Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Practical aspects of the integration of different software in protein structure solution

Vito Calderone

Department of Chemistry, University of Siena, Via Aldo Moro, 53100 Siena, Italy

Correspondence e-mail: vito.calderone@unisi.it

Received 22 January 2004
Accepted 3 August 2004

There is presently an increasing variety of choice in the software for macromolecular phasing and automated model building. In addition to its positive features, this variety poses the problem of which software to use in a specific crystallographic case. Moreover, it must be decided whether a sequence of programs should be used to achieve structure solution more accurately and more rapidly, taking into account the features of the different programs: some software is more suitable for dealing with low-symmetry rather than high-symmetry space groups in the detection of heavy-atom sites, while others can give better estimates of the figures of merit on phases in certain cases (which is crucial in dealing with maximum likelihood) and others are more suitable for chain tracing at low/medium-low resolution than at high resolution or *vice versa*. The 'integrated' choice of different software has become popular among crystallographers, especially when facing crystallographic cases that are not straightforward. A few examples will be presented on the use of different programs to achieve the goal of structure solution and the associated practicalities that can make the difference between solving or not solving a structure.

1. Introduction

In the era of structural genomics and high-throughput structural biology, the crystallographic community feels the need to solve structures in a fast, accurate and automated fashion. For this reason, there has been an increasing need for software that can somehow bypass a great deal of human intervention and 'decide' strategies automatically.

There are presently several programs that are equally accurate and highly automated for protein structure solution and each program has particular characteristics and features that make it more suitable in certain crystallographic cases than in others. In this sense, there is not yet a universal structure-solution piece of software and for this reason it may be a good idea, in many cases, to try an integrated approach, taking advantage of the individual properties and capabilities of each of these programs.

This kind of approach can in turn allow more reliable and more extensive heavy-atom site detection, a better preliminary phase refinement or a more efficient density-modification procedure, all of which have the effect of yielding more accurate phases, which eventually also has a positive effect on model building by making it faster and more efficient.

It is important in some cases to try not to stick to a single piece of software but also to try to use all the options that each

doi:10.1107/S0907444904019055

Table 1
MAD phasing and data-collection statistics for *E. coli* AphA.

Values in parentheses refer to the highest resolution shell.

	BR peak (λ = 0.91957 Å)	BR inflection (λ = 0.9204 Å)	BR remote (λ = 0.88561 Å)	AU remote (λ = 1.03770 Å)
Space group	$P2_12_12$			$I222$
Unit-cell parameters (Å, °)	a = 49.50, b = 92.62, c = 138.25, $\alpha = \beta = \gamma$ = 90			a = 49.28, b = 92.46, c = 138.18, $\alpha = \beta = \gamma$ = 90
Resolution (Å)	25.0–2.2 (2.23–2.20)	25.0–2.2 (2.23–2.20)	25.0–2.2 (2.23–2.20)	25–1.69 (1.79–1.69)
Total reflections	256647 (10044)	217935 (8426)	246719 (9458)	135549 (7369)
Unique reflections	32501 (1272)	32868 (1294)	32383 (1281)	31884 (2579)
Overall completeness (%)	98.5 (96.6)	99.3 (96.9)	97.8 (96.8)	98.2 (92.0)
Anomalous completeness (%)	93.6 (92.6)	91.5 (90.5)	96.5 (96.5)	—
$R_{\rm sym}$ (%)	9.2 (33.0)	7.3 (23.9)	11.4 (41.5)	7.5 (37.5)
$R_{\rm anom}$ (%)	9.5 (15.8)	8.9 (16.1)	7.8 (36.4)	—
Multiplicity	7.3 (7.0)	6.6 (6.5)	7.6 (7.4)	4.3 (2.9)
$I/\sigma(I)$	17.1 (3.9)	16.9 (4.2)	13.7 (2.8)	7.2 (1.8)
FOM (before solvent flattening)	0.28	0.27	0.18	—
FOM (all) (before solvent flattening)	0.47	—	—	—

program allows; the use of default options is again very practical and simple and works well in some cases, but it is often a good idea to try the secondary options by attempting to discover the 'hidden' buttons or keywords in the GUIs or scripts of each piece of software. It is possible that both the automatic and the integrated approaches may lead to structure solution, but is also possible that in some cases the integrated approach gives better phases, which in turn allows a faster and more accurate chain-tracing compliant with the tight schedule imposed by scientific competition. Therefore, what can at first appear to be a 'waste' of time can in the end reveal a more efficient way to solve a crystallographic problem.

