Acta Cryst. (1998). D**54**, 828–833

STEP – a Trial-and-Error Procedure for Crystal Structure Determination. II. The Determination of Two Small Protein Structures

LI LIPU,* YI JIAN, ZHANG XIAOFENG AND HOU YONGGENG

The State Key Laboratory for Structural Chemistry of Unstable and Stable Species, Institute of Chemistry, Chinese Academy of Sciences, Beijing 10080, People's Republic of China. E-mail: houyg@infoc3.icas.ac.cn

(*Received 15 October 1997; accepted 24 December 1997*)

Abstract

This paper describes the difficulties in the process of using the trial-and-error *SYSTEM*90 program to determine *ab initio* the structures of two small proteins App [Woolfson & Yao (1990). *Acta Cryst.* A**46**, 409–413] and rubredoxin [Sheldrick *et al.* (1993). *Acta Cryst.* D**49**, 18–23] with high-resolution data. Some strategies for overcoming the difficulties are discussed and the upgraded *SYSTEM*95 program was used successfully to determine the two structures. The most characteristic feature of this structure-determination process is that the two proteins are treated as unknown structures with only their chemical compositions and high-resolution data sets known. A new figure of merit R(sc), replacing the old figure of merit, XDFOM, is quite effective in picking out a good set of phases in the multi-solution stage when the phases are overconsistent. Controlling the Fourier recycling technique and the residuals can separate the mixture of structures and the enantiomorph and finally give one absolute structure. The results are compared with known structures to verify their reliability.

1. Introduction

The combination of direct methods and computer techniques has changed the face of small-molecule X-ray crystallography and has become the main method in the determination of small structures. In biology, it has become important to obtain structural information on substances such as oligonucleotides and peptides, which play a major role in the origin of life and physiological action. These substances are larger than the largest small molecules but smaller than macromolecules, and it is possible to acquire high-resolution data from them with recent developments in crystal-growth and data-collection techniques. The possibility of applying direct methods to macromolecular structures has become a hot topic in the field of crystallography. Woolfson & Yao (1990) solved the structure of the 36-residue protein App with 0.98 Å resolution data by the use of the *SAYTAN* program. Sheldrick *et al.* (1993) solved two small proteins, crambin and rubredoxin, using phase-annealing and Patterson methods. Weeks *et al.* (1995) solved the crambin structure directly by the use of the program *Shake-and-Bake* (*SnB*) with 0.83 Å resolution data. Full least-squares matrix refinement of this kind of molecular structure is also possible. This paper describes a different approach (the *SYSTEM*95 program) to solve the structures of two small proteins, App and rubredoxin, using high-resolution data.

2. *SYSTEM*95 program generation

*SYSTEM*90 (Hou *et al.*, 1990) is a powerful program for solving small-molecule structures and, in theory, is also applicable to larger structures with high-resolution data. In the *SYSTEM*90 program, a set of strong reflections sufficient to solve the structure is divided into a hierarchy of smaller soluble and connected subsystems. Within each subsystem, the reflections are required to be well connected with each other, given that the phases of all reflections in the previous subsystems are known. A trial-and-error procedure is then employed to provide an approximate solution to an overdetermined set of equations. Subsequently, phases are refined by one of two available tangent formulae and then assessed for plausibility by figures of merit. However, some difficulties arose in the *ab initio* solution of the structure of the proteins App and rubredoxin. These problems are discussed and strategies are designed to solve them.

2.1. *Change of the size of the subsystem*

The number of reflections and Σ_2 relationships in the first subsystem is so great for the two protein structures that the phases in the subsystem are difficult to determine and require a great deal of computer time. The size of the subsystem is controlled by 'ap' such that all the reflections in the first subsystem must have $\langle\alpha(\mathbf{h})\rangle$ [see (2)] greater than ap (default value 5). The smaller the value of ap, the smaller the number of reflections contained in the first subsystem. Therefore, by making the default value of ap (in the range 1.0–5.0) decrease as the size of the determined structure increases, we find the phases are successfully determined in the system and much computer time is saved.

