a refined model, required less than 6 h computation time.

6.2. C1 subunit of α-crustacyanin (PDB code 1i4u)

This structure (Gordon *et al.*, 2001) contained 362 residues (3382 non-H atoms) in the asymmetric unit with space group $P2_12_12_1$. The native data were collected to 1.15 Å resolution and anomalous scattering data to 1.77 Å. The substructure of 12 S atoms was determined using *SnB* (Weeks & Miller, 1999) with anomalous scattering data. *SnB* and other methods were tried to solve this structure from the positions of S atoms, but could not obtain a density map that was good enough to be interpretable. The same positions of S atoms were taken as a starting fragment into *ACORN* and the initial phases were refined using DDM. A density map of excellent quality was provided immediately. The major part of the atomic model for the protein was built automatically using *wARP*.

6.3. A protein structure solved by Esnouf (2001)

The determination of this structure followed a similar pattern to that in §6.2. The structure contained 164 residues (1318 non-H atoms) in the asymmetric unit with space group $P2_12_12_1$. The data were collected at 0.9 Å resolution. *SnB* was used and located 16 S atoms, but could not give any further progress. *ACORN* started from these S positions and solved the structure by DDM using about 2 min CPU time.

6.4. A SeMet protein structure solved by Davies (2001)

This was an SeMet protein structure of 125 residues with space group $P2_12_12_1$ with data at 1.0 Å resolution. A single Se atom was located using the Patterson method and DDM in *ACORN* solved the structure in a few minutes.

7. Determining substructures

Substructures can be solved using either direct or Patterson methods. It is difficult to locate S atoms from anomalous scattering data because the anomalous signal from S atoms is very weak. Therefore, the collection of good-quality anomalous scattering data is the most critical part of the process. If MAD data are available then better estimates of contributions from anomalous scattering atoms can be obtained using *REVISE* in *CCP*4. Alternatively, single random-atom searching in *ACORN* can quickly locate the heavy atoms from anomalous scattering data (SAD or MAD) or isomorphous replacement data (SIR or MIR). A test was carried out on a known structure (Muchmore *et al.*, 1998) with MAD data from 21 Se atoms. The reflections used in *ACORN* were only those in the 10.0–3.5 Å resolution range. *REVISE* was first applied to obtain improved estimates of anomalous contributions, F_m, from Se atoms. The *E* values were calculated from F_m using *ECALC* in *CCP*4 and input to *ACORN*. Single random-atom searching gave a good starting position and DDM provided a solution using only 1 h of CPU time. The *E* map showed 18 Se atoms at the top of the peak list; the other three Se atoms were in lower positions because they were disordered.

Dodson (2002) has performed more tests using *ACORN* successfully to determine substructures from ten P to 166 Se atoms with SAD or MAD data.

8. Discussion

The results of applying *ACORN* to solve known and unknown structures have proved that *ACORN* is a flexible and efficient procedure to tackle the phase problem in the determination of protein structures with atomic resolution data and substructures with SAD, MAD, SIR or MIR data at much lower resolutions. DDM is a fast and powerful density-modification approach and it can naturally be extended to work at lower resolutions (from 1.5 to 2.0 Å) with more sophisticated algorithms. *ACORN-MR* is based on medium-strength reflection correlation-coefficient searching based on *E* values, a process that is slow compared with *AMoRe* (Navaza, 2001). Improvements are required to speed up this process in *ACORN-MR*.

I gratefully acknowledge my collaboration with Professors Michael Woolfson, Keith Wilson and Eleanor Dodson. Their input has been essential in making *ACORN* available for general use. I would also like to thank Drs Katherine McAuley, Elspeth Gordon, Robert Esnouf and Gideon Davies for allowing me to present the results of using *ACORN* to solve their structures. My thanks are also due to Drs James Foadi and Zheng Chao-de for many discussions. I appreciate very much the support from BBSRC for three years and from CCP4 for one year.

References

Cowtan, K. (2001). *Acta Cryst.* D**57**, 1435–1444.
Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.
Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Acta Cryst.* D**56**, 232–237.
Davies, G. (2001). Private communication.
Ding, J., Frasere, M. E., Meyer, J. H. & Bartlett, P. A. (1998). *J. Am. Chem. Soc.* **120**, 4610–4621.
Dodson, E. J. (2002). In preparation.
Esnouf, R. (2001). Private communication.
Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Jia-xing, Y. & Chao-de, Z. (2000). *Acta Cryst.* D**56**, 1137–1147.
Gordon, E. J., Gordon, A. L., McSweeney, S. & Zagalsky, P. F. (2001). *Acta Cryst.* D**57**, 1230–1237.

McAuley, K. E., Yao, J. X., Dodson, E. J., Lehmbeck, J., Astergaard, P. R. & Wilson, K. S. (2001). *Acta Cryst.* D**57**, 1571–1578.
Muchmore, C. R., Krahn, J. M., Kim, J. H., Zalkin, H. & Smith, J. L. (1998). *Protein Sci.* **7**, 39–51.
Murshudov, G. N., Melik-Adamyan, W. R., Grebenko, A. I., Barynin, V. V., Vagin, A. A., Vainshtein, B. K., Dauter, Z. & Wilson, K. S. (1992). *FEBS Lett.* **302**, 127–131.
Navaza, J. (2001). *Acta Cryst.* D**57**, 1367–1372.
Oldfield, T. (2001*a*). *Acta Cryst.* D**57**, 1421–1427.
Oldfield, T. (2002). *Acta Cryst.* D**58**, 963–967
Pearl, L. & Blundell, T. (1984). *FEBS Lett.* **174**, 96–101.
Perrakis, A., Morris, R. M. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.
Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423–431.
Walsh, M. A., Schneider, T. R., Sieker, L. C., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1998). *Acta Cryst.* D**54**, 522–546.
Weeks, C. M., De Titta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.
Weeks, C. M. & Miller, R. (1999). *Acta Cryst.* D**55**, 492–500.
Yao, J. X. (1981). *Acta Cryst.* A**37**, 642–644.