Three examples will be shown here of crystallographic cases where the above-mentioned integrated approach has proved to be successful.

The first concerns a protein called AphA from *Escherichia coli*; it is an acid Mg phosphatase capable of hydrolysing several different phosphomonoesters as well as catalysing phosphate transfer to hydroxyl groups of organic compounds. Furthermore, AphA seems to be involved in the parental strand recognition of the DNA-replication origin. AphA is an oligomeric protein comprising four identical monomers of approximately 25 kDa each and is only present in some bacterial pathogens (Calderone, Forleo *et al.*, 2004; Forleo *et al.*, 2003).

The second example concerns a superoxide dismutase-like (SOD-like) protein from *Bacillus subtilis*; it shares sequence identity ranging from 45 to 30% with Cu and Zn SODs from other bacterial organisms. Of the bacterial proteins, it is the only one that does not conserve two of the residues of the copper-binding site and is reported to have an unknown function (Banci *et al.*, 2004)

The third example concerns another protein that is involved in copper homeostasis inside the cell; it is a truncated form (36 residues instead on 52 of the wild type) of copper thionein from yeast. This protein is capable of binding from six to eight $\mathrm{Cu^I}$ atoms per molecule through its ten cystein residues (Calderone, Dolderer *et al.*, 2004).

2. Data collection and processing

2.1. AphA from *E. coli*

A three-wavelength MAD experiment at the Br edge was performed at 100 K on a single derivatized AphA crystal using the rotation method at the EMBL X-31 PX beamline at DESY (Hamburg, Germany). The bromide-derivatized AphA crystal diffracted to 2.2 Å resolution and belongs to space group $P2_12_12$ (unit-cell parameters a = 49.50, b = 92.62, c = 138.25 Å), with two molecules in the asymmetric unit and a solvent content of about 60%.

A second three-wavelength MAD data set was collected at 100 K at the ESRF ID-29 beamline (Grenoble, France) from the $\mathrm{AuCl_3}$ derivative, which diffracted to 1.69 Å resolution. The space group was $I222$, with one molecule in the asymmetric unit and a solvent content of about 60%. This latter data set was not useful for solving the structure and has been used to refine the AphA structure at higher resolution.

Table 1 shows the data-collection statistics for the three wavelengths of the bromide derivative and for the remote wavelength of the gold derivative. The PDB codes for the bromide and gold derivatives are 1n9k and 1n8n, respectively.

2.2. SOD-like protein from *B. subtilis*

A SAD experiment at the Zn edge was performed on a crystal grown in the presence of zinc using the rotation method at the ELETTRA XRD-1 beamline (Trieste, Italy) at 100 K.

The crystal diffracted to 1.8 Å resolution and belongs to space group $P1$ (unit-cell parameters a = 38.22, b = 61.11, c = 64.91 Å, α = 84.35, β = 76.02, γ = 90.42°), with four molecules in the asymmetric unit and a solvent content of about 45%.

Table 2 shows the data-collection statistics. The PDB code is 1s4i.

Table 2
SAD data-collection statistics for SOD-like protein from *B. subtilis*.

Values in parentheses refer to the highest resolution shell.

	Zn peak (λ = 1.281 Å)
Space group	*P*1
Unit-cell parameters (Å, °)	a = 38.22, b = 61.11, c = 64.91, α = 84.35, β = 76.02, γ = 90.42
Resolution (Å)	37.0–1.8
Total reflections	232658 (8554)
Unique reflections	49677 (3958)
Overall completeness (%)	94.6 (92.4)
Anomalous completeness (%)	89.8 (71.8)
$R_{\rm sym}$ (%)	3.4 (23.2)
$R_{\rm anom}$ (%)	4.1 (18.2)
Multiplicity	4.7 (2.2)
$I/\sigma(I)$	13.8 (2.6)
FOM (before solvent flattening)	0.23

Table 3
Peak and remote data-collection statistics for yeast copper thionein.

Values in parentheses refer to the highest resolution shell.