ISSN 0907-4449

2.2. *Improvement of the weighting scheme*

*SYSTEM*90 used a new weighted tangent formula to refine the phases which, in early 1991, was shown to solve the App structure successfully. The form of the weighting scheme is

$$W(\mathbf{h}) = \min\left[\frac{\alpha(\mathbf{h})}{\langle\alpha(\mathbf{h})\rangle} + C_W, 1.0, \frac{\langle\alpha(\mathbf{h})\rangle}{\alpha(\mathbf{h})} + C_W\right], \quad (1)$$

where

$$\tan(\varphi_{\mathbf{h}}) = \frac{\sum_{\mathbf{k}} \kappa(\mathbf{h},\mathbf{k})\sin(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})}{\sum_{\mathbf{k}} \kappa(\mathbf{h},\mathbf{k})\cos(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}})} = \frac{T(\mathbf{h})}{B(\mathbf{h})}$$

$$\alpha(\mathbf{h})^2 = T(\mathbf{h})^2 + B(\mathbf{h})^2$$

$$\langle\alpha(\mathbf{h})\rangle = \left[\sum_{i=1}^{m}\kappa(\mathbf{h},\mathbf{k}_i)^2 + \sum_{i=1}^{m}\sum_{j=1}^{m}\kappa(\mathbf{h},\mathbf{k}_i)\kappa(\mathbf{h},\mathbf{k}_j) \times \frac{I_1|\kappa(\mathbf{h},\mathbf{k}_i)|I_1|\kappa(\mathbf{h},\mathbf{k}_j)|}{I_0|\kappa(\mathbf{h},\mathbf{k}_i)|I_0|\kappa(\mathbf{h},\mathbf{k}_j)|}\right]^{1/2}.$$

The constant C_W gives control over the phase refinement and we have found values between 0.2 and 0.3 to be best for giving steady refinement. However, the weighting scheme failed for the rubredoxin structure. Two weights $W(\mathbf{k})$ and $W(\mathbf{h}-\mathbf{k})$, which have the same form as $W(\mathbf{h})$ in (1), were cancelled in the original $\alpha(\mathbf{h})$ expression

$$\alpha(\mathbf{h}) = \Big(\Big\{\sum_{\mathbf{k}} W(\mathbf{k})W(\mathbf{h}-\mathbf{k})\kappa(\mathbf{h},\mathbf{k}) \times \sin[\varphi(\mathbf{k}) + \varphi(\mathbf{h}-\mathbf{k})]\Big\}^2 + \Big\{\sum_{\mathbf{k}} W(\mathbf{k})W(\mathbf{h}-\mathbf{k})\kappa(\mathbf{h},\mathbf{k}) \times \cos[\varphi(\mathbf{k}) + \varphi(\mathbf{h}-\mathbf{k})]\Big\}^2\Big)^{1/2}, \quad (2)$$

where

$$\kappa(\mathbf{h},\mathbf{k}) = |E(\mathbf{h})E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|, \quad (3)$$

so the value of $\alpha(\mathbf{h})$ in the weighting function is determined by

$$\alpha(\mathbf{h}) = \Big(\Big\{\sum_{\mathbf{k}}\kappa(\mathbf{h},\mathbf{k})\sin[\varphi(\mathbf{k}) + \varphi(\mathbf{h}-\mathbf{k})]\Big\}^2 + \Big\{\sum_{\mathbf{k}}\kappa(\mathbf{h},\mathbf{k})\cos[\varphi(\mathbf{k}) + \varphi(\mathbf{h}-\mathbf{k})]\Big\}^2\Big)^{1/2}. \quad (4)$$

Refining the phases replacing the old weighting factor with the new one gave good results. For instance, when the old weighting factor is applied to solve the structure of App, the calculated E map contains 129 (116) correct atoms by comparison with the known structure. The first number is the number of correct atoms of one enantiomorph, and the number in parentheses is the number of correct atoms of a second enantiomorph. Using the new form to solve the App structure, we obtained 135 (117) correct atoms from the calculated E map by comparison with the known structure. Furthermore, the fluctuation of the figure of merit ABSFOM, defined in (7), during the process of refinement is also less than in the original case which means that ABSFOM is similar for the correct trials. Using the new form, not only is the structure of rubredoxin solved successfully, but the solution of the structure of App is improved. The details of the structure-determination process are discussed later.

