Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Matrix methods for solving protein substructures of chlorine and sulfur from anomalous data

Rudolf A. G. de Graaff,* Mark Hilge, Jaco L. van der Plas and Jan Pieter Abrahams

Leiden Institute of Chemistry, Gorlaeus Laboratories, University of Leiden, PO Box 9502, 2300 RA Leiden, The Netherlands

Correspondence e-mail: rag@chema110.leidenuniv.nl

Received 17 July 2001
Accepted 4 October 2001

The weak signal obtained from the anomalous scattering (at $\lambda = 1.54$ Å) of naturally occurring elements such as sulfur, phosphorus and ordered solvent chloride ions is used to determine the atomic positions of these atoms. Two examples are discussed: the sulfur and chlorine substructure of tetragonal hen egg-white lysozyme and an oligonucleotide containing ten P atoms. The substructure of lysozyme was also solved from Cu $K\alpha$ radiation data collected on a standard rotating-anode generator. The results presented here are an illustration of the power of the matrix methods, which are to be implemented in next distribution of the direct methods package *CRUNCH*.

1. Introduction

Traditionally, the Patterson function has been used to determine the positions of the heavy-atom sites in SIR and MIR, as well as those of the anomalous scatterers in SAD and MAD experiments. However, if many sites are to be identified solution of the Patterson is not straightforward. An alternative is the use of direct methods. As early as 1989, Mukherjee, Helliwell & Main published a paper on the use of *MULTAN* to locate anomalous scatterers based on the differences between $I(+H)$ and $I(-H)$ caused by anomalous dispersion (Mukherjee *et al.*, 1989).

Recent developments in the field of direct methods have led to spectacular results. Substructures as large as 160 Se atoms have been determined (Hauptman, 2000). These successes prompted the next development: phasing native protein data using the weak anomalous signal of elements such as sulfur, chlorine and phosphorus to determine the positions of these atoms in the cell. If the fraction of anomalous scatterers is not smaller than approximately 1%, then generally the phasing power is sufficient to obtain the entire structure. Procedures such as this have the advantage of avoiding the necessity of preparing and measuring heavy-atom derivatives.

The first report using this approach was by Hendrickson & Teeter (1981) and described the structure determination of crambin from the anomalous scattering of six S atoms. Wang discussed the structure determination of a small protein (12 kDa) using simulated data and the signal from two sulfurs (Wang, 1985). In 1999, a paper appeared discussing the *ab initio* structure determination of lysozyme from the anomalous signal only (Dauter *et al.*, 1999). Synchrotron data of wavelength 1.54 Å and resolution 1.53 Å were used. The so-called 'half-baked' program *SHELXD* developed by Sheldrick (1997, 1998*a*,*b*) generated coordinates for 17 of the S/Cl atoms present. The positions obtained proved to be sufficiently accurate to enable solution of the phase problem.

Recently, using the same approach, the structure of the Z-DNA hexamer duplex d(CGCGCG)$_2$ was determined based on the anomalous dispersion signal of the P atoms present in the compound. The oligonucleotide crystallizes in space group $P2_12_12_1$. The asymmetric unit contains 290 atoms, ten of which are phosphorus. Synchrotron data to a resolution of 1.50 Å were collected at a wavelength of 1.54 Å. The anomalous dispersion signal of phosphorus is even weaker than that of sulfur and chlorine (the $\delta f''$ values at 1.54 Å are 0.43, 0.56 and 0.70, respectively). However, because of the relatively small size of the oligonucleotide, the phase problem in this case is easier to solve (Dauter & Adamiak, 2001).

This paper deals with a successful attempt to address the phase problem by a fundamentally different direct method. The matrix method developed by the authors is especially suited to problems of this sort, showing a comparatively high hit rate. Moreover, we were able to solve the sulfur/chlorine substructure of lysozyme from in-house data.

2. The method

A comprehensive description of the ideas and algorithms used is given in two papers (van der Plas *et al.*, 1998*a*,*b*). Here, a brief outline only is presented.