	Cu peak (λ = 1.37 Å)	Cu remote (λ = 0.919 Å)
Space group	$P4_332$	$P4_332$
Unit-cell parameters (Å, °)	$a = b = c$ = 62.21, $\alpha = \beta = \gamma$ = 90	$a = b = c$ = 62.16, $\alpha = \beta = \gamma$ = 90
Resolution (Å)	31.1–1.65	27.8–1.44
Total reflections	170252 (22414)	173047 (25070)
Unique reflections	5475 (764)	7922 (1117)
Overall completeness (%)	100 (100)	100 (100)
Anomalous completeness (%)	98.6 (99.7)	–
$R_{\rm sym}$ (%)	8.1 (33.7)	7.0 (35.8)
$R_{\rm anom}$ (%)	9.5 (13.6)	–
Multiplicity	31.1 (29.3)	21.8 (22.4)
$I/\sigma(I)$	8.4 (2.1)	9.3 (2.1)
FOM (before solvent flattening)	0.25	—

Table 4
Refinement statistics for *E. coli* AphA.

Resolution range (Å)	25.0–1.69
$R_{\rm cryst}/R_{\rm free}$ (%)	17.7/20.6
Protein atoms	1644
Ions	1
Water molecules	283
R.m.s.d. bond lengths (Å)	0.01
R.m.s.d. bond angles (°)	1.4

2.3. Truncated form of yeast copper thionein

Two diffraction experiments at 100 K were performed using the rotation method at EMBL BW7A beamline at DESY (Hamburg, Germany); the first was carried out at the copper-edge wavelength (1.370 Å) and the second at 0.919 Å.

The first crystal diffracted to 1.7 Å resolution and the second diffracted to 1.4 Å resolution; both crystals belonged to the cubic space group $P4_332$ (unit-cell parameters $a = b = c$ = 62.17 Å, $\alpha = \beta = \gamma$ = 90°), with one molecule in the asymmetric unit and a solvent content of about 50%.

Table 3 reports the data-collection statistics for both data sets. The PDB code is 1rju.

All the above-mentioned data sets were processed using the program *MOSFLM* (Leslie, 1991) and scaled using the program *SCALA* (Evans, 1997) with the TAILS and SECONDARY corrections on (the latter restrained with a TIE SURFACE command) to achieve an empirical absorption correction.

3. Structure solution and different approaches to model building

3.1. AphA from *E. coli*

The phasing of AphA from *E. coli* was performed on the bromide-derivative MAD data with the program *SOLVE* (Terwilliger & Berendzen, 1999) assuming 20 bromide anions