2.3. *A new figure of merit*

The old figure of merit XDFOM is unable to pick out the best solutions of the two structures. The original XDFOM is defined by

$$\text{XDFOM} = (1 + C)\text{ABSFOM}/(\varphi_0 R), \quad (5)$$

where

$$R = \sum_{\mathbf{h}}|\alpha(\mathbf{h}) - \langle\alpha(\mathbf{h})\rangle| \Big/ \sum_{\mathbf{h}}\langle\alpha(\mathbf{h})\rangle \quad (6)$$

$$\text{ABSFOM} = \sum_{\mathbf{h}}[\alpha(\mathbf{h}) - \alpha(\mathbf{h})_r] \Big/ \sum_{\mathbf{h}}[\alpha(\mathbf{h})_{\text{est}} - \alpha(\mathbf{h})_r] \quad (7)$$

$$\varphi_0 = \frac{\sum_{\mathbf{h}}\left|\sum_{\mathbf{k}}E(\mathbf{k})E(\mathbf{h}-\mathbf{k})\right|}{\sum_{\mathbf{h}}\left[\sum_{\mathbf{k}}|E(\mathbf{k})E(\mathbf{h}-\mathbf{k})|^2\right]^{1/2}}. \quad (8)$$

The constant $C = 1.25$ if there are heavy atoms in the structure and $C = 0$ otherwise. The reason for the failure is not only due to the two structures each containing one metal atom but also, in the case of App, because it belongs to the symmorphic space group $C2$. Thus, there is a tendency for phases to be refined to values that make the phase relationships work too well and the tangent formula overconsistent. The values of $\alpha(\mathbf{h})$ are too large and the value of R can be close to 1 or more. In such conditions the old figure of merit is too insensitive. A scaling factor Scal is now used to define a new residual R(sc) as

$$\text{Scal} = \sum_{\mathbf{h}}\alpha(\mathbf{h}) \Big/ \sum_{\mathbf{h}}\langle\alpha(\mathbf{h})\rangle \quad (9)$$

$$R(\text{sc}) = \sum_{\mathbf{h}}|\alpha(\mathbf{h})/\text{Scal} - \langle\alpha(\mathbf{h})\rangle| \Big/ \sum_{\mathbf{h}}\langle\alpha(\mathbf{h})\rangle. \quad (10)$$

The inverse of the residual R(sc) is treated as a new figure of merit for these special structures (XDFOM) and works quite effectively in picking out the best solutions. For other structures the original XDFOM is retained and is effective.

2.4. *Further development of the phase-estimating equations*

The phase-estimating equations used in *SYSTEM*95 are of the form in which least-squares methods are used to minimize RP,

$$\text{RP} = \text{PS} \times \text{RS} \geq 0, \quad (11)$$

where

Table 1. *The weighting system used in (14) for the cases when reflections have general phases or are restricted to the pairs (0°, 180°) and (90°, 270°)*

Reflection type	$S(\mathbf{h})$	$C(\mathbf{h})$
General	1	1
(0°, 180°)	0	1
(90°, 270°)	1	0

$$\mathrm{PS} = \frac{\sum_{\mathbf{l}} \left| \sum_{\mathbf{k}} E(\mathbf{k}) E(\mathbf{l}-\mathbf{k}) \right|}{\left[\sum_{\mathbf{l}} \sum_{\mathbf{k}} |E(\mathbf{k}) E(\mathbf{l}-\mathbf{k})|^2 \right]^{1/2}}. \tag{12}$$

The $\mathbf{l}$ are the indices of the weak reflections for the system. The form of PS is the same as the figure of merit PSI0 in the *MULTAN* program, but is used in the phase-estimating equations to keep the strong reflections in the determined system in touch with the weak reflections in the system, the phases of which are not being determined. Since the first subsystem is the basis for phase determination of other subsystems, PS is only calculated in the first subsystem to speed up the computational process. In other subsystems PS is set to 1.