Let **A** be a Karle–Hauptman matrix of order M with $M > N$, the number of atoms in the unit cell of the structure under consideration. Under the condition that the phases and magnitudes of the elements are exact, the rank of **A** is N and **A** is semi-positive definite. These properties form the basis of an iterative process of phase refinement.

Fig. 1 shows a simple flow chart illustrating the matrix method. Starting from a model containing a number of atoms equal to the number of sites to be found, randomly positioned in the asymmetric unit, phases are calculated for all reflections present in a given Karle–Hauptman matrix. The matrix is constructed to maximize the average E value of the reflections contained therein. During the refinement the phases are changed, minimizing the number and the magnitudes of the negative eigenvalues of the matrix. The refinement is continued until a local optimum is reached.

A map is then calculated, new atomic positions are obtained and the resulting model is evaluated by calculating R_2 and a correlation function. R_2 is defined as

$$R_2 = \frac{\sum_{\mathbf{H}} [|E_o(\mathbf{H})|^2 - \eta^2 |E_c(\mathbf{H})|^2]^2}{\sum_{\mathbf{H}} |E_o(\mathbf{H})|^4},$$

where the suffixes o and c designate the observed and calculated normalized structure factors, and η^2 is the fraction of the scattering power used for the calculation of E_c. The quotient of R_2 and the correlation is used as a figure of merit (Fom). If the model is of insufficient quality, 30% of the atoms are removed from the model (Sheldrick, 1998*a*,*b*). The choice of the atoms selected for deletion is random. The resulting atomic positions are used as the new basis for the matrix refinement.

Table 1
The in-house data.

Given are $R_{\rm merge}$, $I/\sigma(I)$, the redundancy and the completeness as a function of the resolution.

$d_{\rm min}$	$R_{\rm merge}$	$I/\sigma(I)$	Redundancy	Completeness (%)
3.49	0.041	70.0	21.8	99.7
2.81	0.044	73.5	21.9	99.3
2.46	0.051	67.8	21.8	99.4
2.24	0.059	60.3	21.6	98.0
2.09	0.066	55.2	21.4	98.2
1.96	0.108	23.3	19.8	97.5
1.87	0.095	41.9	21.3	96.6
1.79	0.118	34.0	21.0	95.7
1.72	0.139	28.3	20.9	96.5
1.66	0.178	16.8	18.0	89.3
All *hkl*	0.053	59.8	20.9	97.1

Table 2
Lysozyme: ten trials using the synchrotron data.

The number of atoms found, the average error, the maximum error and the Fom are given both before and after a final optimization step

	Before optimization				After optimization			
Trial	Found	Error (Å)	Max. error (Å)	Fom	Found	Error (Å)	Max. error (Å)	Fom
1	13	0.36	0.98	0.56	17	0.10	0.23	1.03
2	—	—	—	0.24	—	—	—	0.26
3	—	—	—	0.26	—	—	—	0.30
4	—	—	—	0.26	—	—	—	0.27
5	—	—	—	0.25	—	—	—	0.30
6	14	0.32	0.73	0.57	17	0.10	0.25	1.01
7	13	0.35	0.88	0.55	17	0.10	0.25	1.03
8	13	0.41	1.06	0.50	17	0.10	0.28	1.04
9	13	0.30	0.99	0.51	17	0.10	0.24	1.03
10	13	0.34	0.96	0.51	16	0.09	0.16	1.03

For a given trial, the cycles are stopped if the average variation of the Fom is less than 2% over 40 consecutive cycles or if no stable solution has been found within 400 cycles. The model is next optimized by careful peak-picking, maximizing the same figure of merit.

In common with most current phase-refinement procedures, the method uses alternate modification of the model in real and reciprocal space (Abrahams & de Graaff, 1998). An important difference is that in our procedure, refinement in reciprocal space can be carried through until convergence is reached, without over-refinement occurring.

3. The data

For details of the measurements and processing of the data collected at the Brookhaven synchrotron we refer to the paper by Dauter *et al.* (1999).