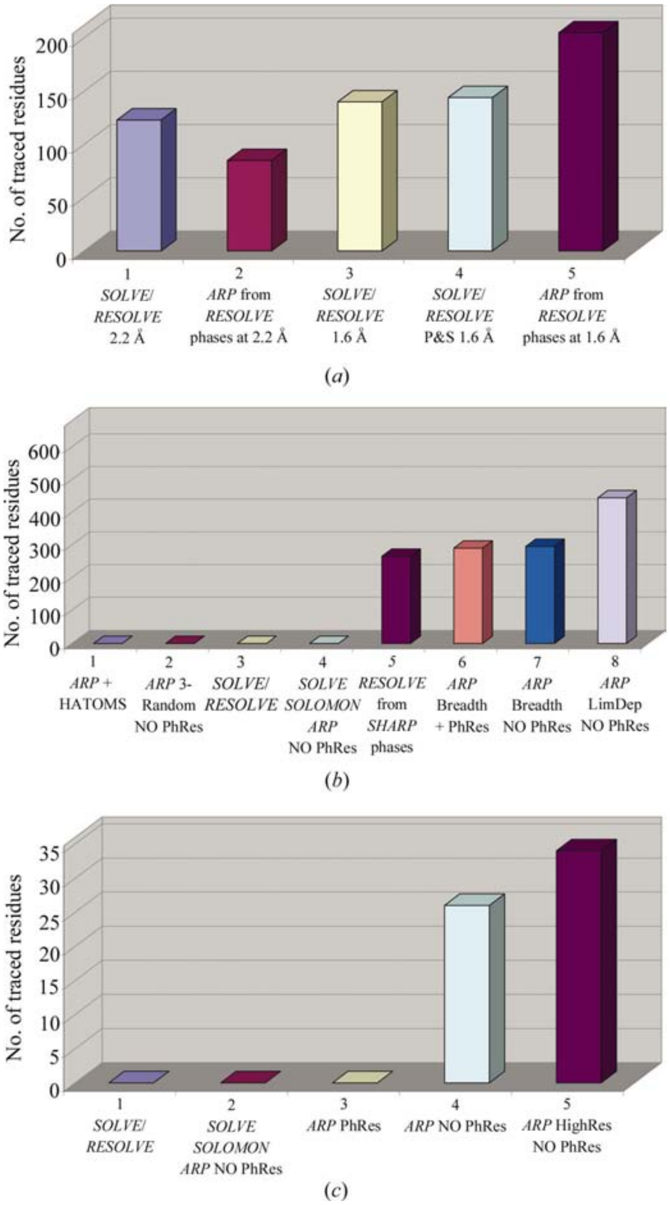

Figure 1
(*a*) Histogram representation of the number of residues built *versus* the phasing and building approach used for *E. coli* AphA (see text for details). (*b*) Histogram representation of the number of residues built *versus* the phasing and building approach used for SOD-like protein from *B. subtilis* (see text for details). (*c*) Histogram representation of the number of residues built *versus* the phasing and building approach used for yeast copper thionein (see text for details).

per asymmetric unit. The best solution yielded 18 Br atoms having good occupancies and displacement parameters; nine of these sites were related to the others by a non-crystallographic twofold axis. Density modification with NCS averaging was then applied, assuming two molecules in the asymmetric unit with a solvent content of 60%. The resulting electron-density map was of sufficient quality to allow partial tracing of the protein main chain (about 55% of the residues without side chains for each of the two chains in the asymmetric unit) with the program *RESOLVE* (Terwilliger, 2000, 2003).

Another approach was to use the solvent-flattened phases from *RESOLVE* and feed them into *ARP*/*wARP* 6.0 (Perrakis *et al.*, 1999), still using the same data at 2.2 Å resolution; this approach was less efficient and it was only possible to obtain about 40% of the residues without side chains for each of the two molecules in the asymmetric unit.

In order to try and improve phases by extending the resolution, one monomer from the partial solution of the NaBr derivative was then used as a starting model for molecular replacement on the remote-wavelength gold-derivative data at higher resolution with the software *AMoRe* (Navaza, 1994). The rotation function had the highest peak with a good correlation coefficient and the following translation function provided one clear solution which, after rigid-body refinement, gave a correlation coefficient of 42.6 and an *R* factor of 0.49.

This partial model was then combined with the gold-derivative data using *SIGMAA* (Read, 1986) to yield *SIGMAA*-weighted phases and figures of merit (FOMs); these phases were then fed into *RESOLVE* using the standard tracing protocol and the prime-and-switch option whose target is to reduce model bias. The number of residues traced was about 70 and 75%, respectively.

The best result in terms of the number of traced residues was obtained by feeding the molecular-replacement solution into *ARP*/*wARP* 6.0 without using phase restraints (*i.e.* without restraining phases to the Hendrickson–Lattmann phase probability distribution) and using the limited depth-first algorithm. The automatic building of the molecule was