The RS factor in the phase-estimating equation refers only to the internal relations of the system. It has been developed to have three forms. The first form is the left-hand side of the phase-estimating equation used in *SYSTEM*90, *i.e.*

$$\mathrm{RS} = \sum_{\mathbf{h}} B(\mathbf{h}) |\alpha(\mathbf{h})_e - \langle \alpha(\mathbf{h}) \rangle| \Big/ \sum_{\mathbf{h}} B(\mathbf{h}) \langle \alpha(\mathbf{h}) \rangle, \tag{13}$$

where

$$\alpha(\mathbf{h})_e = \Big(S(\mathbf{h}) \Big\{ \sum_{\mathbf{k}} \kappa(\mathbf{h}, \mathbf{k}) \sin[\varphi(\mathbf{k}) + \varphi(\mathbf{h}-\mathbf{k})] \Big\}^2 + C(\mathbf{h}) \Big\{ \sum_{\mathbf{k}} \kappa(\mathbf{h}, \mathbf{k}) \cos[\varphi(\mathbf{k}) + \varphi(\mathbf{h}-\mathbf{k})] \Big\}^2 \Big)^{1/2} \tag{14}$$

$$B(\mathbf{h}) = I_1[\langle \alpha(\mathbf{h}) \rangle] / I_0[\langle \alpha(\mathbf{h}) \rangle]. \tag{15}$$

$S(\mathbf{h})$ and $C(\mathbf{h})$ are weights for the real and imaginary parts of (14) for different reflections, as indicated in Table 1. The second form is

$$\mathrm{RS} = \sum_{\mathbf{h}} B(\mathbf{h}) \Big\{ \Big| \sum_{\mathbf{k}} K(\mathbf{h}, \mathbf{k}) \sin[\varphi(\mathbf{k}) + \varphi(\mathbf{h}-\mathbf{k})] - \langle \alpha(\mathbf{h}) \rangle \sin \varphi(\mathbf{h}) \Big| + \Big| \sum_{\mathbf{k}} K(\mathbf{h}, \mathbf{k}) \cos[\varphi(\mathbf{k}) + \varphi(\mathbf{h}-\mathbf{k})] - \langle \alpha(\mathbf{h}) \rangle \cos \varphi(\mathbf{h}) \Big| \Big\} \Big/ \sum_{\mathbf{h}} B(\mathbf{h}) \alpha(\mathbf{h}) \tag{16}$$

and the third form

$$\mathrm{RS} = \sum_{\mathbf{h}} B(\mathbf{h}) \langle \alpha(\mathbf{h}) \rangle \Big/ \sum_{\mathbf{h}} B(\mathbf{h}) \alpha(\mathbf{h}). \tag{17}$$

The three forms limit the phases of reflections within the system in different ways and have different advantages.

Table 2. *Figure-of-merit data for some of direct-methods trials for the App structure case.*

Set	Loop	Scale	R	XDFOM
1	10	1.707	0.435	2.299
2	13	1.878	0.401	2.494
3	12	1.685	0.434	2.303
4	15	1.763	0.407	2.455
5	10	1.888	0.389	2.570
6	6	3.264	0.109	−9.173
7	12	1.794	0.416	2.407
8	7	1.756	0.423	2.365
9	17	1.823	0.382	2.618
10	10	1.587	0.462	2.162
11	4	3.265	0.109	−9.173
12	8	1.712	0.434	2.306
13	3	3.264	0.109	−9.173
14	12	1.880	0.385	2.596
15	5	1.730	0.435	2.301
16	9	1.890	0.392	2.552
17	6	1.833	0.421	2.378
18	8	1.766	0.418	2.390
19	14	1.795	0.393	2.543
20	6	1.734	0.431	2.319
21	8	3.264	0.109	−9.172
22	11	1.780	0.418	2.392
23	22	1.777	0.401	2.492
24	11	1.795	0.384	2.606
25	9	2.202	0.338	2.958
26	10	1.646	0.450	2.222
27	4	1.769	0.431	2.319
28	5	1.831	0.426	2.348
29	3	3.264	0.109	−9.173
30	3	3.264	0.109	−9.173
31	12	1.739	0.411	2.433
32	4	2.378	0.299	3.340
33	4	2.378	0.297	3.368
34	6	1.799	0.423	2.367
35	6	1.735	0.434	2.305
36	9	1.693	0.432	2.315
37	20	1.761	0.408	2.452
93	8	1.662	0.442	2.260
94	3	3.264	0.109	−9.172
95	5	1.794	0.424	2.361
96	3	3.265	0.109	−9.172
97	9	1.727	0.428	2.338
98	8	2.377	0.299	3.345
99	3	1.730	0.443	2.257
100	4	3.265	0.109	−9.173

They can be selected according to judgement. These new phase-estimating equations use not only the internal relations but also the external relations of the determined system. This has proved to be more effective in practice than the old form of *SYSTEM*90.