Locally, the data were collected on an FR591 copper-source rotating-anode generator (Nonius, Delft, The Netherlands) equipped with Osmic mirrors and a MAR345 image plate. Before the measurements, a lysozyme crystal (dimensions 0.2 × 0.2 × 0.2 mm) was transferred to the cryoprotectant solution containing 30% glycerol and flash-frozen in the Cryostream (N_2, 100 K): 279 1° scans were taken, using an exposure time of 20 min. The resolution of the data is 1.66 Å.

The data were processed carefully using a beta version of *HKL*2000 (Otwinowski & Minor, 1997). Table 1 shows $R_{\rm merge}$, $I/\sigma(I)$, the redundancy and the completeness as a function of the resolution.

For all calculations of E values from the anomalous difference data the program suite *DREAR* was used as implemented in *SnB* version 2.1 (Blessing & Smith, 1999). Fig. 2 contains a comparison of the lysozyme data from Brookhaven and our home measurements on the rotating-anode generator. Depicted are the ratio of the signal to noise for the difference structure factors, averaged in 25 resolution bins. As is to be expected, the synchrotron measurements yield data with a higher signal-to-noise ratio than a standard laboratory source.

4. Results

Matrices of order $2N$ were used throughout for phase refinement. A maximum of 400 cycles per trial was chosen, alternating between phase refinement and model evaluation and modification in each cycle.

4.1. Lysozyme, synchrotron data

Ten trials, each based on a fresh random start, resulted in six solutions. The values of Fom range from 1.00 to 1.04. The positions of the ten S atoms and seven solvent chlorines present are identified correctly, except in solution number 10 where only 16 atoms are placed correctly. A typical value of Fom for a non-solution is 0.27. The quality of the coordinates is very good; on average, the error in the coordinates is about 0.1 Å. Table 2 summarizes the results. The errors quoted are obtained by comparison with the refined coordinates (Dauter *et al.*, 1999).

4.2. The oligonucleotide

Five trials generated five solutions for the phosphorus substructure of the oligonucleotide. Foms range between 1.02 and 1.07. Compared with the refined phosphorus coordinates, the best solution has an average absolute error of 0.08 Å; the maximum error is 0.14 Å.

We extended the phosphorus substructure to the complete structure using the 1.54 Å data as well as data measured at a wavelength of 0.98 Å to a resolution of 0.95 Å. *AUTOFOUR* (Kinneging & de Graaff, 1984) is a standard 'small-molecule' approach to the problem of finding the complete structure from a small fragment. It is implemented in the direct-methods package *CRUNCH* (de Gelder *et al.*, 1993). Extension to the full structure by *AUTOFOUR* is automatic if the 0.95 Å data are used. The complete structure was obtained; the R_2 value based on E calculated for the final model was 14%.

Unsurprisingly, using data with a resolution of 1.5 Å this approach did not yield a model of the same quality. However, coordinates are found for all 290 atoms, resulting in a value of R_2 of 32% for the model obtained, indicating the model to be substantially correct. We have not tried to refine this model.

In this case, obtaining more accurate coordinates of all the atoms implies using a more sophisticated means of identifying atomic positions than a simple peak search.

4.3. Lysozyme, in-house data

As we expected this gave a lower hit rate, 20 trials were calculated instead of ten. Table 3 shows the results. Five good solutions to the phase problem were obtained. Non-solutions are clearly distinct from successful trials. However, the lower information content of the local data is reflected in both the lower hit rate and the slightly larger error in the coordinates obtained: 0.2 Å on average. Also, the solutions with 16 atoms placed correctly are indistinguishable from those with all 17 atoms present.