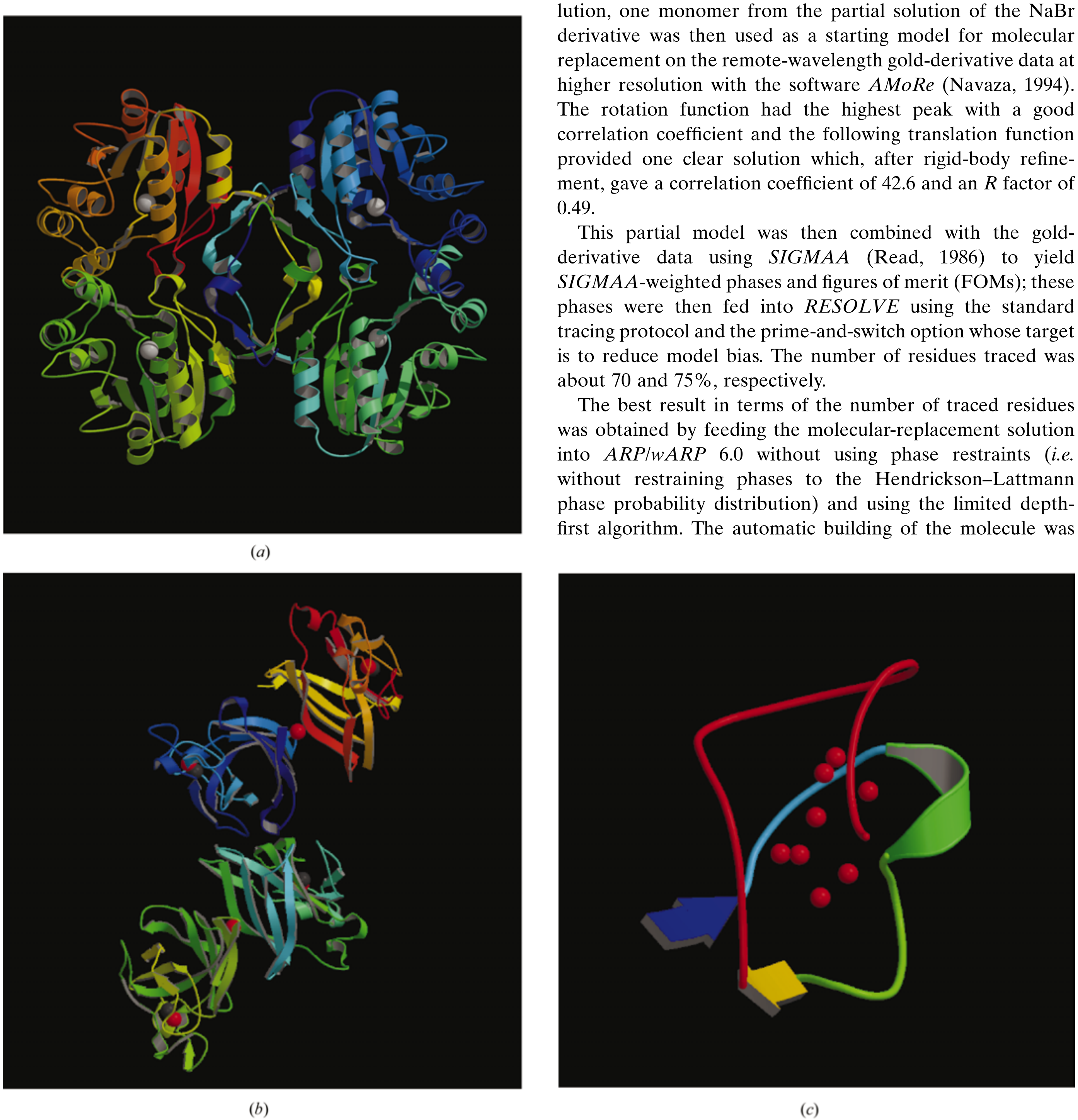

Figure 2
(*a*) Overall view of the functional tetramer of *E. coli* AphA. (*b*) Overall view of the asymmetric unit of SOD-like protein from *B. subtilis*. (*c*) Overall view of the structure of yeast copper thionein.

Table 5
Refinement statistics for SOD-like protein from *B. subtilis*.

Resolution range (Å)	37.0–1.8
$R_{\text{cryst}}/R_{\text{free}}$ (%)	22.0/25.7
Protein atoms (four molecules)	4520
Ions	6
Water molecules	398
R.m.s.d. bond lengths (Å)	0.02
R.m.s.d. bond angles (°)	2.7

able to assign about 95% of the structure (205 residues out of the 212 expected). The remaining residues were then built manually.

Table 4 shows the refinement statistics for the gold derivative.

Fig. 1(*a*) shows a histogram view of the different tracing approaches carried out on AphA from *E. coli*.

Fig. 2(*a*) shows the physiological tetramer built starting from the monomer obtained as described above and the magnesium-binding sites.

3.2. SOD-like protein from *B. subtilis*

The detection of the six heavy-atom sites was carried out using the program *SHELXD* (Schneider & Sheldrick, 2002).

The preliminary phases were then refined with the program *SHARP* (de La Fortelle & Bricogne, 1997) and solvent flattening was performed with *SOLOMON* (Abrahams & Leslie, 1996); the following chain-tracing protocols were attempted on the resulting phases.