The four strategies above have been incorporated into *SYSTEM*90 to generate a new program *SYSTEM*95, which has now successfully determined two protein structures with high-resolution native data. The software is easy to use and the mode of the structure-determination process is controlled by the code *EFM* following the keyword MOD, where

E = 0, 1, 2 representing different forms of the phase-estimating equations;

Table 3. *Figure-of-merit data for some of the direct-methods trials for the rubredoxin structure case*

Set	Loop	Scale	R	XDFOM
1	15	2.104	0.407	2.456
2	6	2.000	0.450	2.220
3	8	2.081	0.426	2.345
4	4	2.080	0.444	2.251
5	7	2.013	0.441	2.267
6	4	2.531	0.361	2.767
7	4	2.530	0.360	2.776
8	4	2.555	0.358	2.796
9	6	2.531	0.361	2.771
10	11	2.529	0.363	2.752
43	6	1.957	0.441	2.269
44	14	2.151	0.408	2.452
45	7	2.531	0.358	2.792
46	14	2.106	0.410	2.440
47	6	2.546	0.351	2.852
48	4	2.054	0.435	2.299
49	7	1.917	0.456	2.195
50	4	2.162	0.422	2.371
51	5	2.527	0.359	2.788
65	9	1.919	0.440	2.275
66	7	2.060	0.442	2.265
67	5	2.066	0.420	2.383
68	8	2.136	0.426	2.345
69	13	2.542	0.350	2.858
70	7	2.004	0.456	2.193
71	4	2.067	0.455	2.198
72	4	2.001	0.457	2.189
73	6	1.950	0.448	2.231
93	4	2.531	0.364	2.750
94	9	2.081	0.432	2.317
95	5	2.530	0.357	2.801
96	5	2.553	0.357	2.799
97	10	2.539	0.345	2.900
98	4	2.131	0.433	2.308
99	13	2.239	0.415	2.411
100	5	2.027	0.450	2.224

F = 0, 1 representing different refinement processes (SETF or SWTF, respectively);

M = 0, 1 representing different figures of merit in the recognition of the correct solution.

When M = 1, the inverse of $R(sc)$ has been used to pick out the correct solutions.

MOD 000, which is the default mode of *SYSTEM*95, has proved to be quite effective in the solution of many common structures. However, for the solution of macromolecular structures in symmorphic space groups, MOD 011 works more effectively according to our experimental results.

3. Direct-methods determination for App and rubredoxin structures

App (Glover *et al.*, 1983) is a 36-residue zinc protein. We used native protein data with 0.99 Å resolution comprising 1510 strong reflections with 72625 triplet relationships and 301 weak reflections with 26627 triplet relationships to determine the structure using MOD 011. Table 2 shows the figures of merit of some solutions of the 100 complete trials. It can be seen that set 33 corresponds to the largest XDFOM and sets 98 and 32 correspond to the second largest. The calculated E maps contain 135 (117) correct atoms in set 33, 127 (121) correct atoms in set 98 and 127 (117) correct atoms in set 32 by comparison with the known structure. The first number is the number of correct atoms of one enantiomorph, and the number in parentheses is the number of correct atoms of a second enantiomorph. Since set 33 resolved the enantiomorph best, the effectiveness of R(sc) was confirmed.

Rubredoxin (Dauter *et al.*, 1992) is a 52-residue iron-containing protein. We used native protein data with 0.99 Å resolution in space group $P2_1$ comprising 2015 strong reflections with 150935 triplet relationships and 391 weak reflections with 88391 triplet relationships to determine the structure by MOD 011. Table 3 shows the figures of merit of some solutions of the 100 complete trials. It can be seen that set 97 corresponds to the lowest R(sc) and sets 69 and 47 correspond to the second lowest. The calculated E maps contain 162 (109) correct atoms in set 97, 143 (141) correct atoms in set 69 and 148 (137) correct atoms in set 47 by comparison with the known structure. The effect of R(sc) was again verified, as set 97 best breaks the enantiomorphic ambiguity.