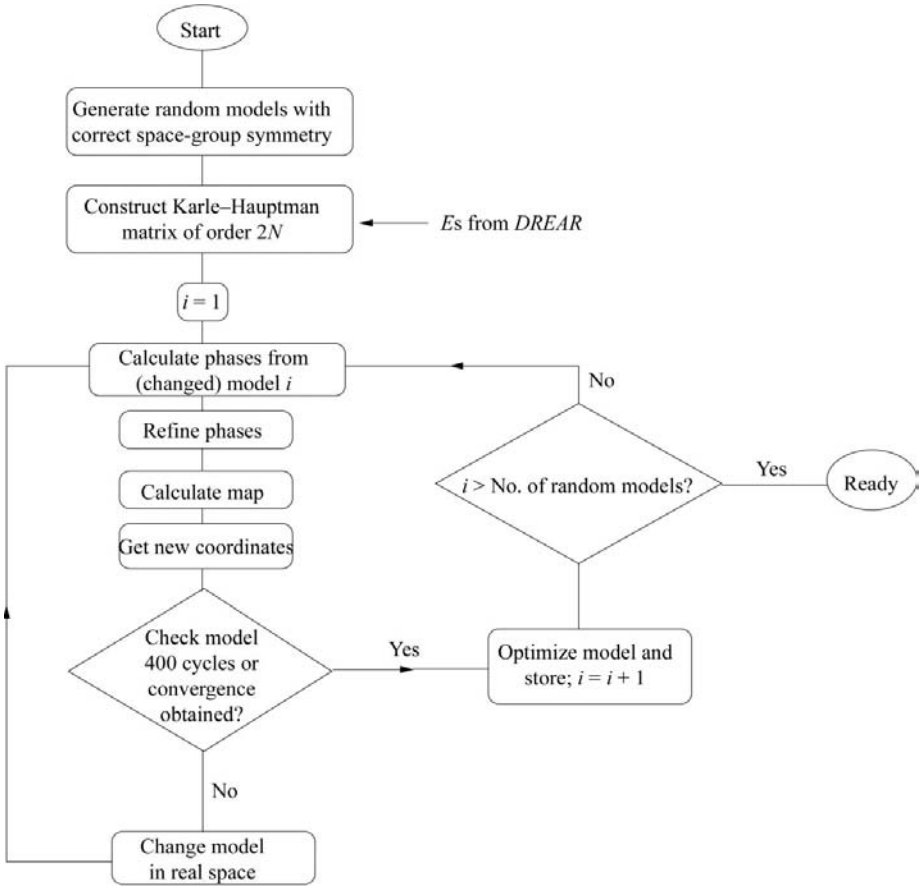

Figure 1
A flow chart of the matrix method.

Table 3
Lysozyme: in-house data.

A selection of trials, including the five solutions obtained. The number of atoms found, the average error, the maximum error and the Fom are given both before and after a final optimization step.

	Before optimization				After optimization			
Trial	Found	Error (Å)	Max. error (Å)	Fom	Found	Error (Å)	Max. error (Å)	Fom
1	—	—	—	0.19	—	—	—	0.37
2	—	—	—	0.20	—	—	—	0.43
3	15	0.25	0.46	0.44	16	0.17	0.28	0.71
4	15	0.27	0.58	0.48	17	0.18	0.44	0.73
6	14	0.27	0.57	0.43	16	0.18	0.44	0.72
8	—	—	—	0.23	—	—	—	0.40
9	15	0.24	0.55	0.42	17	0.18	0.44	0.71
11	—	—	—	0.20	—	—	—	0.39
15	16	0.25	0.66	0.54	17	0.17	0.47	0.71
20	—	—	—	0.19	—	—	—	0.39

Starting from the solution with the highest Fom phases are calculated in *MLPHARE* (Otwinowski, 1991). Only the individual *B* factors of the 17 atoms were refined; the coordinates were fixed to the values obtained from the matrix method. The overall acentric phasing power was 2.2. The phases were improved by 20 conventional cycles of density modification using *DM* (Cowtan & Zhang, 1999). After 570 cycles of *ARP*/*wARP*, *warpNtrace* (Perrakis *et al.*, 1997) built a model containing 125 residues, including all ordered side chains (97% of the complete molecule). Phasing and density-modification steps were carried out using defaults. Anisotropic refinement of the model using *REFMAC* (Murshudov *et al.*, 1999) resulted in a final R of 16.1% ($R_{\rm free}$ = 21%). The map correlations between the final $2F_o - F_c$ map and maps based on the initial phases, the phases after refinement of individual *B* factors, after density modification and after *warpNtrace* are 0.29, 0.47, 0.64 and 0.97, respectively. Figs. 3 and 4 illustrate the progress during the structure determination.