The first attempt was to trace the chain using *ARP*/*wARP* 6.0, starting from the heavy-atom sites and the experimental structure-factor amplitudes without phase restraints; no residues were built this way.

The second attempt was to carry out a three-randomization run using *ARP*/*wARP* 6.0 without phase restraints; the randomization corresponds to a crude simulated-annealing procedure, which aims to drive the model out of possibly wrong local minima. This attempt also turned out to be unsuccessful, since no residues were built.

Another failed attempt with no residues built started from the known heavy-atom positions obtained from *SHELXD*, using them as input for *SOLVE*/*RESOLVE*.

In a fourth attempt, the known heavy-atom positions were again refined with *SOLVE*, but *SOLOMON* was then applied to perform density modification; the modified phases obtained in this way were then fed into *ARP*/*wARP* 6.0, but no residues were built.

A further attempt started from the phases obtained from *SHARP* and fed them into *RESOLVE*; this time, a chain was traced accounting for about 40% of the total number of residues.

A sixth attempt was to start again from the *SHARP* phases but this time to feed them into *ARP*/*wARP* 6.0, using the breadth-search algorithm and applying phase restraints (*i.e.* restraining phases to the Hendrickson–Lattmann phase probability distribution); the result was the building of about 45% of the total number of residues.

The seventh attempt was the same as the previous one but this time no phase restraints were applied; this approach resulted in some more residues being built.

The last and most successful attempt involved running *ARP*/*wARP* 6.0 on the *SHARP* phases using the limited depth-search algorithm without phase restraints; this approach gave about 75% of the total residues built.

Table 5 reports the refinement statistics.

Fig. 1(*b*) shows a histogram view of the different tracing approaches carried out on this protein.

Fig. 2(*b*) shows the four molecules in the asymmetric unit with the six Zn atoms.

3.3. Yeast copper thionein

This crystallographic case seemed to be straightforward, since the number of anomalous scatterers accounted for about 12% of the weight of the protein; for this reason, the anomalous signal was outstanding, being about 15–20% of the total signal. Despite this fact, several attempts with the most widely used software in protein crystallography proved to be unsuccessful.

The successful data set had a very high redundancy compared with the data sets collected previously and slightly better data-collection statistics. Therefore, seven of the eight copper positions were found using the anomalous dispersion method at the single wavelength of the copper edge (1.370 Å) with the program *SOLVE*; the preliminary phases obtained (FOM = 0.25) were then improved with the density-modification technique to an FOM of 0.78, using a solvent content of 50%, with the program *RESOLVE*.

Using these phases, several attempts at tracing have been performed.

The first was to use the chain-tracing routine of *RESOLVE*, but it was not possible to trace any residues in the electron-density map.

Two further unsuccessful results were obtained when the phases refined with *SOLVE* were density-modified with *SOLOMON* and then fed into *ARP*/*wARP* 6.0, with phase restraints in one case and without phase restraints in the other.

The best result was obtained when the phases from *RESOLVE* were used as input into *ARP*/*wARP* 6.0 using the limited depth-first search algorithm without phase restraints: 24 out of the total 36 residues of the protein were traced without side chains. The electron-density map now clearly showed the position of the eighth Cu atom, which was further confirmed by the presence of eight large peaks in the anomalous Fourier difference map.

When using the data set at higher resolution (1.4 Å) starting from the partial model available and using *ARP*/*wARP* 6.0 without phase restraints, the tracing was then of 34 residues out of 36. The two remaining residues were then added and all the side chains were placed manually.

Table 6 reports the refinement statistics for the high-resolution data set.

Fig. 1(*c*) shows a histogram view of the different tracing approaches carried out on yeast copper thionein.

Table 6
Refinement statistics for yeast copper thionein.

Resolution range (Å)	27–1.44
$R_{\rm cryst}/R_{\rm free}$ (%)	14.4/17.0
Protein atoms	256
Ligand atoms	8
Water molecules	64
R.m.s.d. bonds (Å)	0.02
R.m.s.d. angles (°)	2.0

Fig. 2(*c*) shows the overall structure of this protein coordinating the eight Cu atoms.