The results above show that *SYSTEM*95 is effective in solving these two structures. However, in terms of the value of Scal and the mixture of enantiomorphs in the E map, we find that the phases we determine give over-consistent relationships and the problem of the loss of enantiomorphic information needs to be solved. This is discussed in detail in the next section.

4. Method of model determination and results

It is difficult to resolve the correct enantiomorph from the resultant mixture. We attempt to solve the problem using the following steps based on the Fourier method and the stereochemistry.

4.1. *Step 1*

The largest fragment in the E map, calculated from the best direct-methods phase set, is included in the weighted recycling cycle with the metal atoms indicated by the large peaks, according to the chemical composition of the molecule. A larger fragment of the structure will usually thus be revealed, and the value of R will usually decrease. This overall process is repeated until the largest fragment of the structure develops no further. In this way, the correct component in the mixture of enantiomorphs develops so that the enantiomorphic ambiguity is resolved.

Table 4. *The process and results of App structure determination*

Method Weighted Fourier recycling	Input data: Atoms in fragment	Input data: Atoms in peak order	Figure of merit: R_f (%)	Figure of merit: R_0(%)
1	Zn+61C	–	56.86	–
2	Zn+106C	–	55.78	–
3	Zn+163C	–	51.30	–
4	Zn+196C	–	49.01	–
5	Zn+257C	–	47.87	–
6	Zn+273C	–	43.68	–
7	Zn+280C	–	42.00	–
8	Zn+289C	–	40.91	–
9	Zn+292C	–	40.59	–
10	–	Zn+180C	–	40.14
11	–	Zn+195C	–	38.41
12	–	Zn+210C	–	37.33
13	–	Zn+226C	–	36.38
14	–	Zn+241C	–	36.14
15	–	Zn+256C	–	36.08
16	–	Zn+256C (other enantiomorph)	–	35.87

Table 5. *The process and results of rubredoxin structure determination*

Method Weighted Fourier recycling	Input data: Atoms in fragment	Input data: Atoms in peak order	Figure of merit: R_f (%)	Figure of merit: R_0 (%)
1	Fe+93C	–	58.25	–
2	Fe+118C	–	56.84	–
3	Fe+138C	–	55.58	–
4	Fe+121C	–	54.42	–
5	–	Fe+39C	–	55.49
6	–	Fe+80C	–	54.14
7	–	Fe+120C	–	51.15
8	–	Fe+160C	–	48.68
9	–	Fe+201C	–	45.86
10	–	Fe+241C	–	43.00
11	–	Fe+281C	–	40.70
12	–	Fe+321C	–	38.97
13	–	Fe+362C	–	37.57
14	–	Fe+402C	–	37.43
15	–	Fe+402C	–	37.09
16	–	Fe+5S+397C	–	33.93
17	–	Fe+5S+397C MET 1	–	33.23

4.2. *Step 2*

Between one-third and two-thirds of the atoms of the largest fragment obtained by step 1 are input into the weighted Fourier recycling cycle, in order of peak height.

4.3. *Step 3*

The weighted Fourier recycling is continued by increasing the number of the input atoms in peak-height order by 5 to 10% of the total number of the non-H atoms in the asymmetric unit. Step 3 is then repeated until the value of R as defined in (6) decreases no further. If the value of R is then larger than 40%, step 2 is repeated. In this way the correct enantiomorph in the mixture becomes more and more dominant.

4.4. *Step 4*

If the structure contains not only metal atoms but also some heavy atoms such as P, S *etc.*, the value of R can decrease further if the corresponding large peaks are governed by good stereochemistry. They can, on this basis, be included in the next weighted Fourier recycling cycle.

The preparation of the Fortran coordinate file for the above recycling procedure is a tiresome procedure. In order to save time we have developed a special program *listing* which automatically generates the data file from the latest Fourier result for the next weighted Fourier recycling cycle.