5. Discussion

For the type of problem discussed, the matrix-based methods proposed by the authors clearly constitute a strong alternative to the probabilistic approach which is common in direct methods. In conventional direct methods, statistical phase relations are used with probabilities calculated from the available magnitudes. Clearly, in the case of SAD and SIR the errors in these magnitudes are very much larger than in the case of conventional *ab initio* small-molecule structure determinations. Therefore, the foundations are undermined by the minimum principle (Debaerdemaeker & Woolfson 1983) as used in *SnB* and by the tangent refinement as used in *SHELXD*.

This fact explains why *SHELXD* solving the lysozyme substructure achieves a hit rate of only 0.2% (Dauter *et al.*, 1999). While the quality of the individual magnitudes is poor, a large number of data per atom are available. The matrix method, which is not based on estimated values of probabilities but on exact properties which are fundamental to the process of diffraction, is very robust. Using the synchrotron data, our experiments show a high hit rate of approximately 60%.

The oligonucleotide poses a much smaller problem: the multiplicity of the general position is only four and only ten phosphorus positions need to be determined. The five solutions mentioned were generated within a quarter of an hour.

We have tried to solve the complete structure *ab initio* from the 0.95 Å data. Both running with default parameters, the current distribution of *CRUNCH* – which as yet does not contain the matrix method – as well as *SnB* 2.1 (Weeks &

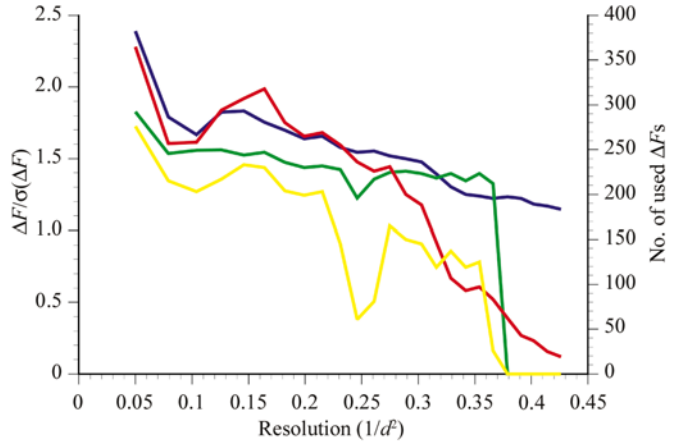

Figure 2
$\Delta F/\sigma(\Delta F)$ for both the synchrotron (blue) and in-house (green) data sets of lysozyme as a function of resolution. Also shown is the number of usable reflections in each case (red, synchrotron; yellow, in-house).