For all the three above-mentioned structures, the refinement was then carried out using *REFMAC*5 (Murshudov *et al.*, 1997) and the manual rebuilding and model visualization were performed with the program *XtalView* (McRee, 1999). The stereochemical quality of the refined models was assessed using the program *PROCHECK* (Laskowski *et al.*, 1993).

4. Conclusions

Automated phasing and model building performed by a single piece of software is a great tool in protein crystallography, but sometimes one program does not work or gives limited results; this situation can be improved by using different strategies. Each program in fact has very particular features, which can simultaneously be weak and strong points, such as the ability to trace better at low than at high resolution or the ability to give more realistic figures of merit on the initial or on the density-modified phases (which is essential when using maximum-likelihood methods).

As shown in the examples above, the best results have been obtained through the combined use of different programs for heavy-atom detection, phasing, density modification and chain tracing.

Furthermore, default options in model-building programs generally work well, but in cases that do not succeed fully in the first place (*e.g.* very limited chain tracing using default options proposed by the program) it could be worth spending some time trying the secondary options, such as the alternative search algorithm (in the case of *ARP*/*wARP*) or the prime-and-switch option (in the case of *RESOLVE*).

As a general rule of thumb, on the basis of the results described above, the limited depth-search algorithm seems to work better than the breadth search; the latter algorithm explores all possible further connections (but only peptide-unit deep) from each built peptide and iteratively eliminates the worst ones until a single chain remains. By never looking further than one peptide unit, this method can be defined as ‘local’ in terms of the geometric features that can be employed. For poor densities a new algorithm (the limited depth-search algorithm) was implemented, which searches deeper into the tree of peptide connections and looks for long fragments of good geometric quality. In the case of very good data at high resolution, however, the breadth-search algorithm seems to work substantially better, although much more slowly, than the limited depth-search algorithm.

In the case of *RESOLVE*, the prime-and-switch option, which is advised when starting from a molecular-replacement partial solution, does not seem to affect the chain tracing. This program seems to work better in the case of medium-resolution data.

Another rule of thumb concerns the phase-restraints option in *ARP*/*wARP* iterative refinement; it usually makes the tracing less efficient, but the result could depend on the accuracy of the estimation of the probability distribution by the phasing programs.

Several people from various groups and institutions in Siena, Florence and Tubingen have made it possible to carry out the work described above; I would like to thank them all for the valuable collaboration and expertise. A special acknowledgement to my supervisor Professor Stefano Mangani for his precious suggestions and support throughout this experience. Another acknowledgment to Professor Ivano Bertini, Professor Lucia Banci and Professor Claudio Luchinat from CERM (University of Florence) for giving me the opportunity to take part to their research projects through valuable collaborations. I also gratefully acknowledge the beamline staff at the ESRF (Grenoble, France) facility, the European Community Access to Research Infrastructure Action of the Improving Human Potential Programme to the EMBL Hamburg Outstation (contract No. HPRI-CT-1999-00017) and the ELETTRA XRD-1 (Trieste, Italy) beamline staff. This work has been financially supported by the Italian MURST COFIN01.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

Banci, L., Bertini, I., Calderone, V., Del Conte, R., Fantoni, A., Mangani, S., Quattrone, A. & Viezzoli, M. S. (2004). Submitted.

Calderone, V., Dolderer, B., Echner, H., Hartmann, H.-J., Del Bianco, C., Luchinat, C., Mangani, S. & Weser, U. (2004). Submitted.

Calderone, V., Forleo, C., Benvenuti, M., Thaller, M. C., Rossolini, G. M. & Mangani, S. (2004). *J. Mol. Biol.* **335**, 761–773.

Evans, P. R. (1997). *Jnt CCP4/ESF–EABCM Newsl. Protein Crystallogr.* **33**, 22–24.

Forleo, C., Benvenuti, M., Calderone, V., Schippa, S., Doquier, J. D., Thaller, M. C., Rossolini, G. M. & Mangani, S. (2003). *Acta Cryst.* D**59**, 1058–1060.

La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

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

Leslie, A. G. W. (1991). *Crystallographic Computing V*, edited by D. Moras, A. D. Podjarny & J. P. Thierry, p. 50. Oxford University Press.

McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.

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

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

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

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.

Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.

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

Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.