*SYSTEM*95 determined the full structure model of App and rubredoxin successfully using the above methods. Table 4 shows the process and results for the App model determination. It can be seen that nine recycling cycles were performed in step 1. In the first cycle the Zn atom and 61 C atoms were found which gave an R value of 56.9%. By the ninth cycle the Zn atom and 292 C atoms were input which gave an R value of 40.6%. In the tenth cycle, step 2 input the Zn and 180 C atoms in peak-height order into the cycle, which gave an R value of 40.1%. Step 3 then increased the number of input atoms by a further 5% of the total number of non-H atoms in the asymmetric unit (302) in the following cycle until the value of R, which was 36.1% in the 15th cycle, no longer decreased. The atoms included in the 15th cycle were the Zn atom and 256 C atoms. In the 16th cycle, the Zn atom and 256 C atoms were again input but the enantiomorph was changed, giving an R value of 35.9%.

The final E map contained 262 correct atoms by comparison with the known structure. The map showed that the main chain of 34 residues was complete except for the Arg and Tyr residues at the N-terminus, which exhibit high thermal motion. The 25-residue side chains were complete.

Table 5 details the rubredoxin determination. Four cycles were performed in step 1 and the R value dropped from 58.3 to 54.4%. In the fifth cycle, step 2 input the Fe atom and 39 C atoms into the cycle, which gave an R value of 55.5%. Step 3 then increased the number of input atoms by a further 10% of the total number of non-H atoms in the asymmetric unit (403) in the following cycle until the value of R, which was 37.1% in the 15th cycle, no longer decreased. In the 16th cycle, the Fe atom and 397 C atoms were input. Five S atoms were indicated by peak height, giving an R value of 33.9%. In the 17th cycle, the condition was identical to the previous cycle except the keyword MET was set to 1, taking the metal effect into consideration.

From the final E map, the number of atoms in the largest fragment was 398 and the value of R was 33.2%. This result is even better than the App structure determination result. The final E map contained 373 correct atoms by comparison with the known structure. The E map shows that the main chain of 51 residues was complete except for the C-terminal residue. The 44-residue side chains were also complete.

The two models (App and rubredoxin) that we have determined are chemically very reasonable, and it is possible to break the enantiomorphic ambiguity. The final R values of the two structures are low, and it is well known that the structures can finally be solved by refinement from such starting points. This illustrates the effectiveness of our methods.

5. Concluding remarks

It is difficult to solve macromolecular structures in symmorphic space groups, *e.g.* $P1$, $C2$, $R3$, because in such space groups the tangent formulae tend to be perfectly consistent, particularly when the structure contains one or more metal atoms. Thus, the old figure of merit XDFOM no longer works properly and a serious enantiomorphic ambiguity occurs. The experimental results of these two native protein structure determinations show that the new program *SYSTEM*95 has conquered the above difficulties. There are four distinctive features in the structure-determination process.

(i) About 3% of direct-methods trials were successful for rubredoxin and App. This is a relatively high rate and certainly sufficient to solve such structures.

(ii) R(sc) can recognize the correct solution when the phases are overconsistent.

(iii) The correct enantiomorph is gradually resolved by inputting first the largest fragment followed by atoms in peak order of the latest E map into the weighted Fourier recycling cycle.

(iv) The value of R is used to assess the model and control its acceptance and thereby the generation of new models.

The project was supported by NSFC, the High Technology Department of the Scientific Committee of China and the Chinese Academy of Sciences. The authors are grateful to Professor M. M. Woolfson and Fan Hai-Fu, who provided data and were most helpful. The authors are also grateful to Professor Fu Heng for his advice and encouragement.

References

Dauter, Z., Seiker, L. C. & Wilson, K. S. (1992). *Acta Cryst.* B**48**, 42–59.

Glover, I., Haneef, I., Pitts, J., Wood, S., Moss, T., Tickle, I. & Blundell, T. L. (1983). *Biopolymers*, **22**, 293–304.

Hou, Y., Gao, M., Li, L. & Hou., P. (1994). *Acta Cryst.* A**50**, 748–753.

Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.

Weeks, C. M., Hauptman, H. A., Smith, G. D., Blessing, R. H., Teeter, M. M. & Miller, R. (1995). *Acta Cryst.* D**51**, 33–38.

Woolfson, M. M. & Yao, J. X. (1990). *Acta Cryst.* A**46**, 409–413.