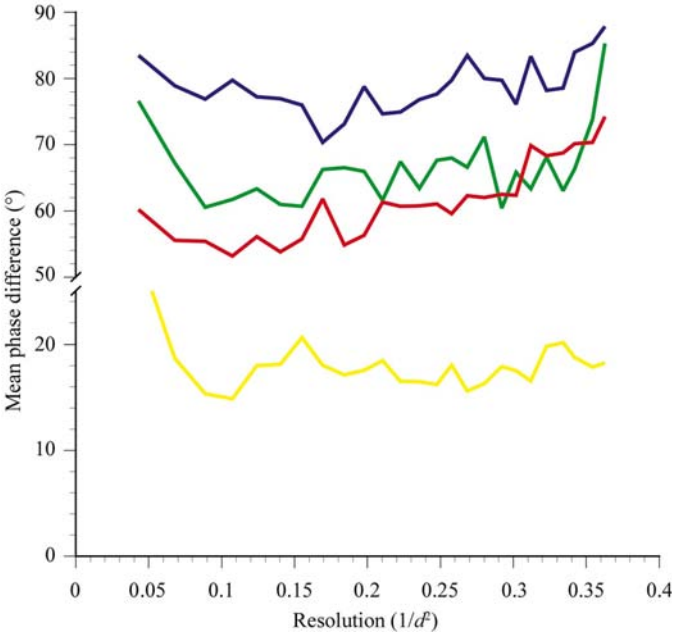

Figure 3
The in-house data. Average phase errors as a function of resolution after phase calculation in *MLPHARE* (*CRUNCH*; blue), after refinement of individual *B* factors for the S/Cl atoms (*MLPHARE*; green), after density modification (*DM*; red) and finally after automatic model building (*warpNtrace*; yellow). The phases are compared with those of the anisotropically refined structure.

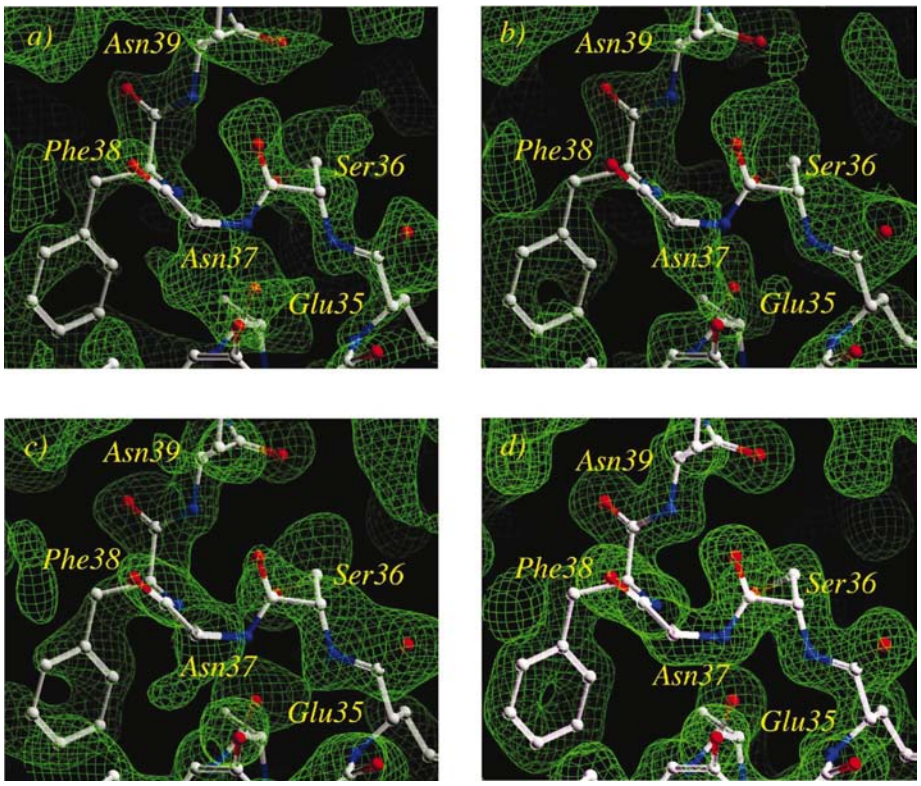

Figure 4
The in-house data. A region of the map calculated to the maximum resolution of 1.66 Å; (*a*) after phase calculation in *MLPHARE*, (*b*) after refinement of individual *B* factors for the S/Cl atoms, (*c*) after density modification and finally (*d*) after automatic model building.

Miller, 1999) failed to give a solution. In view of this, it is well worth considering solving the phase problem in this and similar cases by first finding the anomalous scatterers and building from there. If data are available at atomic resolution using the atomic positions of the phosphorus to find the other atoms only takes a few minutes.

The matrix method solves the substructure of lysozyme from our in-house data. As far as the authors are aware, this is the first time the sulfur/chlorine substructure of lysozyme has been determined from anomalous data measured on a common laboratory setup. We have found the careful measurement and processing of the data to be of crucial importance. In particular, systematic errors have to be eliminated or reduced by collecting a data set of high multiplicity. Initial attempts, albeit using an older version of *HKL*2000, failed to solve the substructure from in-house data obtained from a 160° scan. This is consistent with the results obtained by Dauter *et al.* (1999). An experiment attempting to solve the substructure omitting 30 reflections corresponding to *d* values of 8.0 Å and higher yielded just one solution from 20 trials.

Obviously, the upper limit of 400 cycles we have used is arbitrary; more cycles would very likely result in more solutions per trial. It is important to note that once a set of phases and coordinates corresponding to an approximate solution has been obtained, matrix refinement in reciprocal space does not lead to divergence.

Each cycle of our procedure implies many calculations. On a Pentium III 766 MHz machine a cycle takes about 1 min. The order of the algorithm is N^3, where N is now the number of anomalous scatterers in the unit cell. The cost-determining step in the method is the calculation of the eigenvalues and eigenvectors of the matrix. Currently, we are considering various ways of speeding up the calculations. Possibilities are firstly setting all E values below a certain threshold to zero in the Karle–Hauptman matrix. The eigenvalue problem could then be solved by sparse-matrix routines. Secondly, and perhaps more promising, we aim to use the eigenvectors and values of one cycle as an approximation to those of the next, finding updated values by iteration.

In conclusion, our results compare very favourably with those obtained previously using more conventional probabilistic direct methods. The phase refinement in reciprocal space is resolution-independent; however, the current implementation requires the sites to be determined to be resolved. A version of the software for general use, also suitable for finding heavy-atom positions from SIR data, will become part of the distribution of the direct-methods package *CRUNCH* (http://chema110.leidenuniv.nl/~rag).

The authors would like to thank Dr Z. Dauter of the Brookhaven National Laboratory, New York, USA, for generously providing us with the synchrotron data for both lysozyme and the oligonucleotide.

References

Abrahams, J. P. & de Graaff, R. A. G. (1998). *Curr. Opin. Struct. Biol.* **8**, 601–605.
Blessing, R. H. & Smith, G. D. (1999). *J. Appl. Cryst.* **32**, 664–670.
Cowtan, K. D. & Zhang, K. Y. J. (1999). *Prog. Biophys. Mol. Biol.* **72**, 245–270.
Dauter, Z. & Adamiak, D. A. (2001). *Acta Cryst.* D**57**, 990–995.
Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.
Debaerdemaeker, T. & Woolfson, M. M. (1983). *Acta Cryst.* A**39**, 193–196.
Gelder, R. de, de Graaff, R. A. G. & Schenk, H. (1993). *Acta Cryst.* A**49**, 287–293.
Hauptman, H. (2000). Personal communication.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
Kinneging, A. J. & de Graaff, R. A. G. (1984). *J. Appl. Cryst.* **17**, 364–366.

Mukherjee, A. K., Helliwell, J. R. & Main, P. (1989). *Acta Cryst.* A**45**, 715–718.
Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). *Acta Cryst.* D**55**, 247–255.
Otwinowski, Z. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Scattering and Anomalous Replacement*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Perrakis, A., Sixma, T. K., Wilson, K. S. & Lamzin, V. S. (1997). *Acta Cryst.* D**53**, 448–455.
Plas, J. L. van der, de Graaff, R. A. G. & Schenk, H. (1998*a*). *Acta Cryst.* A**54**, 262–266.
Plas, J. L. van der, de Graaff, R. A. G. & Schenk, H. (1998*b*). *Acta Cryst.* A**54**, 267–272.
Sheldrick, G. M. (1997). *Proceedings of the CCP4 Study Weekend. Recent Advances In Phasing*, edited by K. S. Wilson & G. Davies, pp. 147–158. Warrington: Daresbury Laboratory.
Sheldrick, G. M. (1998*a*). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 401–411. Dordrecht: Kluwer Academic Publisher.
Sheldrick, G. M. (1998*b*). *Riding the Fence Between Large and Small Molecules.* Annu. Meet. Am. Crystallogr. Assoc., Workshop WK02.
Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.
Weeks, C. M. & Miller, R. (1999). *J. Appl. Cryst.* **32**, 120–124.